Summer 8-15-2015

Neural Mechanisms of Working Memory Cortical Networks

Charalampos Papadimitriou
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Neurosciences

Dissertation Examination Committee:
Lawrence H. Snyder, Chair
Richard Abrams
Todd Braver
Vitaly Klyachko
Camillo Padoa-Schioppa

Neural Mechanisms of Working Memory Cortical Networks

by
Charalampos Papadimitriou

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in St. Louis
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

August 2015
St. Louis, Missouri

Table of Contents

List of Figures

Endure. In enduring, grow strong.

Dak'kon

Acknowledgements

I thank the people who – through their support, encouragement, or challenge – have instilled and continue to instill ambitions in me. I also thank the people who provided the resources, both intellectual and financial, necessary to complete this work.

The list of people who significantly contributed to the completion of this work through one of the aforementioned methods is long. The list includes Larry Snyder, whose infectious curiosity makes every question interesting, and the members of my thesis committee – Vitaly Klyachko, Todd Braver, Camillo Padoa-Schioppa, and Richard Abrams – who contributed their time to review and challenge my work, making it stronger. My family, and especially my mother Eleni Poutakidou, makes another significant part of this list. Because of their encouragement I continue to attempt things more difficult than I would on my own. Because of their expectations I continue to set the bar higher than I would without them. And because of their financial support I have been able to 'play', pursuing those things of interest without compromise.

Many others, not listed here, have made significant contributions that lead to the completion of this work, and although it is not possible to create an exhaustive list, I nevertheless hope this short message adequately conveys my appreciation.

ABSTRACT OF THE DISSERTATION

Neural Mechanisms of Working Memory Cortical Networks

by

Charalampos Papadimitriou

Doctor of Philosophy in Biology and Biomedical Sciences

Neurosciences

Washington University in St. Louis, 2015

Professor Lawrence H. Snyder, Chair

This dissertation is aimed at understanding the cortical networks that maintain working memory information. By leveraging patterns of information degradation in spatial working memory encoding we reveal new neural mechanisms that support working memory function and challenge existing models of working memory circuits.

First we examine how interference from previous memoranda influences memory of a currently remembered location. We find that memory for a currently remembered location is biased toward the previously memorized location. This interference is graded, not all-or-none. Interference is strongest when the previous and current targets are close and activate overlapping populations of neurons. Contrary to the attractive behavioral bias, the neural representation of a currently remembered location in the frontal eye fields appears to be biased *away* from the previous target location, not toward it. We reconcile this discrepancy by proposing a model in which receptive fields of memory cells converge toward memorized locations. This reallocation of neural

resources at task-relevant parts of space reduces overall error in the memory network but introduces systematic behavioral biases toward prior memoranda.

We also find that attractive behavioral bias asymptotically increases as a function of the memory period length. Critically, the increase in bias depends only on the *current* trial's memory period. That is, the effect of the previous target progressively increases in the current trial after that target's memory has become irrelevant. We modeled this finding using a two-store model with a transient but unbiased visual sensory store and a sustained store with constant bias. Initially behavior is driven by the veridical visual sensory store and is therefore unbiased. As the visual sensory store decays in the current trial, behavioral responses are increasingly driven by the sustained but biased store, leading to an asymptotic increase of behavioral bias with increasing memory period length.

Finally, we look at how memory activity is encoded over long (15 second) memory periods. Memory cells tend to turn on early in the memory period and stay active for a fixed amount of time. Most memory cells shut off prior to the end of the memory period. Within each cell, offset times are repeatable from one trial to the next. Across cells, offset times are broadly distributed throughout the entire memory period. Once a cell shuts off, it remains off for the rest of the memory period. On the one hand, these findings challenge the leading model for working memory, the attractor network framework, which predicts a single homogenous time course from all cells. On the other hand, the findings also show that the patterns of activity seen in memory circuits are much more structured than the heterogeneous patterns suggested by the leading competitors to the attractor models. Our findings are not predicted by current models

of working memory circuits and indicate that new network models need to be developed.

Chapter 1

Introduction

1.1 Preface

Working memory is the ability to temporarily maintain and transform information. Many cognitive tasks and high order cognitive functions rely on working memory (Engle and Kane, 2004; Miyake and Shah, 1999; Raghubar et al., 2010; Unterrainer and Owen, 2006). Functions such as planning a future action (Altgassen et al., 2007; Cohen, 1996; Unterrainer and Owen, 2006), comparing evidence to make a decision (Payne and Bettman, 1992; Payne et al., 1990), or language processing and comprehension (Daneman and Carpenter, 1980; Daneman and Merikle, 1996; Just and Carpenter, 1992) require information to be stored and transformed in goal-oriented ways. Without working memory, many high order cognitive functions such as these would not be possible, and behavior would instead be driven only by immediately present stimuli.

Spatial working memory is a type of working memory for remembering locations in space. Spatial working memory is necessary for motor tasks that involve movements toward or relative to locations of stimuli that are no longer visible. It is an important area of study because impairments in spatial working memory are major factors of many disorders such as Alzheimer's disease and schizophrenia (Green et al., 2000; Park and Holzman, 1992). Spatial working memory is a great model system for studying working memory function. Unlike many other types of working memory with categorical memoranda – such as remembering particular items or words – spatial working memory uses continuous memoranda (locations in space). Continuous memoranda allow working memory circuits to be probed more finely than categorical memoranda. Animals can be easily trained to perform spatial working memory tasks, allowing us to directly interrogate neural circuits with electrophysiology and compare these direct measures with human fMRI.

The work presented in this dissertation is aimed at understanding the neural substrate that supports working memory function in non-human primates. Information stored in spatial working memory is not maintained indefinitely, but instead decays. This decay is reflected both in behavioral responses and in spatial working memory neural correlates. By examining behavioral responses and their neural correlates in a spatial working memory task, we can determine how information that was previously (but no longer) relevant interferes with the formation, storage, or retrieval of information that is currently relevant. We also ask how memory is maintained over long (15 second) time

periods. The patterns we observe reveal important properties of the neural circuits that support spatial working memory.

1.2 The role of prefrontal cortex in working memory

Lesion studies in prefrontal cortex (PFC) produce significant working memory deficits, implicating the PFC as an important structure for working memory maintenance and function (e.g. Funahashi et al., 1993a; Jacobsen, 1936s; Milner, 1963). Neurophysiology studies in monkeys also implicate the PFC as a key region involved in maintenance of spatial information. Dorsolateral PFC (dlPFC) and frontal eye fields (FEF) show a sustained increase in firing rates during spatial working memory tasks (Bruce and Goldberg, 1985; Chafee and Goldman-Rakic, 1998; Constantinidis et al., 2001; Ferrera et al., 1999; Funahashi et al., 1989, 1993b; Fuster and Alexander, 1971; Kojima and Goldman-Rakic, 1982; di Pellegrino and Wise, 1993; Sommer and Wurtz, 2001; Takeda and Funahashi, 2002, 2004; Umeno and Goldberg, 2001). FEF is also involved in transforming visual signals into saccadic commands (Bruce and Goldberg, 1985; Schall, 1991; Sommer and Wurtz, 2000). During memory tasks that allow for saccade plan generation early in the memory period, fMRI signals in human FEF show increased coherence with a network of oculomotor areas (supplementary eye fields, dorsal anterior cingulate) involved in maintaining saccade goals. In contrast, during memory tasks that prevent saccade planning until late in the trial, fMRI signals in FEF instead show increased coherence with a different network of areas (dlPFC, superior frontal sulcus, posterior parietal cortex) thought to be involved in sustaining covert attention at a

particular spatial location (Corbetta et al., 2002; Curtis et al., 2005). These results suggest that dlPFC and FEF play an important role in maintaining the perceived position of a stimulus, and that FEF is involved in transforming that information into a saccade plan that can be maintained over time.

1.3 Working memory neural circuits

A number of models for working memory have been proposed (e.g. Baddeley, 2012; Durstewitz et al., 2000; Miyake and Shah, 1999; Mongillo et al., 2008). To support memory function, neural circuits need to have properties that allow for the retention of information after stimulus removal. One way to store information about a stimulus is through the short-term modification of synapses (e.g. Mongillo et al., 2008). For example, strengthening of synapses between neurons responsive to a stimulus location may allow for maintaining stimulus information with minimal firing rate changes during the memory period. Even without an ongoing or sustained change in firing rate during the memory period, synaptic modification allows for stimulus information to be recovered through broad activation of the memory circuit at the end of the memory period.

On the single cell level, models for maintaining information may employ cellular bistability as a mechanism for maintenance (Lisman et al., 1998; Marder and Abbott, 1996). In these models, individual cells have two or more stable states. A strong signal due to a stimulus may switch a cell from a low firing 'down' state to a high firing 'up'

state. The cell will then remain in the 'up' state until some other strong external signal causes it to change back to the 'down' state.

Information maintenance may also rely on network-level mechanisms that arise from the combined dynamics of many cells. Network architectures that support memory maintenance typically come in two flavors: feed-forward and recurrent networks. Feed-forward mechanisms such as synfire chains (see Abeles, 1991) work by passing information from one population of cells to the next. Information can persist in these networks for as long as it takes for activity to propagate from the first neural population in the chain to the last. Unlike feed-forward networks, recurrent networks are constructed with cells that have reciprocal connections to one another (see Hebb, 2005). After a population of cells is initially activated through external input, each cell in the population provides excitatory input through connections that feed back onto that cell population, maintaining the active state. Typically such networks also include inhibitory connections to prevent cell excitation from increasing indefinitely. When inhibitory and excitatory connections in recurrent networks are balanced such that the network dynamics contain stable equilibria of active states, they are referred to as attractor networks (e.g. Amit, 1992; Amit and Brunel, 1997; Brunel, 1996; Compte et al., 2000; Wang, 2009).

A large body of experimental work has focused on directly probing working memory circuits by using electrophysiology techniques to determine the network properties and architectures actually employed by prefrontal circuits to maintain information. Much of the electrophysiology literature on spatial working memory is based on tasks with short

memory periods, often between 1 and 3 seconds. Elevated neuronal firing rates observed in these tasks are sustained without appreciable decay for the entirety of the memory period. These results support neural attractor networks. The recurrently connected continuous attractor network stands out as the premier framework for most computational studies of spatial working memory. Continuous attractor networks have been widely successful in modeling phenomena related to spatial working (e.g. Compte et al., 2000; Wang, 2009; Wimmer et al., 2014; see also Durstewitz et al., 2000). Individual excitatory nodes within these networks are tuned to particular portions of the visual field. When arranged topographically, these nodes encode memory locations with a spatially localized bump of elevated activity. This bump is a stable attractor state in the system dynamics and can therefore persist indefinitely over time. Stability is achieved through a fully connected recurrent connectivity structure with strong excitatory connections between cells representing parts of visual space that are close together and weaker excitatory connections between cells representing parts of visual space that are farther apart. The excitation is balanced by inhibitory circuits with broad connections that keep the excitatory activity in check and prevent the bump from spreading across the entire network (Figure 1.1). A number of studies have found anatomical (Kritzer and Goldman-Rakic, 1995; Levitt et al., 1993) and physiological (Gonzalez-Burgos, 2000) evidence of the local recurrent connections necessary for recurrent attractor networks. Neural activity dynamics of attractor network cells are generally simple. Cells quickly increase firing rate to a steady-state asymptote and maintain that state until the end of the memory period. The time course of activity increase is the same for all cells.

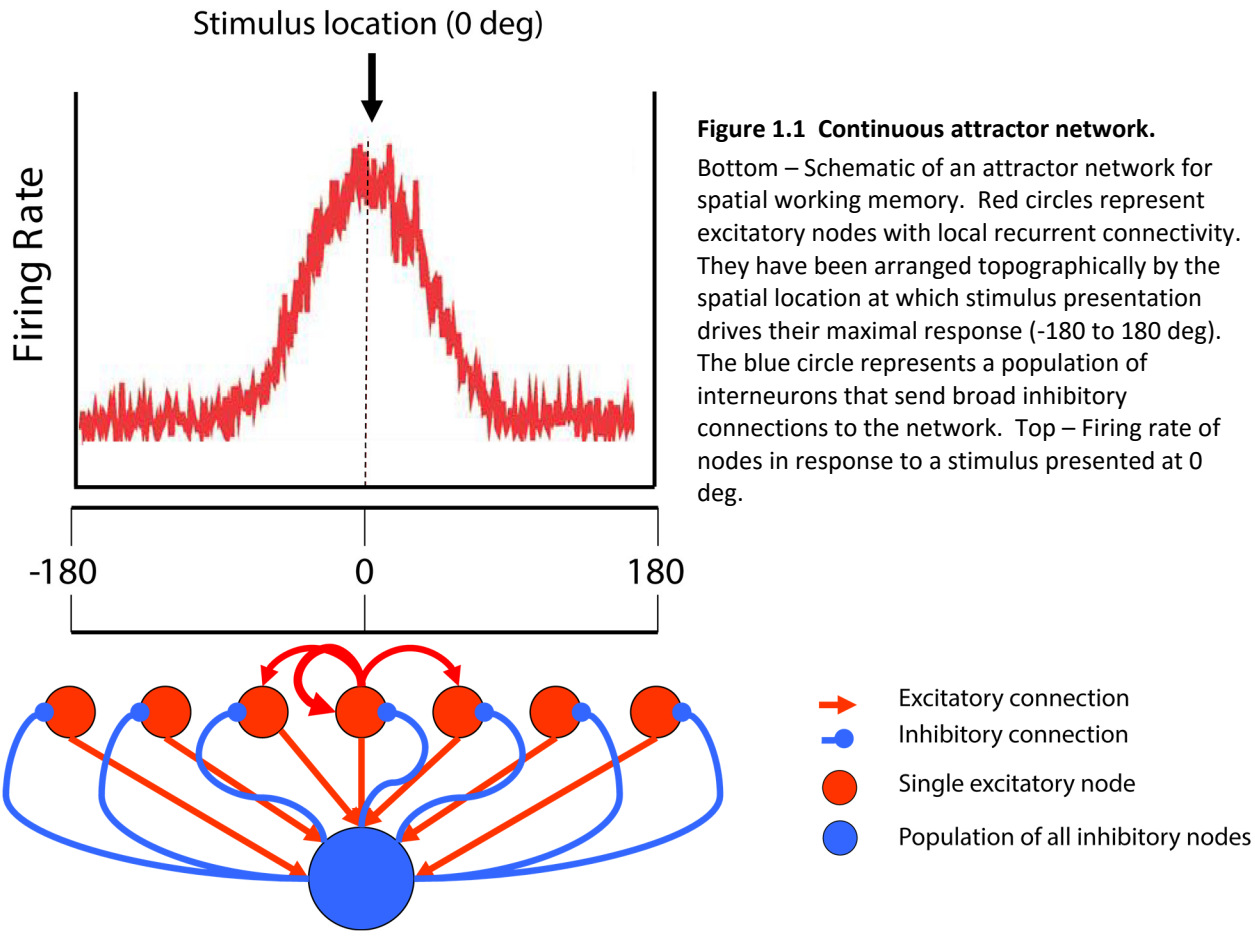


Figure 1.1 Continuous attractor network.
Bottom – Schematic of an attractor network for spatial working memory. Red circles represent excitatory nodes with local recurrent connectivity. They have been arranged topographically by the spatial location at which stimulus presentation drives their maximal response (-180 to 180 deg). The blue circle represents a population of interneurons that send broad inhibitory connections to the network. Top – Firing rate of nodes in response to a stimulus presented at 0 deg.

Other studies looking at working memory responses paint a different picture than the well-behaved, sustained responses of attractors that are homogeneous from one cell to the next. They instead show that recorded memory cells appear to turn on and off multiple times during the memory period and the time course of activity is heterogeneous across cells (Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010). At any given point during the memory period, individual cells may or may not encode the memory location, but population activity across all cells still provides robust memory encoding.

In this dissertation we examine degradation of spatial working memory information in behavior and in prefrontal memory circuits. The patterns of degradation we observe reveal properties of working memory cortical circuits that allow us to distinguish between working memory models.

1.4 How does memory degrade?

Although working memory can maintain information that information is not maintained indefinitely. Generally there are two main qualitative reasons for working memory decay. Firstly, information may compete for limited working memory resources. That is, memory circuits may not have the capacity to encode and maintain all of the information present in a set of stimuli. In the context where task-irrelevant information and task-relevant information compete, the irrelevant information is said to interfere with memory of the relevant information. Secondly, memory may also degrade over time due to an imperfect memory system. That is, random noise due to imperfections in the information retention mechanisms may increasingly contaminate a memory representation over time, degrading stored information. In the following subsections we discuss each type of memory degradation in detail.

1.4.1 Interference

Working memory may be degraded due to interference or distraction from task-irrelevant information presented before the memoranda is presented or during the

memory period. Irrelevant information may cause interference by overwriting relevant memory information, adding non-systematic noise, or adding a systematic bias. Distinguishing between these possibilities based on categorical memory tasks that measure only the proportion of correct responses can be difficult. Behavioral studies on memory interference have focused on the spatial working memory system because it provides continuous measures for both stimuli and responses, and therefore is well-suited to studying partial memory degradation (Chumbley et al., 2008; Macoveanu et al., 2007). (With categorical memory, there is only a binary readout – either memoranda are remembered or not remembered --- and so partial degradation cannot be tested.) These studies focus on interference within a trial, asking how flashing a task-irrelevant distractor at a location in space can interfere with the memory of the spatial location of a target stimulus when both target and distractor are presented in the same trial. They show that the presentation of a distractor does not overwriting the memory location completely but instead leads to an intermediate compromise between the memory location and the distractor location. That is, the memory of the target stimulus' location is biased toward the distractor location.

Continuous attractor circuits as described in Compte et al. (2000) show a similar pattern in simulations involving distractors. When two competing representations (e.g. location of a target stimulus and location of a distractor) in a line attractor are far apart in the encoded space, strong global inhibitory connections quash the weaker of the two representations. In that case, the network settles on encoding either the target location or the distractor location. However, when the two competing representations are close together, neurons representing the intermediate space receive high local excitatory

input from nearby neurons activated by either of the two stimuli. As a result, the network settles to encode an intermediate location between the distractor and target, consistent with the results in behavioral studies.

In addition to within-trial interference from distractors, information previously held in memory has been shown to interfere with current memory performance, a phenomenon known as proactive interference. A classic example of this effect from human literature is described by Underwood (1957), who performed experiments requiring memorization of nonsense syllable lists and found that the more sessions his subjects participated in the worse they performed. He attributed this to interference from syllables memorized in prior sessions. Animal studies have identified proactive effects that degrade memory performance during categorical memory tasks (Dunnett and Martel, 1990; Edhouse and White, 1988; Jarvik et al., 1969; Moise, 1976; see also Jonides and Nee, 2006). In these studies, memorized information from the immediately preceding trial causes interference. These studies typically employ delayed-match-to-sample or recent-probes tasks with a small number of (often just two) categorical responses, such that errors in these tasks are all-or-none. As a result, these studies do not address whether and how proactive interference may lead to partial memory degradation and what form that degradation of information may take in behavioral and neural responses. Whether information degradation due to proactive interference is similar to that caused by distractors is an open question. Furthermore, the neural mechanisms that lead to partial interference have not been identified.

In Chapter 2 of the dissertation we use a memory-guided saccade task to determine how interference from previously relevant information can cause degradation of spatial working memory. We find that behavioral responses are biased toward the previously remembered locations. While the spatial profile of the bias is consistent with predictions from an attractor network, a single memory attractor does not account for the temporal profile of the bias. Our findings suggest that working memory circuits may rely on two memory stores, a sustained store with biased representations due to proactive interference and a quickly decaying but unbiased visual sensory store (e.g. an Iconic store).

Iconic and intermediate memory stores

Classic human psychology studies have identified iconic memory, a short-term store that retains the visuo-spatial characteristics of stimuli for a brief interval after stimulus offset (e.g. Sperling, 1960). Early researchers observed that when a brief visual display is presented people believe that they see more information than they are able to report. That is, they sensed that some of the information that is available to them after the display is removed fades before they are able to report it (e.g. Cattell, 1883; Erdmann and Dodge, 1898). George Sperling used partial report experiments to assess this phenomenon experimentally. In these experiments subjects were presented with a visual display with multiple rows of letters for 50 ms. In the whole-report condition subjects were asked to report the identity and spatial position of as many letters as possible. In the partial-report condition after the visual display offset subjects were cued on which row of the visual display they were to report. Subjects were able to report

only 38% of letters in the whole-report condition but 76% of letters in the partial-report condition. Because subjects did not know which row would be cued prior to the visual display offset in the partial-report condition, they had to memorize the entire array of letters, and Sperling argued that the amount of remembered information in this condition was representative of the information available to subjects immediately after the visual display offset. That is, immediately after visual display offset subjects remembered at least 76% of the information of the entire visual display (including all rows) but that information quickly faded to only 38% prior to subjects being able to report it.

In a series of follow-up experiments researchers found that iconic memory has virtually unlimited capacity (Dick, 1974), decays in under one second (Averbach and Coriell, 1961), depends on exposure parameters such as luminance (Averbach and Sperling, 1961), and does not survive eye movements (Irwin, 1992; Irwin and Andrews, 1996). Iconic memory was also found to be susceptible to masking. That is, information in iconic memory may be degraded or wiped out by presenting irrelevant stimuli in the visual field of the memory stimulus before or after the memory stimulus presentation (Averbach and Coriell, 1961; Sperling, 1963). These properties suggest that iconic memory is a low-level sensory memory.

Other visual short-term memory stores intermediate to iconic memory and working memory have been identified in behavioral (Pinto et al., 2013; Sligte et al., 2008; Vandenbroucke et al., 2014; see also Griffin and Nobre, 2003; Makovski and Jiang, 2008; Makovski et al., 2008) and neurophysiological studies (Sligte et al., 2009; see

also Bisley et al., 2004; Pasternak and Greenlee, 2005). These stores have longer time constants than the iconic store, with some studies showing information retention for up to 4 seconds. In Chapter 2 we show how a short-term store such as these interacts with a longer sustained memory store to drive behavioral responses.

Receptive field changes in prefrontal cortex

The results presented in Chapter 2 lead us to hypothesize that the bias of the sustained memory store is driven by attractor network dynamics. In Chapter 3 we turn our attention to neural correlates of this bias. To test our hypothesis we record from frontal eye fields, an area known to show sustained neural activity in memory tasks and to drive saccadic responses. We find that neural correlates of proactive interference do not fit attractor network predictions. Instead our findings suggest that changes in cells' receptive fields lead to the behavioral bias we observe.

Receptive field changes have been found to occur during tasks requiring execution of saccades and spatial attention. In these tasks neurons in some visual areas and areas involved in working memory appear to temporarily shift their receptive fields (Colby and Goldberg, 1992; Connor et al., 1997; Sommer and Wurtz, 2006; Tolias et al., 2001; Walker et al., 1995; Zirnsak et al., 2014). Initial findings suggested that receptive fields shift in the direction of the saccade vector (Colby and Goldberg, 1992; Sommer and Wurtz, 2006; Walker et al., 1995). That is, neurons predictively remap to respond to the part of space they will encode after the saccade is completed. These findings have given rise to the interpretation that receptive field changes may play a role in stabilizing visual space during eye movements.

This interpretation is challenged by findings like those in Zirnsak et al. (2014). In this study receptive field changes in FEF have been mapped more rigorously than in previous studies. The authors find that rather than shifting in the direction of the saccade vector, receptive fields converge toward the vector's endpoint. That is, the visual field around the saccade endpoint or attended location is represented by an increased number of neurons. This could serve as a way to devote more resources to that part of visual space, and may be a correlate of spatial attention. Because spatial attention and spatial working memory are closely related, receptive field changes may also occur in spatial memory tasks. Merrikhi et al. (2014) have shown preliminary evidence of this in area V4.

Receptive field shifts have been linked to mislocalizations of stimuli around the time of a saccade (Hamker et al., 2008; Ross et al., 1997, 2001). In spatial tasks, behavioral responses and perception of stimuli are mislocalized toward the saccadic endpoint or attended location. This attractive mislocalization is opposite to what may be expected from the observed pattern of receptive field shifts. That is, shifting of a neuron's receptive field center in a particular direction relative to the visual space is equivalent to shifting of the visual space in the opposite direction relative to the neuron's receptive field center. A neuron with a receptive field center that has been shifted in one direction will respond to a visual stimulus as if that stimulus has shifted in the opposite direction. Therefore behavioral responses should instead be mislocalized away from the location of receptive field convergence.

Because receptive field shifts in prefrontal cortex have been found in tasks that require saccadic execution, spatial attention, and possibly spatial memory, the encoding and readout of memory information may be modified by these effects. Thus, receptive field changes may be reflected in behavioral responses and prefrontal neural activity in working memory tasks.

In Chapter 3 we show that it is receptive field changes that drive interference effects from the previous trial, not attractor dynamics. Contrary to our predictions, the neural correlates of behavioral bias due to proactive interference we identify in frontal eye fields are not consistent with predictions from attractor networks. Instead, neural proactive effects can best be modeled using persistent receptive field convergence toward memory targets. We show that convergence of receptive fields improves performance of a model memory network but leads to systematic effects of previous memoranda.

1.4.2 Decay over time

In addition to degradation from interference, information stored in working memory circuits may also be degraded due to imperfections in the memory retention mechanism. One way this can occur is when sustained firing rates in memory neurons decay over time. Although a number of studies show that neural responses are generally sustained without much decay for the full duration of memory periods tested, due to the relatively short memory periods (1 to 3 seconds) typically tested these studies may not be sensitive to a slow decay rate.

Information degradation can also occur without decay in neural activity. For example in a continuous attractor network that indefinitely sustains elevated memory activity information degrades due to 'random drift'. The bump of elevated activity initially encoding a memory location is allowed to drift from encoding one location of space to encoding another. Neural noise from external inputs to the circuit can cause the bump to drift randomly, and over time the total amount of possible drift increases. Therefore, as the time between stimulus presentation and behavioral response increases, error in the encoded memory also increases. Evidence of random drift has been shown by Wimmer et al. (2014) in PFC memory representations during a spatial working memory task.

In Chapter 4 we look at spatial working memory decay due to the passage of time. We use long memory periods of 15 seconds to determine whether decay predictions of attractor network models are consistent with decay in cortical memory networks. We find that memory cell responses are not indefinitely sustained nor do they have a common rate of decay. Our results support a model with great heterogeneity in the decay rate of each cell, but repeatable trial-to-trial decay within each cell. Memory trace acquisition and degradation is highly structured with cells generally turning on early in the memory period, maintaining the memory signal for a fixed period, and then shutting off for the remainder of the trial.

References

Abeles, M. (1991). Corticonics: Neural circuits of the cerebral cortex (Cambridge: Cambridge University Press).

Altgassen, M., Phillips, L., Kopp, U., and Kliegel, M. (2007). Role of working memory components in planning performance of individuals with Parkinson's disease. Neuropsychologia *45*, 2393–2397.

Amit, D. (1992). Modeling brain function: The world of attractor neural networks (Cambridge: Cambridge University Press).

Amit, D.J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex *7*, 237–252.

Averbach, E., and Coriell, A.S. (1961). Short-term memory in vision. Bell Syst. Tech. J. *40*, 309–328.

Averbach, E., and Sperling, G. (1961). Short-Term Storage of information in Vision. Inf. Theory 196–211.

Baddeley, A. (2012). Working Memory : Theories, Models, and Controversies.

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

Bisley, J.W., Zaksas, D., Droll, J.A., and Pasternak, T. (2004). Activity of neurons in cortical area MT during a memory for motion task. J. Neurophysiol. *91*, 286–300.

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

Bruce, C.J., and Goldberg, M.E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. *53*, 603–635.

Brunel, N. (1996). Hebbian learning of context in recurrent neural networks. Neural Comput. *8*, 1677–1710.

Cattell, J. (1883). Uber die Tragheit der Netzhaut und des Sehcentrums. Philos. Stud. *3*, 94–127.

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

Chumbley, J.R., Dolan, R.J., and Friston, K.J. (2008). Attractor models of working memory and their modulation by reward. Biol. Cybern. *98*, 11–18.

Cohen, G. (1996). Memory in the Real World (Psychology Press).

Colby, C., and Goldberg, M. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. Science *255*, 90–92.

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

Connor, C.E., Preddie, D.C., Gallant, J.L., and Van Essen, D.C. (1997). Spatial attention effects in macaque area V4. J. Neurosci. *17*, 3201–3214.

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

Corbetta, M., Kincade, J.M., and Shulman, G.L. (2002). Neural systems for visual orienting and their relationships to spatial working memory. J. Cogn. Neurosci. *14*, 508–523.

Curtis, C.E., Sun, F.T., Miller, L.M., and D'Esposito, M. (2005). Coherence between fMRI time-series distinguishes two spatial working memory networks. Neuroimage *26*, 177–183.

Daneman, M., and Carpenter, P. a. (1980). Individual differences in working memory and reading. J. Verbal Learning Verbal Behav. *19*, 450–466.

Daneman, M., and Merikle, P.M. (1996). Working memory and language comprehension: A meta-analysis. Psychon. Bull. Rev. *3*, 422–433.

Dick, A.O. (1974). Iconic memory and its relation to perceptual processing and other memory mechanisms. Percept. Psychophys. *16*, 575–596.

Dunnett, S.B., and Martel, F.L. (1990). Proactive interference effects on short-term memory in rats: I. Basic parameters and drug effects. Behav. Neurosci. *104*, 655–665.

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

Edhouse, W. V., and White, K.G. (1988). Cumulative proactive interference in animal memory. Anim. Learn. Behav. *16*, 461–467.

Engle, R., and Kane, M. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. Psychol. Learn. Motiv.

Erdmann, B., and Dodge, R. (1898). Psychologische Untersuchengen uber das Lesen auf experimenteller Grundlage. (Niemeyer: Halle).

Ferrera, V.P., Cohen, J.K., and Lee, B.B. (1999). Activity of prefrontal neurons during location and color delayed matching tasks. Neuroreport *10*, 1315–1322.

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

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

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

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

Gonzalez-Burgos, G. (2000). Horizontal Synaptic Connections in Monkey Prefrontal Cortex: An In Vitro Electrophysiological Study. Cereb. Cortex *10*, 82–92.

Green, M., Kern, R., Braff, D., and Mintz, J. (2000). Neurocognitive deficits and functional outcome in schizophrenia. Schizophr. Bull.

Griffin, I.C., and Nobre, A.C. (2003). Orienting attention to locations in internal representations. J. Cogn. Neurosci. *15*, 1176–1194.

Hamker, F.H., Zirnsak, M., Calow, D., and Lappe, M. (2008). The Peri-Saccadic Perception of Objects and Space. PLoS Comput. Biol. *4*, e31.

Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

Hebb, D. (2005). The organization of behavior: A neuropsychological theory (New York: Wiley).

Irwin, D.E. (1992). Memory for position and identity across eye movements. J. Exp. Psychol. Learn. Mem. Cogn. *18*, 307–317.

Irwin, D.E., and Andrews, R. V (1996). Integration and accumulation of information across saccadic eye movements. Atten. Perform. Inf. Integr. Percept. Commun. *16*, 125–155.

Jacobsen, C.F. (1936). Studies of cerebral function in primates, 1. The function of the frontal association areas in monkeys. J. Compo Psychol.

Jarvik, M.E., Goldfarb, T.L., and Carley, J.L. (1969). Influence of interference on delayed matching in monkeys. J. Exp. Psychol. *81*, 1.

Jonides, J., and Nee, D.E. (2006). Brain mechanisms of proactive interference in working memory. Neuroscience *139*, 181–193.

Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C.D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. J. Neurosci. *30*, 916–929.

Just, M. a, and Carpenter, P. a (1992). A capacity theory of comprehension: individual differences in working memory. Psychol. Rev. *99*, 122–149.

Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–50.

Kritzer, M.F., and Goldman-Rakic, P.S. (1995). Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. J. Comp. Neurol. *359*, 131–143.

Levitt, J.B., Lewis, D.A., Yoshioka, T., and Lund, J.S. (1993). Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46). J. Comp. Neurol. *338*, 360–376.

Lisman, J., Fellous, J., and Wang, X. (1998). A role for NMDA-receptor channels in working memory. Nat. Neurosci. 1, 273-275.

Macoveanu, J., Klingberg, T., and Tegnér, J. (2007). Neuronal firing rates account for distractor effects on mnemonic accuracy in a visuo-spatial working memory task. Biol. Cybern. *96*, 407–419.

Makovski, T., and Jiang, Y. V (2008). Proactive interference from items previously stored in visual working memory. Mem. Cognit. *36*, 43–52.

Makovski, T., Sussman, R., and Jiang, Y. V (2008). Orienting attention in visual working memory reduces interference from memory probes. J. Exp. Psychol. Learn. Mem. Cogn. *34*, 369–380.

Marder, E., and Abbott, L. (1996). Memory from the dynamics of intrinsic membrane currents. PNAS 93, 13481-13486

Merrikhi, Y., Parsa, M., Albarran, E., and Noudoost, B. (2014). Maintenance of spatial information gates the processing of incoming visual information in area V4. In Society for Neuroscience.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. Arch. Neurol.

Miyake, A., and Shah, P. (1999). Models of working memory: Mechanisms of active maintenance and executive control.

Moise, S.L. (1976). Proactive effects of stimuli, delays, and response position during delayed matching from sample. Anim. Learn. Behav. *4*, 37–40.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Park, S., and Holzman, P. (1992). Schizophrenics show spatial working memory deficits. Arch. Gen. Psychiatry.

Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.

Payne, J.W., and Bettman, J.R. (1992). Behavioral Decision Research: A Constructive Processing Perspective. Annu. Rev. Psychol. *43*, 87–131.

Payne, J.W., Bettman, J.R., and Johnson, E.J. (1990). The adaptive decision maker: Effort and accuracy in choice. In Insights in Decision Making: A Tribute to Hillel J. ..., R. Hogarth, ed. (University of Chicago Press), pp. 129–153.

Di Pellegrino, G., and Wise, S.P. (1993). Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J. Neurosci. *13*, 1227–1243.

Pinto, Y., Sligte, I.G., Shapiro, K.L., and Lamme, V. a (2013). Fragile visual short-term memory is an object-based and location-specific store. Psychon. Bull. Rev. *20*, 732–739.

Raghubar, K.P., Barnes, M. a., and Hecht, S. a. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. Learn. Individ. Differ. *20*, 110–122.

Ross, J., Morrone, M.C., and Burr, D.C. (1997). Compression of visual space before saccades. Nature *386*, 598–601.

Ross, J., Morrone, M.C., Goldberg, M.E., and Burr, D.C. (2001). Changes in visual perception at the time of saccades. Trends Neurosci. *24*, 113–121.

Schall, J.D. (1991). Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields. J Neurophysiol *66*, 559–579.

Sligte, I.G., Scholte, H.S., and Lamme, V. a F. (2008). Are there multiple visual short-term memory stores? PLoS One *3*, 2–10.

Sligte, I.G., Scholte, H.S., and Lamme, V. a F. (2009). V4 activity predicts the strength of visual short-term memory representations. J. Neurosci. *29*, 7432–7438.

Sommer, M. a, and Wurtz, R.H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. Nature *444*, 374–377.

Sommer, M.A., and Wurtz, R.H. (2000). Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. J Neurophysiol *83*, 1979–2001.

Sommer, M.A., and Wurtz, R.H. (2001). Frontal eye field sends delay activity related to movement, memory, and vision to the superior colliculus. J Neurophysiol *85*, 1673–1685.

Sperling, G. (1960). The information available in brief visual presentations. Psychol. Monogr. Gen. Appl.

Sperling, G. (1963). A Model for Visual Memory Tasks. Hum. Factors J. Hum. Factors Ergon. Soc. *5*, 19–31.

Takeda, K., and Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. J Neurophysiol *87*, 567–588.

Takeda, K., and Funahashi, S. (2004). Population vector analysis of primate prefrontal activity during spatial working memory. Cereb Cortex *14*, 1328–1339.

Tolias, A.S., Moore, T., Smirnakis, S.M., Tehovnik, E.J., Siapas, A.G., and Schiller, P.H. (2001). Eye Movements Modulate Visual Receptive Fields of V4 Neurons. Neuron *29*, 757–767.

Umeno, M.M., and Goldberg, M.E. (2001). Spatial processing in the monkey frontal eye field. II. Mem. Responses. J Neurophysiol *86*, 2344–2352.

Underwood, B.J. (1957). Interference and forgetting. Psychol. Rev. *64*, 49–60.

Unterrainer, J.M., and Owen, A.M. (2006). Planning and problem solving: From neuropsychology to functional neuroimaging. J. Physiol. Paris *99*, 308–317.

Vandenbroucke, A.R.E., Sligte, I.G., Barrett, A.B., Seth, A.K., Fahrenfort, J.J., and Lamme, V. a F. (2014). Accurate metacognition for visual sensory memory representations. Psychol. Sci. *25*, 861–873.

Walker, M.F., Fitzgibbon, E.J., and Goldberg, M.E. (1995). Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye movements. J Neurophysiol *73*, 1988–2003.

Wang, X.-J. (2009). Attractor Network Models. In Encyclopedia of Neuroscience, L.R. Squire, ed. (Oxford: Oxford: Academic Press), pp. 667–679.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. *17*, 431–439.

Zirnsak, M., Steinmetz, N.A., Noudoost, B., Xu, K.Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. Nature *507*, 504–507.

Chapter 2

Ghosts in the Machine: Memory Interference from the Previous Trial

2.1 Abstract

Previous memoranda can interfere with the memorization or storage of new information, a concept known as proactive interference. Studies of proactive interference typically use categorical memoranda and match-to-sample tasks with categorical measures such as the proportion of correct to incorrect responses. In this study we instead train five macaques in a spatial memory task with continuous memoranda and responses, allowing us to more finely probe working memory circuits. We first ask whether the memoranda from the previous trial result in proactive interference in an oculomotor delayed response task. We then characterize the spatial and temporal profile of this interference, and ask whether this profile can be predicted by an attractor network model of working memory.

We find that memory in the current trial shows a bias toward the location of the memorandum of the previous trial. The magnitude of this bias increases with the duration of the memory period within which it is measured. Our simulations using standard attractor network models of working memory show that these models easily replicate the spatial profile of the bias. However, unlike the behavioral findings, these attractor models show an increase in bias with the duration of the *previous* rather than the current memory period. To model a bias that increases with current trial duration we posit two separate memory stores, a rapidly-decaying visual store that resists proactive interference effects and a sustained memory store that is susceptible to proactive interference.

2.2 Introduction

Working memory involves maintenance of goal-related information in an active state over a short period of time. It has been implicated in higher cognitive functions (Miyake and Shah 1999, Engle and Kane 2004, Raghubar et al. 2010, Unterrainer and Owen 2006) such as planning (Cohen 1996, Altgassen et al. 2007, Unterrainer and Owen 2006), decision-making (Payne et al. 1990, Payne and Bettman 1992) and language comprehension (Daneman and Carpenter 1980, Just and Carpenter 1992, Daneman and Merikle 1996), and hence an understanding of the circuitry of working memory is critical for understanding higher cognition. Spatial working memory provides an excellent model system for this purpose. The memoranda – locations in space – are continuous and well-defined, as are the responses that provide a read out for these

memoranda. Animals can be easily trained to perform spatial working memory tasks, and neural circuits can then be directly interrogated using various electrophysiological techniques. A number of models for working memory have been proposed (Durstewitz et al. 2000, Miyake and Shah 1999, Baddeley 2012). Here we present findings that can constrain these models and illuminate the mechanisms by which memory is degraded.

Working memory may be degraded due to interference or distraction from task-irrelevant events that occur before or during the memory period. Prior computational and behavioral studies on memory interference have focused on the spatial working memory system because it provides continuous measures for both stimuli and responses, and is well-suited to studying partial memory degradation (Compte et al. 2000, Macoveanu et al. 2007, Chumbley et al. 2008). These studies focus on interference within a trial, asking how irrelevant information and distractor stimuli interfere with the memory information of a target stimulus when both are presented in the same trial.

The recurrently connected continuous attractor network (e.g. Wang 2009, Compte et al. 2000) is the premier framework for most computational studies of working memory. Continuous attractor networks have been widely successful in modeling phenomena related to working memory. Individual excitatory nodes within these networks are tuned to particular portions of the visual field. When arranged topographically, these nodes encode memory locations with a spatially localized bump of elevated activity. This bump is a stable attractor state in the system dynamics and can therefore persist indefinitely over time. Stability is achieved through a fully connected recurrent connectivity structure with strong excitatory connections between cells representing

parts of visual space that are close together and weaker connections between cells representing parts of visual space that are farther apart. The excitation is balanced by inhibitory circuits with broad connections that keep the excitatory activity in check and prevent the bump from spreading across the entire network. Continuous attractor models have been used to explore within-trial interference from distractor stimuli. The connectivity structure and bump attractor state in these models predict distractor effects with a very specific spatial profile. The models predict an attractive bias of memory toward the distractor location, with strong bias when remembered and distractor locations are close together and weaker bias when they are progressively farther apart (Compte at al. 2000). The spatial profile of within-trial interference arises because when two bumps of activity – one representing the true remembered location and one representing the distractor response – are close together, nodes between the two bumps receive the most excitatory input, and as a result the bumps merge to form a single bump at an intermediate location. When the two peaks are far apart, intermediate nodes are not excited and the global inhibition quashes the weaker bump (usually the distractor bump) before it has a chance to distort the bump representing the correct memory location. The predicted spatial pattern of within-trial interference has been replicated in behavioral studies (Macoveanu et al. 2007, Chumbley et al. 2008).

In addition to within-trial interference from distractors, information previously held in memory has been shown to interfere with current memory performance, a phenomenon known as proactive interference. A classic example of this effect from human literature is described by Underwood (1957), who performed experiments requiring memorization of nonsense syllable lists and found that the more sessions his subjects participated in

the worse they performed due to interference from syllables memorized in prior sessions. Animal studies have identified proactive effects that degrade memory performance during categorical memory tasks (Jarvik et al. 1969, Moise 1976, Edhouse and White 1988, Dunnett & Martel 1990; see also Jonides & Nee 2006). In these studies, memorized information from the immediately preceding trial causes interference. These studies typically employ delayed-match-to-sample or recent-probes tasks with a small number (often just two) categorical responses, and errors in these tasks are all-or-none.

In this study we looked at memory errors due to sequential, between-trial memory interference – whether and how what is remembered on one trial interferes with memory on the next trial – using a continuous oculomotor delayed response task in which subjects maintain a target location in working memory and then indicate the contents of their memory with a rapid eye movement (saccade). We found that saccades were biased toward the location of the target of the previous trial. In contrast with most previous studies, the continuous nature of our task reveals that between-trial proactive interference is graded as a function of the relative distance between the current and previous memory locations. The specific spatial profile of graded interference we identify strongly resembles within-trial errors caused by distractors, indicating that the neural mechanisms involved in between-trial proactive interference and interference due to distractors presented while a memory is already being maintained are the same or strongly related. Attractor network dynamics inherently produce errors with the same spatial structure when multiple memory traces compete within an attractor circuit (e.g. Compte et al. 2000). We therefore propose a specific mechanism for the

generation of proactive interference in working memory circuits, showing that a residual ghost of activity from the previous trial competing with the current target representation in an attractor circuit can reproduce the spatial structure of the proactive bias.

We hypothesized that as the memory trace of the previous trial decays over the course of the delay period, proactive interference would decrease. Surprisingly, proactive interference instead increased asymptotically with delay length. This finding is not readily reproduced by a single sustained store. We model this effect using two stores: a biased sustained memory store (e.g., working memory) and an unbiased transient visual sensory store (e.g., something similar to iconic memory). The sustained store depends on attractor dynamics that can maintain a memory trace for long periods but as a result is susceptible to proactive interference. The visual sensory store can maintain a trace for only a few seconds, but because it is unbiased, it can protect memory representations against bias while it is available.

2.3 Materials and Methods

All experiments were conducted with the approval of the IACUC at Washington University in St. Louis.

Five male macaques were trained on a center-out memory-guided saccade task requiring them to remember peripheral locations distributed on a circle. Target locations were continuously distributed around the circle. Once the macaques became proficient at the task (> 85 % success rate) we recorded the end points of saccades made to the remembered target locations.

Each subject sat in a primate chair in a dark room and was head-fixed securely in a straight-ahead position. Visual stimuli were projected onto a white screen 20 cm from the subject. For subjects C and W, eye-position was recorded with 0.05 degree resolution every 2 ms using a field coil system. For monkeys L, D and R, an ISCAN infrared video eye tracking system was used to record eye-position. Stimuli were controlled by custom software.

Behavioral Task

Each trial began with presentation of a central fixation target and subjects were required to maintain fixation (within 4 degrees) until it disappeared. After the subject acquired fixation a peripheral target was presented for 150 ms at a fixed eccentricity (between 10° and 15° depending on the subject) while the subject continued to fixate. In most experiments the target location was randomly selected in each trial from 360 locations spaced 1° apart. A delay period between 0 to 6 seconds (depending on the experiment and subject) followed target presentation, after which the central fixation target disappeared and the subject was required to make a saccade to the remembered peripheral target location. Saccadic responses within 6 to 10 degrees (depending on the animal) of visual angle of the target were accepted as correct. The peripheral target reappeared 300 ms after the subject's response, and the subject was rewarded for making a corrective saccade to the target and maintaining eye position within 6 degrees of the target for 300ms. An intertrial interval (ITI) period between 2 to 6.5 seconds (depending on the experiment and subject) followed the corrective response. Correct trials were rewarded with delivery of water. Memory period errors occurred when fixation was broken before the central fixation target disappeared. These failure trials

were tallied but excluded from all analyses except analysis of the error distribution. Because memory information may not have been properly encoded in failure trials, success trials following a failure trial were also excluded when considering the effect of previous target location.

Data Analyses

We measured the response error in each trial. As compared to visually-guided saccades, subjects make systematic and variable errors while making saccades to remembered target locations. Due to the systematic errors, saccades to targets in the upper visual field tend to be hypermetric, saccades to targets in the lower visual hemifield tend to be hypometric, and saccades to targets on the horizontal meridian tend to be upward. These systematic errors have been shown to be influenced only by the early part (~400ms to 800ms) of memory delay suggesting that memory processes are not major sources of these errors. However, variable error in memory tasks is influenced by delay period length over several seconds and as a result can provide information about memory decay (White & Sparks 1986, White et al. 1994, Gnadt et al. 1991). We therefore excluded systematic error effects and focused specifically on variable response error in our analyses.

Saccade angular directions were obtained during the 100 ms to 300 ms interval following the first saccade to the target. We calculated saccadic response error as the difference between the saccade direction and target direction in each trial. Systematic error, which was relatively constant over the duration of the experiment, was removed from saccadic error to obtain residual variable response error. When many repetitions of discrete target locations were used, the systematic error was computed as the mean

saccade endpoint for each target location. The corresponding mean was then subtracted from the saccade endpoint on each individual trial to obtain the non-systematic error. For continuous target locations, the systematic error was computed by spatially low-pass filtering the saccade endpoints, expressed as a function of target location, using the MATLAB loess smooth function. The resulting estimate of systematic error was then subtracted from each individual saccade endpoint to obtain the non-systematic error, which we call residual error. For each trial we also calculated the previous trial target direction relative to the current trial target direction by taking the difference between previous and current target directions. We then fit these data to the Gabor function in Equation 1.

$$y = height * e^{-(width*x)^2} * \sin(width * x) \qquad (1)$$

where y represents the residual error on each trial and x represents the relative direction of the previous trial's target. When reporting bias effect sizes we use the peak-to-peak distance of Equation 1 which is equivalent to *0.793*height*.

Continuous attractor network simulations

For attractor network simulations we used a continuous attractor network as described in Compte et al. (2000). The network consisted of 1024 excitatory nodes, each tuned to a location in space, and 256 inhibitory nodes. Each node modeled an excitatory or inhibitory neuron. The network code is available at http://eye-hand.wustl.edu/supplemental/SpatialWMNet_Published_EyeHand.zip and parameters and simulation paradigm can be found in the included parameters.ini file.

We simulated pairs of oculomotor delayed response trials, with each trial including a target presentation and a memory period. Each simulation began with 250ms of spontaneous firing prior to the start of the first trial. A target stimulus was then presented at a spatial location for 250ms in the form of a current injection of 70pA to the 100 model neurons most closely tuned to that spatial location. After the memory period (500ms to 3000ms, depending on the experiment) a stop signal (140pA to all excitatory neurons for 100ms) representing the end of the first trial turned off the sustained activity in the network. After an intertrial interval of 150ms the next trial in the simulation began and a target stimulus with identical amplitude and duration was presented at a different location in the network. The network was read out after the second memory period ended. The population activity of the network was first smoothed across time (50 ms) and spatial location (30 adjacent neurons). A Gaussian was then fit to the time-slice at the end of the second trial's memory period and the memory was read out as the center of the Gaussian.

Two store model of short-term memory

We modeled the temporal aspect of the bias effect using a model with a quickly decaying visual sensory store and a sustained working memory store. The target location estimate in the visual store was assumed to be veridical while the sustained store's estimate was taken to have a Gabor-like spatial profile. Each store's activity was modeled as a decaying exponential. To read out a target location at a given time t, each store's activity at t, was first normalized by the activity during stimulus presentation. The target location was then determined by the equation

$$T_E(t) = T_V w(t) + T_M\big(1 - w(t)\big) \qquad (2)$$

where $T_E(t)$ is the estimated target location, T_V is the true target location held in the visual store, and T_M is the biased target location held in the sustained store and determined by a Gabor function with 8 degrees peak-to-peak based on behavioral data shown in Figure 2.5. The quantity $w(t)$ is the ratio of normalized activity in the visual sensory store $V(t) = V_0 e^{-t/\tau_v}$ and sustained memory store $M(t) = M_0 e^{-t/\tau_m}$ at time t

$$w(t) = \frac{e^{-t/\tau_v}}{e^{-t/\tau_m}} \qquad (3)$$

At time $t = 0$, $w(t) = 1$ and $T_E(t)$ is entirely determined from the content of the visual store and is therefore unbiased. As activity in the visual store decays, $w(t) \to 0$ and $(1 - w(t)) \to 1$, shifting the dependence of $T_E(t)$ to the biased memory of the sustained store.

Based on the large or infinite decay time constants of attractor models and our observations using neural recordings of spatial working memory circuits, we conservatively set the sustained store's decay time constant to be $\tau_m = 15$ seconds; the model results were essentially identical for any value above 10 seconds. In order to reproduce the time course shown in Figure 2.5, we set $\tau_v = 1.7$ seconds.

2.4 Results

2.4.1 Attractive bias toward previous trial target direction

We measured errors in saccade responses in an oculomotor delayed response task as a function of the distance of the previous target from the current target. Error was defined as the angle between the target and the saccade made to the memorized target

location. We subtracted systematic errors related to the current target direction, leaving only the residual error (see Methods). We then tested whether this residual error was related to the location of the previous target, relative to the location of the current target. Figure 2.1 (left) shows the data from one animal, averaged in 30 deg bins and fit to a Gabor function. When the target in the preceding trial was clockwise from the target in the current trial, mean residual error in saccade responses was also clockwise, that is, towards the previous target. When the target in the preceding trial was counterclockwise from the target in the current trial, mean residual error was counterclockwise, again towards the previous target. Thus we found that memory-guided saccade responses are biased in the direction of the target in the previous trial.

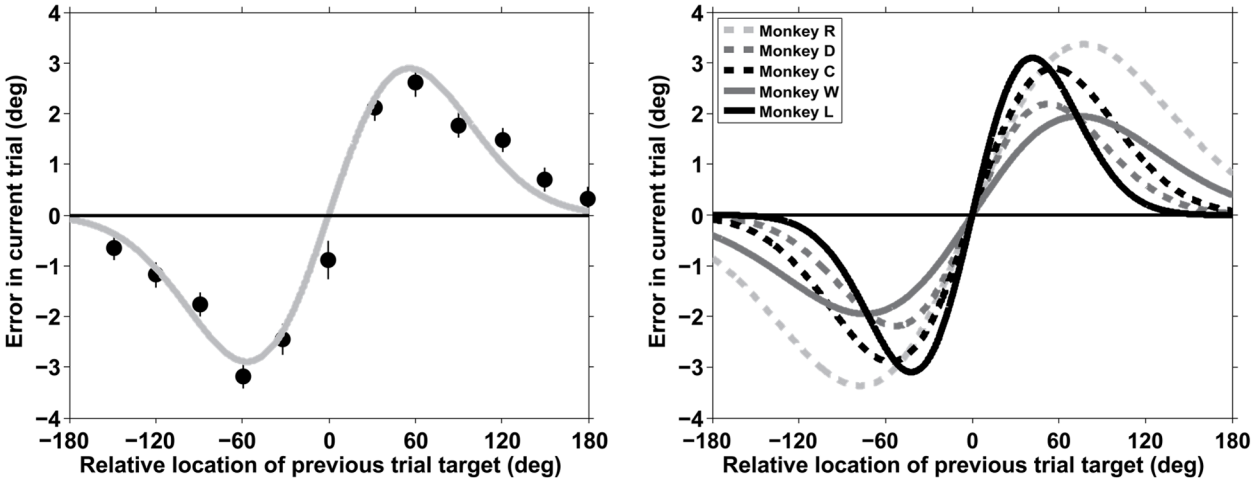


Figure 2.1 Memory bias due to a prior memory.

Left - The influence of previous target on the memory of current targets for monkey C in 7437 trials. The x-axis is the location of the target in the previous trial relative to the location of the target in the current trial, and the y-axis is the mean residual error (see Materials and Methods). The grey line shows the Gabor fit of the data (peak-to-peak = 5.7, $p < 0.005$), and the error bars show the standard error. Right – The Gabor fits for five monkeys. Fits for each monkey were highly significant ($p < 0.005$).

This was true for all five monkeys tested. In each case, Gabor fits to both the raw and binned data were highly significant, accounting for 2 to 7 percent of the variance of the

raw data and 76 to 96 percent of the variance of the binned data points. The peak attractive influence of the previous target occurred when the previous target was between 41 and 78 degrees from the current target.

Bias is unimodal

The small bias toward the previous trial direction could be due to small bias in each and every trial, or due to a large error that occurs only rarely, e.g., an occasional response directed to the previous target location instead of the current location. To examine these possibilities we viewed the distribution of errors for trials in which the previous target was between +35 and +85 degrees from the current target (Figure 2.2). Saccades directed to the previous target would therefore appear at around +60 deg, resulting in a bimodal distribution of errors. The distribution is instead unimodal, with a mode at +1.29 deg and a mean of +2.21 deg, supporting the view that there is a small systematic bias toward the previous target on every trial.

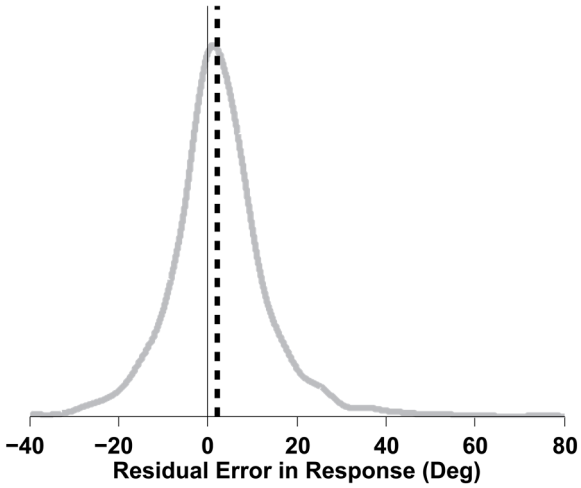


Figure 2.2. Bias is unimodal.
Residual error distribution for mean relative target location between +35 and +85 degrees for all animals. Current target direction was rotated to 0 degrees. The distribution is unimodal with a mode at +1.29 deg and a mean at 2.21 deg indicating a small bias toward the previous target in each trial.

Long-term changes or single-trial effects?

The mechanisms responsible for the response bias observed could reflect persistent, long-term changes that build up over time. Such changes might be beneficial for a system that commonly encounters similar information from one trial to the next. Alternatively the response bias could be due to interference from residual information encoded during the previous trial. We first tested the temporal dynamics of the response bias by varying the length of the intertrial interval while keeping the memory delay fixed. The response bias dropped with increasing ITI in all three animals tested (Figure 2.3).

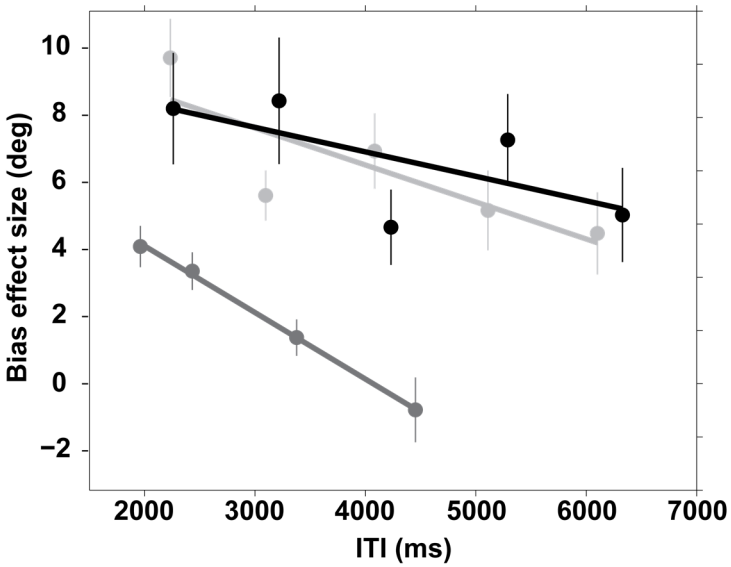


Figure 2.3. Response bias as a function of ITI.
The response bias decreased with increasing ITI in each of the three animals tested. Monkey C (dark grey) -2.00 deg/sec, $p < 0.0005$, N = 2958; monkey R (light grey) -1.13 deg/sec, $p = 0.09$, N = 9026; monkey D (black) -0.73 deg/sec, $p = 0.22$, N = 6437.

If the response bias were due to long-term persistent changes, then consecutive repetitions of target locations might increase the size of the bias (Verstynen & Sabes, 2011). To determine whether the bias toward previous target locations was due to a short-term effect or to long-term plasticity, we presented targets at the same location for one to four consecutive trials, followed by a trial with a target ±60 degrees away. We measured the response bias on the final trial of each set and asked whether it depended

on the number of consecutive repeated presentations that preceded it. Figure 2.4 shows data from two macaques comparing the response bias after a single target trial versus two or more consecutive trials using that same target. The effects were statistically indistinguishable in both monkeys. As shown, there is no significant change in bias in either subject (monkey C, $p > 0.33$, $N = 830$; monkey D, $p > 0.48$, $N = 974$). This finding together with the finding that response bias decreases with increasing ITI provides strong evidence that the bias effect is not due to long-term plasticity. Instead, the response bias is more likely due to interference from the previous trial, possibly due to persistent activity of the previous memory trace.

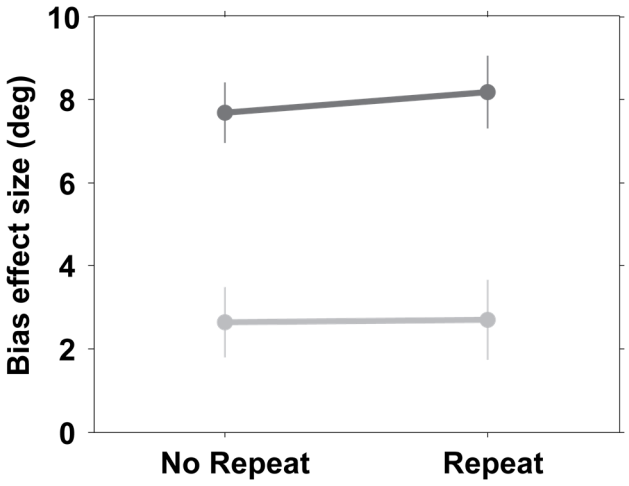


Figure 2.4. The effect of target repetition on response bias.

The bias in trials when a previous target location is repeated is no different from the bias in trials after two or more trials target location repetitions (monkey C (dark) $p > 0.33$, $N = 830$; monkey D (light) $p > 0.48$, $N = 947$). In this experiment targets were always presented either at the same location as or 60 degrees away from the preceding trial's target.

Effect of memory delay length on bias

We asked whether the length of the memory period would affect the bias caused by the previous trial. Simple models that employ a slowly decaying memory trace predict that the bias will decrease with longer memory periods. Longer periods lead to greater decay and hence a reduced effect on the following trial. If, on the other hand, the memory trace does not decay with time (as in an attractor network, for example), then the bias effect should remain constant with increasing delay times. Remarkably, we found

instead that the bias *increased* asymptotically with delay length, with a mean time constant of 1.7 s (Figure 2.5).

In the previous description, we assume that it is the delay length of the *previous* trial that is the key factor affecting bias. In the experiments of Fig. 2.5, memory delay was held constant within a block of trials, that is, every trial had the same delay length. Therefore the dependence of the bias on delay length could reflect the length of the memory delay on either the previous or the current trial. Holding a location in memory for a longer duration on the *previous* trial might result in a greater bias on the subsequent trial. Alternatively, the bias might increase with time within the current trial, independent of the length of the previous trial. We found the latter to be the case.

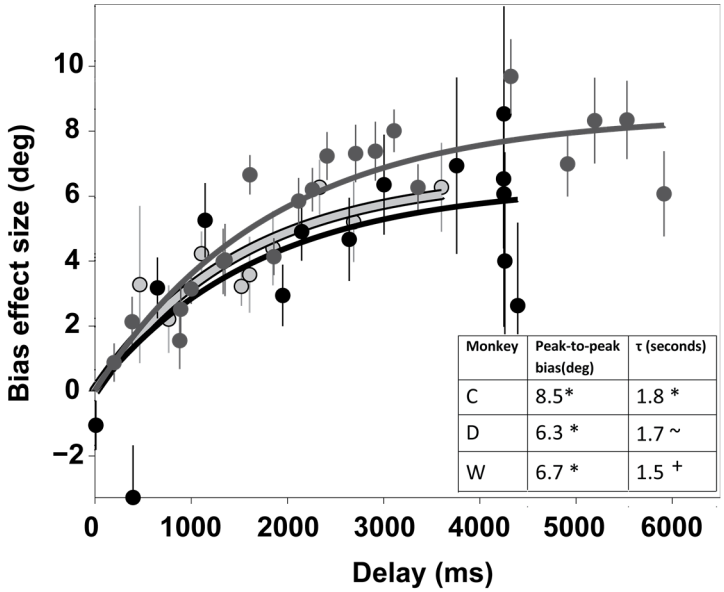


Figure 2.5. Bias as a function of memory delay.
Effect of increasing delay period on peak-to-peak bias for monkey C (dark grey, $p < 0.0001$), monkey D (black, $p < 0.002$), and monkey W (light grey, $p < 0.02$). The inset shows fit coefficients for each monkey (* $p < 0.002$, + $p = 0.06$, ~ $p = 0.28$).

We randomly interleaved short (0.8 s) and long (3.2 s) delay lengths within a block of trials, and then split the data into four sets: (i) short-delay trials followed by long-delay trials, (ii) long-delay trials followed by long-delay trials, (iii) long-delay trials followed by short-delay trials and (iv) short-delay trials followed by short-delay trials. Figure 2.6 shows a representation of trial lengths on the left and the corresponding mean response

bias on the right, averaged over three animals. When the current trial delay length was long, peak response bias was large regardless of the length of the previous delay (short-long 10.2 deg, long-long 10.1 deg). When the current delay was short, peak response bias was small, again regardless of the previous delay (long-short 4.4 deg, short-short, 4.5 deg). An ANOVA (two factors, previous and current delay length; three repeated measures) indicates that only the main effect of the current delay is significant ($F(1,11) = 21.38$, $p < 0.002$). The same result was also obtained for each individual animal. Thus, the bias from the previous trial grows over time within the current trial, and is unaffected by whether the previous target was held in memory for a long or short time.

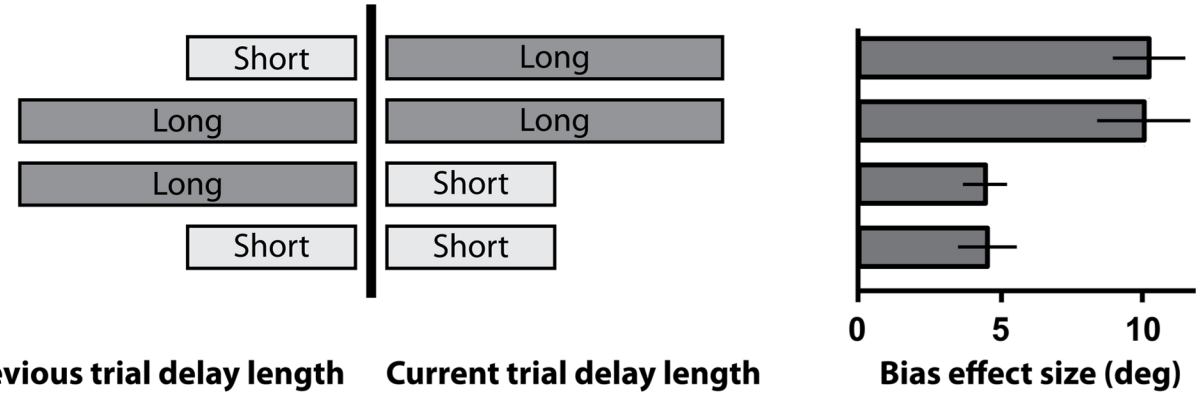


Figure 2.6. Comparison of bias as a function of previous and current delay.

Conditions from top to bottom: (i) Short-delay trials followed by long-delay trials (10.2 deg), (ii) long-delay trials followed by long-delay trials (10.1 deg), (iii) long-delay trials followed by short delay trials (4.4 deg), (iv) short-delay trials followed by short-delay trials 4.5 deg). Only the main effect of the current delay is significant ($F(1,11) = 21.38$, $p < 0.002$).

Neither saccade execution nor memory maintenance are necessary for bias

In most of our experiments, a visual target was presented, the location of that target was maintained in memory, and a saccade was directed to the remembered location. To determine whether either memory maintenance or saccade execution are necessary to produce spatial bias on the subsequent trial, standard memory-guided saccade trials were interleaved with memory-only trials and saccade-only trials. Memory-only trials

differed from standard memory trials in that the subject was required to continue fixation at the end of the memory period rather than making a saccadic response. Saccade-only trials differed from standard memory trials in that no memory cue was presented and the subject performed a visually-guided saccade at the end of the delay period. Biases occurred in standard memory trials that followed either a memory-only or a saccade-only trial, with effect sizes of 7.1 and 6.5 deg, respectively (fit $p < 0.005$ in both cases; data from 383 and 376 trials, respectively, performed by monkey C). This indicates that neither a memory period nor a saccade is required in the previous trial in order to produce an attractive bias.

A second experiment confirmed these results. In the random saccade experiment, standard memory-guided saccade trials were interleaved with trials that were identical to the standard memory trials up until the time of the go cue (fixation offset). At this time a second "random" target appeared, and the animal was required to saccade to this new location rather than to the memorized location. The new location was statistically independent of the memorized location. We found that a standard memory-guided trial that followed a random saccade trial showed a bias toward both the previous trial's memory location (bias = 5.5, fit $p < 0.005$, Figure 2.7 top left) and the previous trial's saccade location (bias = 5.6, fit $p < 0.005$, Figure 2.7 bottom right). This experiment shows that biases can be driven by either memorizing a target without actually moving to it, or by moving to a target without having previously memorized it.

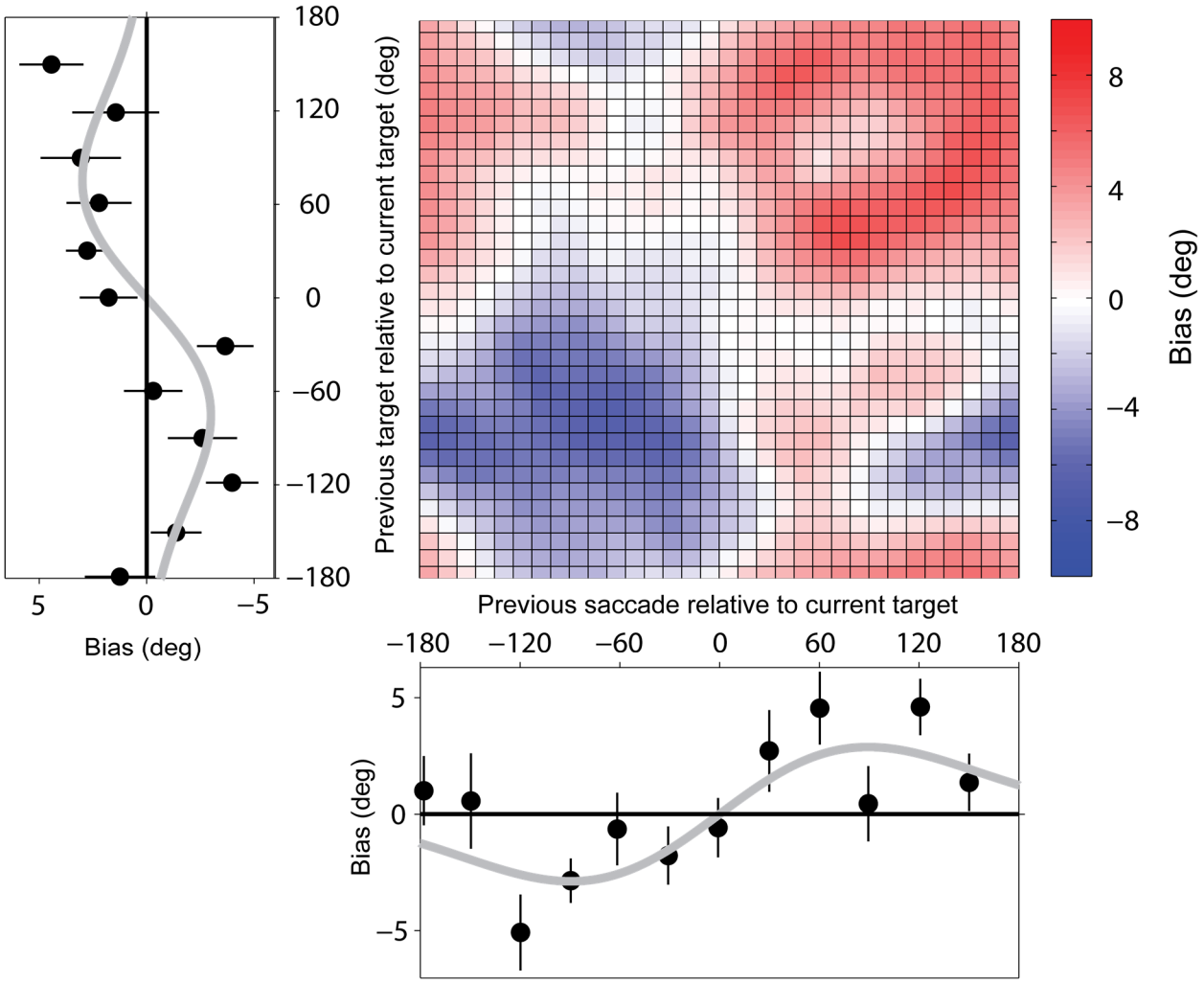


Figure 2.7. Both previous memory location and previous saccade direction bias a subsequent memory-guided response.

Standard memory-guided saccades were interleaved with trials in which a target was memorized, but then, instead of making a memory-guided saccade to that memorized target, a new target was presented and animals made a visually-guided saccade to it (see text). We asked how the two processes, memorizing a previous target and making a saccade in a particular direction, affect a subsequent ("current") memory-guided saccade. The surface fit shows that the current memory-guided saccade is biased by both the previous visually-guided saccade (abscissa) as well as by the previously memorized target (ordinate). The marginal plots show that the effect sizes are similar, with a previous saccade direction bias of 5.5 deg ($p < 0.005$) and a previous memory location bias of 5.6 deg ($p < 0.005$).

There are at least two possible interpretations of these results. First, memorizing a target location and executing a saccade may each, by themselves, be sufficient to bias a memory-guided saccade on a subsequent trial. Alternatively, it may be that merely presenting a visual stimulus is sufficient to generate a bias. There are two arguments for the latter interpretation. First, the magnitude of the effects seen in the random-saccade, memory-only and saccade-only trials are all comparable to the magnitude of effects seen in the standard memory trials. If memory maintenance and saccade execution contributed independently and equally to the bias effects, then we would expect that the magnitude of effects on standard memory trials would be twice as large as the effects in the other trial types. Second, previous studies using distractors show attractive bias toward irrelevant (distractor) stimuli that subjects ignore while performing a standard spatial memory task (Chumbley et al. 2008, Macoveanu et al. 2007). Taken together, the evidence suggests that neither memory maintenance nor saccade execution is necessary to produce bias, and that instead it is the presentation of a visual stimulus that produces interference on a subsequent memory-guided saccade trial.

Reaction time effects

It is possible that the spatial biases we observe are accompanied by reaction time (RT) effects. We asked whether the reaction time of a saccade to a memorized target depends on the relative location of the target on the previous trial. We removed mean reaction time as a function of current target location and computed residual reaction time as a function of previous target location. An ANOVA showed no effect of relative target location on memory-guided saccade RT (previous and current targets within +/- 45 deg of one another: RT = 0.26 ms; from 45 to 135 degrees apart: RT = -0.02 ms; from 135 to

180 degrees: RT = -0.12 ms; $F(2,32121) = 0.19$, $p = 0.83$). Note that these same trials showed clear spatial biases (Fig. 2.1).

2.4.2 Interference in attractor networks

To model the data in these experiments we implemented a continuous attractor network as described in the *Materials and Methods* section and Compte et al. 2000. In this model, memory for a particular location is maintained by strong recurrent excitatory connections between neurons with similar preferred directions. This excitatory input is counteracted by inhibitory connections between neurons with dissimilar preferred directions. The resulting bump in population activity sustains its shape over time, but can drift due to stochastic fluctuations in activity.

The attractor network framework is commonly used for modeling working memory circuits. These models match many of the physiologically observed properties of the neuronal circuits believed to be involved in memory, and successfully account for many of the behavioral phenomena (Durstewitz et al. 2000, Wang 2009). These models display within-trial interference effects with a spatial profile that is reminiscent of the between-trial effects we observe in our experiments. In particular, distractors in continuous attractor models exert an attractive bias that is stronger when the distractor is close to the target location and weaker when it is farther away (Compte et al. 2000). This bias results from local positive feedback loops within the attractor network. Therefore we chose these networks to model our between-trial interference effects, positing that the same structures that give rise to within-trial interference might also generate between-trial effects with a similar spatial profile.

The network stimulation paradigm we used in our simulations was based on our behavioral paradigm. For simplicity, each simulation consisted of just two trials, a previous and current trial. At the start of each simulation a network of 1024 excitatory and 256 inhibitory neurons was randomly initialized and allowed to evolve without any external input for 0.25 s. A target direction was then chosen and presented in the form of a 70 pA, 0.25 s current injection to the 100 excitatory neurons most closely tuned to that location. This was followed by a 1 s memory period during which the elevated activity was maintained in the neurons around the target location. Next, 140 pA of current was injected broadly to all excitatory neurons in the network for a duration of 0.1 s in order to reset the network. This reset signal displaces the network away from the memory-holding attractor state and back to a baseline state after a short time. Following an ITI period of 0.15 s during which network activity was allowed to evolve without any external stimulation, a second target stimulus with an identical profile to the first was presented at a new location. After a delay of 1 s we fit a Von Mises function to the network activity and read out the peak as the network's response in the second trial.

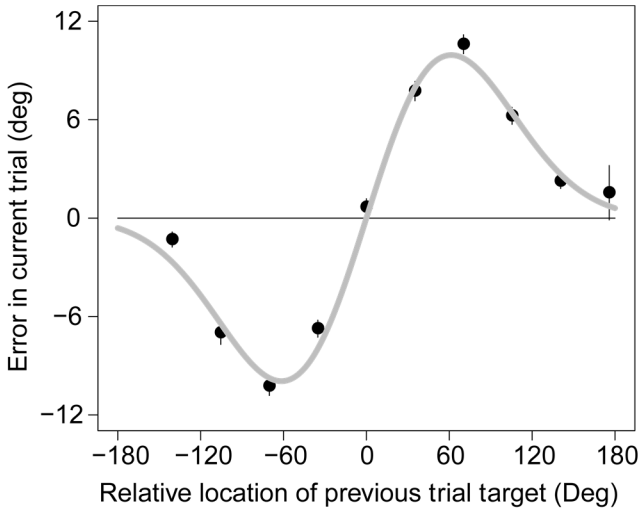


Figure 2.8. Influence of previous target on the memory of current targets as predicted by the continuous attractor network model.
The x-axis is the difference between the current and previous target direction and the y-axis is the mean error, as in Fig. 2.1. The grey line represents the Gabor fit to the model data ($p < 0.0001$) with peak-to-peak = 19.9 deg ($p < 0.0001$).

The attractor network replicated the spatial aspect of the bias effect. Although the reset signal between the first and second trials causes the network to leave its attractor state, when the ITI is short there

is still enough residual activity at the start of the next trial to bias the network toward the previous trial's target location. Figure 2.8 shows response error (the difference between target and response locations) on the second trial as a function of the distance between the first and second target locations. Similar to the behavioral data, the simulation data shows a bias toward the previous target location that is modulated by distance.

The attractor network model also reproduces the finding that bias increases with delay length. After stimulus presentation, the network only gradually settles into the attractor state. The speed at which this occurs depends on network parameters. If the first memory period is short, then the network may not reach the full attractor state. As a result, the carry-over of information from the first to the second trial is reduced for short memory periods (Figure 2.9).

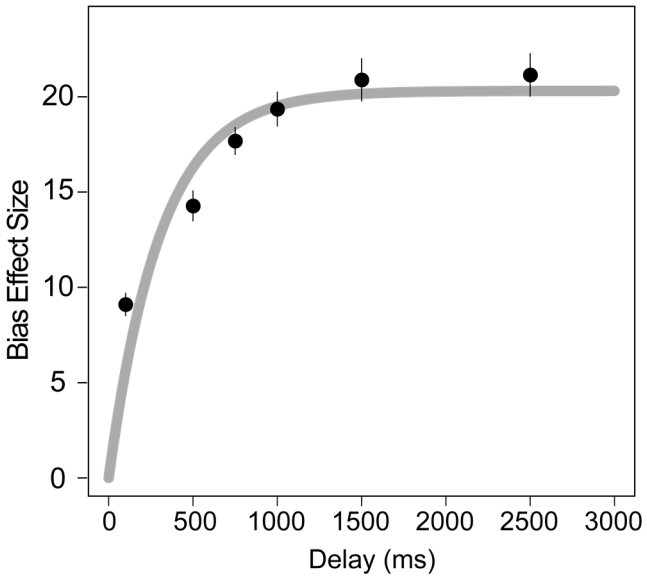


Figure 2.9. Effect of increasing delay period on peak-to-peak bias predicted by the continuous attractor network model.

The black line is the exponential fit to the data ($p < 0.011$) with peak-to-peak bias = 20.3 deg ($p < 0.0002$) and time constant = 308 ms, $p < 0.03$).

Critically, the build-up of the response bias as the memory period increases in duration is related to the duration of the previous trial and not the current trial. Unlike the animal data (Fig. 2.6), the magnitude of the response bias from the network depends on the length of the previous trial, not on the length of the current trial (Figure 2.10). When the previous delay was long, response bias was large regardless of the length of

the current delay (long-short 21.3 deg, long-long 21.6 deg). When the previous delay was short, response bias was small, again regardless of the current delay (short-short 14.8 deg, short-long, 14.6 deg). An ANOVA (two factors, previous and current delay length; three repeated measures) indicates that only the main effect of the previous delay is significant ($F(1,11) = 532.81$, $p < 0.0001$). Thus the network fails to reproduce the results of the mixed delay experiment. These results allow us to reject the hypothesis that the dynamics of a standard attractor model can solely account for the increase in bias as a function of memory period length.

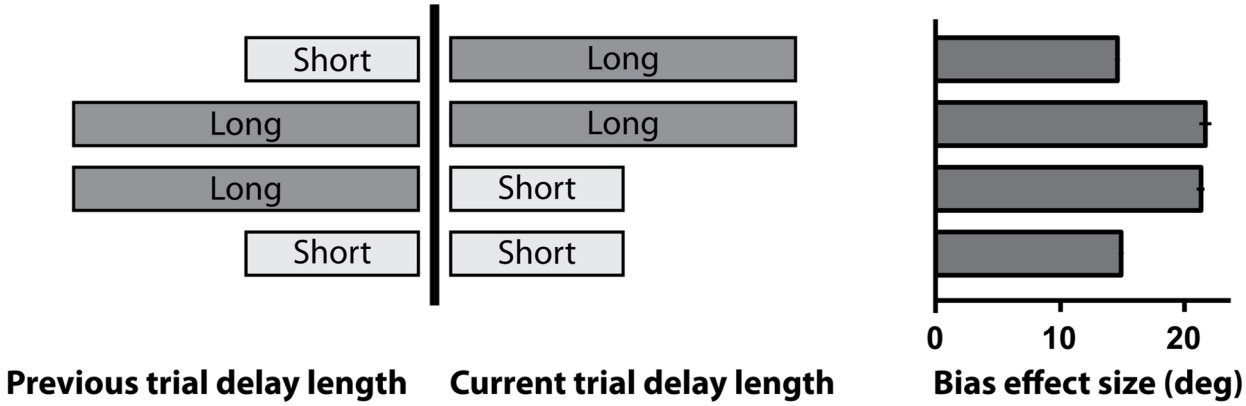


Figure 2.10. Comparison of bias as a function of previous and current delay in network simulations.
Conditions from top to bottom: (i) Short-delay trials followed by long-delay trial (14.6 deg) (ii) long-delay trials followed by long-delay trials (21.6 deg) (iii) long-delay trials followed by short delay trials (21.3 deg) (iv) short-delay trials followed by short-delay trials (14.8 deg). Only the main effect of the previous delay is significant ($F(1,11) = 532.81$, $p < 0.0001$).

2.4.3 Two-store model of short-term memory

To capture the dependence of the response bias on the duration of the current memory period, rather than on the duration of the previous memory period, we modified the attractor model by adding a short-term visual store. The visual store is ephemeral, discharging much more quickly than the main memory store. The visual store is fully

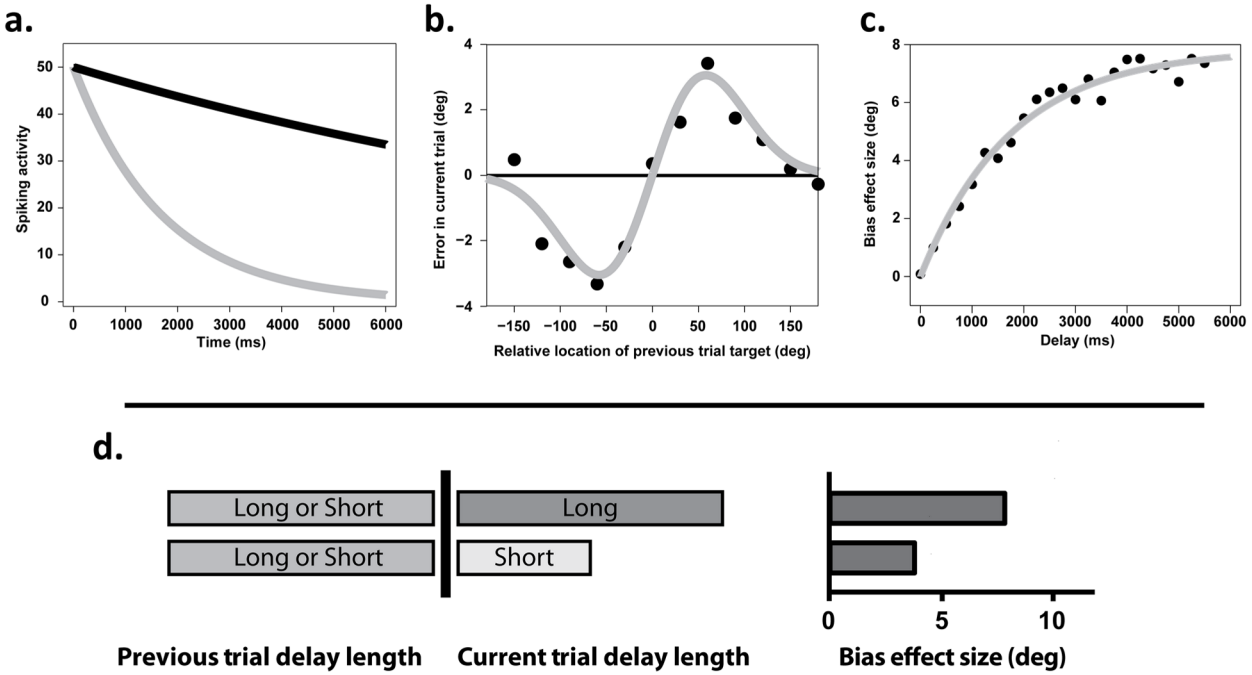


Figure 2.11. Two-store model predictions.

a. Activity profiles of the visual store (grey) and the memory store (black). **b.** Effect of previous trial on the memory of current target for a delay of 3s. **c.** Effect of increasing delay period on the bias. **d.** Predictions of the two-store model in the Long-Short experiment.

charged by the end of the stimulus presentation, but this charge then decays with a fast time constant (Figure 2.11a). This decay is too rapid to provide a conduit for contaminating the current trial with residual activity from the previous trial. The attractor network, as just described in the single store model, has a very long time constant and so the stimulus can be maintained indefinitely. In the two-store model, the target location estimate is determined by the fractionally weighted estimates of the veridical visual sensory store and the biased sustained store. For visually-guided and shorter memory-guided trials, the veridical visual circuit accounts for all or most of the contribution to the target estimate resulting in little bias or no bias. For longer trials, activity in the visual circuit has decayed and therefore the biased sustained memory circuit contributes the most strongly to the target estimate, resulting in a biased output (Figure 2.11b). Thus, as the duration of the memory period increases and the activity in

the visual store drops exponentially, the response bias grows as a saturating exponential (Figure 2.11c). Critically, the response bias of this model, like that of our subjects, depends on the length of the current trial, not the previous trial (Figure 2.11d). Parameters in this model are constrained as described in *Materials and Methods*, with the visual-store time constant as the only free parameter. In order to reproduce the time course shown in Figure 2.5, we set $\tau_v = 1.7$ seconds in the simulations shown in Fig. 2.11.

We asked whether the visual store was, like human iconic memory (see Discussion), disrupted by the presentation of a visual mask. If so, then we would expect that a mask would force the system to rely on the (biased) working memory store, thereby sharply increasing the bias seen early in the memory period. If, on the other hand, the visual store is robust, then a mask will not change the early bias. To test this, monkey R participated in an additional experiment in which, in half of the trials, a mask was presented 67ms after the memory target onset. The target was presented for 50ms and the mask (a dense display of line-segments at random orientations, with the same color as the target) was presented for 100 ms. Memory periods were 600, 1200, or 2400 ms. Without a mask, bias at 600 and 1200 ms was 41% and 44%, respectively, of the full bias effect measured at 2400 ms (see also Fig. 2.5). With a mask, bias at 600 and 1200 ms increased to 54% and 51% of the full bias, respectively. These increases were not significant ($p = 0.40$ and 0.64 respectively, $N = 7373$), suggesting that the transient visual store in our two-store model is relatively robust to masking effects.

2.5 Discussion

In this study we investigated how spatial stimuli held in working memory are influenced by memoranda from the previous trial. We found that responses are biased toward previous stimulus locations and that this bias is strongest when the previous trial target and current trial target are separated by about 60 deg of arc. The bias toward the previous target location did not increase with consecutive target presentations at that location, suggesting that the bias is a trial-by-trial effect perhaps due to a residual memory trace rather than a long-term synaptic effect. We also found that the size of the bias increases asymptotically when increasing the memory period of the current trial. We were able to model the spatial but not the temporal aspects of the bias with a standard attractor network. In order to model the increase in bias with the length of the current memory period we used a two-store model with a rapidly decaying visual store and biased sustained memory store.

The response bias that we see from the previous stimulus is similar to proactive interference effects reported in prior literature. Classic examples of proactive interference in humans show long-term buildup of interference over repeated memorization sessions often spanning many days (e.g. Underwood 1957). However, there are also reports of interference effects within a single session. In particular, Dunnet & Martel (1990) report proactive interference due specifically to the immediately preceding trial (but see Wright et al. 2012). In some studies, the magnitude of interference grows asymptotically with the length of the delay period (e.g. Moise 1976, Dunnett & Martel 1990, Edhouse and White 1988), exactly as we show here (Fig. 2.5), although this was not commented on in the original studies. Another point of

similarity between our results and previous studies of proactive interference is that interference effects are larger when the previous and current stimuli share similar properties, such as their spatial location (e.g. Wickens et al. 1963, Makovski & Jiang 2008). Most of these previous studies use categorical memoranda and measure interference effects in percent correct. We use continuous memoranda and see analogous interference effects in accuracy, suggesting that similar mechanisms are at play.

A preceding memory may cause interference by overwriting a current memory, adding non-systematic noise, or adding a systematic bias to the current memory. Distinguishing between these possibilities based only on the proportion of correct responses can be difficult. An advantage of using continuous memoranda and responses is that we can more finely characterize the spatial and temporal structure of the interference and thereby address how previous information interferes with current memory. This in turn provides insight into the underlying circuits for working memory. In particular, we find that the interference is not caused by overwriting of the previous memory or the introduction of non-systematic noise, but instead that previous information combines with and systematically biases current information.

This systematic bias is not due to persistent, long-term adaptation. Verstynen and Sabes (2011) describe such an effect in a visually-guided reaching task. They find that attractive bias increases with repeated presentations of the same target location. This was not the case in the current study (Fig. 2.3). Instead, the bias from the previous trial appears to be a single-trial effect, likely due to either residual activity within the network or from short-term changes in synaptic efficacy from the previous trial.

The spatial profile of the bias effect is similar to that of distractor effects in attractor network simulations (Compte et al. 2000) arising from the recurrent connectivity structure in attractor models. We show that a standard attractor network shows proactive interference effects in the form of an attractive bias toward the previous target location and replicates the spatial aspect of the bias. The animal's bias increases as delay increases, rising with a time constant of just over 1.5 s (Fig. 2.5) and the critical time interval is the duration of the memory period in the current trial (Fig. 2.6). Although the attractor model predicts a similar increase in bias with delay, the bias depends on the duration of the *previous* memory period (Figs. 2.9 and 2.10). This occurs because, with short trials, the memory trace in the attractor model does not have sufficient time to build up to the full attractor state, and as a result there is less residual activity remaining after the reset signal, leading to a weaker effect on the subsequent trial.

We asked whether saccade execution or active memory maintenance were necessary to produce bias. We show that while a visually-guided saccade or memory period will, on their own, bias a subsequent memory-guided saccade, neither are necessary. Furthermore, behavioral studies with distractors show that merely presenting an irrelevant visual stimulus is sufficient to produce an attractive bias (Macoveanu et al. 2007, Chumbley et al. 2008). These findings suggest that visual target presentation, which occurs in both saccade-only and memory-only trials, is the most likely cause for the observed response bias in every case.

A large body of work in humans and non-human primates has shown that the reaction time of responses to visual stimuli depends on the location of previous targets. For

example, reaction times in a center-out visually-guided saccade task depends on where the target lies, relative to the target location on the previous trial (Dorris et al. 1999). Monkeys execute saccades more quickly to a target whose location coincides with the target of the previous trial. This may be related to the phenomenon of inhibition of return, which was first described in humans, although in humans the polarity of the effect is reversed – saccades to targets at repeat locations are slowed rather than sped up (Rafal et al. 1994, Taylor 1997, Taylor & Klein 1998.) These reaction time effects have been modeled as a consequence of transient suppression of cells involved in saccade generation (e.g. Dorris et al. 1999). The same mechanism could also lead to spatial biases in saccades. If so, then reaction time and spatial bias effects should appear in tandem. However, as described in Results, we found spatial but not reaction time effects in the memory-guided saccade task. In saccade-only trials, we found reaction time slowing when the target was in the same direction as the previous trial's target, but not spatial biases (RT difference (same – opposite): 35.7 ms, $p < 0.002$; Peak-to-peak bias: 0.59 deg, $p = 0.74$, $N = 377$). This double dissociation provides strong evidence that spatial bias and reaction time effects rely at least in part on different neuronal mechanisms.

To model a build-up in bias that depends on the *current* trial duration, we posit two separate memory stores, a visual sensory store and a memory store. In this model, the saccade to the remembered target is a combination of information from the two stores. The visual sensory store decays quickly, losing all information within a few seconds after target onset. Classic human psychology studies have identified iconic memory, a short-term store that retains the visuo-spatial characteristics of stimuli for a brief interval

after stimulus offset (e.g. Sperling 1960). Due to its short duration and its apparent inability to survive eye movements (Irwin 1992, Irwin & Andrews 1996), iconic memory is unlikely to show substantial interference from the previous trial and in this respect is a good match for the quickly decaying visual store posited in our model. However, most studies describe iconic memory as completely decaying in 700 ms or less (e.g. Sperling 1960, but see Averbach & Sperling 1961). This is too short to match our data, which requires a time constant closer to 1.7s for the decay of the visual store. In addition, iconic memory is overwritten when a mask is presented following the memory stimulus. Our model predicts that, by overwriting the contents of iconic memory almost immediately after target presentation, the memory-guided saccade will rely only on the longer, biased memory store, and will therefore show substantial bias even with very short memory periods. Contrary to this, we find that a mask has little effect on the early bias. Thus, both the observed time constant and the effect of a visual mask suggest that iconic memory is not involved. Other visual short-term memory stores intermediate to iconic memory and working memory have been identified in behavioral (Sligte et al. 2008, Pinto et al. 2013, Vandenbroucke et al. 2014; see also Griffin & Nobre 2003, Makovski and Jiang 2007; Makovski et al. 2008) and neurophysiological studies (Sligte et al. 2009; see also Pasternak & Greenlee 2005, Bisley et al. 2004). These stores have longer time constants than the iconic store, with some studies showing information retention for up to 4 seconds.

The sustained spatial working memory store holds information for a much longer time than the visual store, and is biased by residual activity from the previous trial. To produce a behavioral response the signals from the visual sensory store and sustained

memory store are combined and weighted by the ratio of their activity. Immediately after target presentation, the visual sensory store is highly active and provides most of the signal used to guide the saccade. After several seconds, this unbiased store has decayed away, and the biased memory store provides most of the signal used to guide the saccade. This model successfully replicates the observed features of memory (Fig. 2.11). These findings suggest that short-term memory circuits may be composed of multiple memory stores with independent decay rates. Furthermore, the visual sensory store may hold a more veridical representation of visual information than the sustained memory store and may act to shield against systematic biases arising from the recurrent attractor circuit dynamics of the sustained store.

2.6 Conclusions

We characterized proactive interference in a simple oculomotor delayed response paradigm and describe several new and significant properties. First, proactive interference is graded, not all-or-none. Second, the spatial profile of proactive interference is well modeled by attractor circuits. Third, proactive interference resembles interference from within-trial distractors which suggests both proactive and distractor interference arise due to the same memory mechanisms. Fourth, proactive interference was manifest only on delayed saccade responses, not on visually-guided saccades. This distinguishes the process from inhibition of return, which affects all saccades. Fifth, proactive interference increased as a function of delay length. This is significant because it implies that not one but two memory stores are in operation: a

short term store (distinct from iconic memory) that is not susceptible to proactive interference, plus a long term store that is susceptible.

References

Altgassen M, Phillips L, Kopp U, Kliegel M. Role of working memory components in planning performance of individuals with Parkinson's disease. *Neuropsychologia* 45:2393-2397, 2007

Averbach E, Sperling G. Short term storage of information in vision. In: *Cherry, C (ed.), Information Theory,* pp196-221, Washington, DC: Butterworth & Co., 1961

Baddeley A. Working Memory: Theories, Models, and Controversies. *Annu. Rev. Psychol.* 63:1-29, 2012

Bisley JW, Zaksas D, Droll JA, Pasternak T. Activity of neurons in cortical area MT during a memory for motion task. *J. Neurophysiol.* 91(1):286-300, 2004

Chumbley JR, Dolan RJ, Friston KJ. Attractor models of working memory and their modulation by reward. *Biol. Cybern.* 98:11-18, 2008

Cohen G. Memory in the real world (2nd ed.) Hove: Psychology Press, 1996

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex* 10:910-923, 2000

Daneman M, Carpenter PA. Individual differences in working memory and reading. *J. Verb. Learning & Verb. Behav.* 19:450-466, 1980

Daneman M, Merikle PM. Working memory and language comprehension: A meta-analysis. *Psychonomic Bull. Rev.* 3(4):422-433, 1996

Dorris MC, Taylor TL, Klein RM, Munoz P. Influence of previous visual stimulus or saccade on saccadic reaction times in monkey. *J. Neurophysiol* 81:2429-2436, 1999

Dunnett SB, Martel FL. Proactive interference effects on short-term memory in rats: I. Basic parameters and drug effects. *Behavioral Neuroscience* 104(5):655-665, 1990

Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. *Nat. Neurosci.* 3(Suppl.):1184-1191, 2000

Edhouse WV, White KG. Sources of proactive interference in animal memory. *J Experimental Psychol.* 14(1):56-70, 1988

Engle RW, Kane MJ. Executive attention, working memory capacity, and a two-factor theory of cognitive control. In: The psychology of learning and motivation Vol 44 (ed. Ross BH), pp145-199 New York: Elsevier Academic Press, 2004

Gnadt JW, Bracewell RM, Andersen RA. Sensorimotor transformation during eye movements to remembered visual targets. *Vision Research* 31(4):693-715, 1991

Griffin IC, Nobre AC. Orienting attention to locations in internal representations. *J. Cognitive Neurosci.* 15(8):1176-1194, 2003

Irwin, DE. Memory for position and identity across eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 307–317(28), 1992

Irwin DE and Andrews RV. Integration and accumulation of information across saccadic eye movements. In: *Attention and Performance XVI (Inui, T. and McClelland, J.L., eds)*, pp. 125–155, MIT Press, 1996

Jarvik ME, Goldfarb TL, Carley JL. Influence of interference on delayed matching in monkeys. *J. Experimental Psychol.* 81(1), 1969

Jonides J, Nee DE. Brain mechanisms of proactive interference in working memory. *Neuroscience* 139:181-193, 2006

Just MA, Carpenter PA. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 1:122-149, 1992

Macoveanu J, Klingberg T, Tegnér J. Neuronal firing rates account for distractor effects on mnemonic accuracy in a visuo-spatial working memory task. *Biol. Cybern.* 96:407-419, 2007

Makovski T, Jiang Y. Distributing versus focusing attention in visual short-term memory. *Psychonomic Bull. Rev.* 14(6):1072-1078, 2007

Makovski T, Jiang Y. Proactive interference from items previously stored in visual working memory. *Memory & Cognition* 36(1):43-52, 2008

Makovski T, Sussman R, Jiang YV. Attention in visual working memory reduces interference from memory probes. *J. Experimental Psychol.* 34(2) 369-380, 2008

Miyake A, Shah P. Models of working memory: Mechanisms of active maintenance and executive control. New York, NY: Cambridge Univ. Press, 1999

Moise SL Jr. Proactive effects of stimuli, delays, and response position during delayed matching from sample. *Animal Learning & Behavior* 4:37-40, 1976

Pasternak T, Greenlee MW. Working memory in primate sensory systems. *Nature* Reviews 6(2):97-107, 2005

Payne JW, Beuman JR, Johnson EJ. The adaptive decision maker: Effort and accuracy in choice. In: Insights in decision making: A tribute to Hillel J. Einhorn (ed. Hogarth RM), pp129-153 Chicago: Univ. Chicago Press, 1990

Payne JW, Bettman JR, Johnson EJ. Behavioral decision research: A constructive processing perspective. *Annu. Rev. Psychol.* 43:87-131, 1992

Pinto Y, Sligte IG, Shapiro KL, Lamme VAF. Fragile visual short-term memory is an object-based and location-specific store. *Psychonomic Bull. Rev.* 20:732-739, 2013

Rafal RD, Egly R, Rhodes D. Effects of inhibition of return on voluntary and visually guided saccades. *Can. J. Exp. Psychol.* 48:284–300, 1994

Raghubar KP, Barnes MA, Hecht SA. Working memory and mathematics: A review of developmental, individual, difference, and cognitive approaches. *Learning and Indiv. Differences* 20:110-122, 2010

Sligte IG, Scholte HS, Lamme VAF. Are there multiple visual short-term memory stores? *PLoS One* 3(2):e1699, 2008

Sligte IG, Scholte HS, Lamme VAF. V4 activity predicts the strength of visual short-term memory representations. *J. Neurosci.* 29(23):7432-7438, 2009

Sperling G. The information available in brief visual presentations. *Psychological Monographs: General and Applied* 74(11):1-29, 1960

Taylor, TL. Generating and Measuring Inhibition of Return (PhD thesis). Halifax, Nova Scotia, Canada: Dalhousie University, 1997

Taylor TL, Klein RM. On the causes and effects of inhibition of return. *Psychonomic Bull. Rev.* 5(4): 625–643, 1998

Underwood BJ. Interference and forgetting. *Psychological Review* 64(1):49-60, 1957

Unterrainer JM, Owen AM. Planning and problem solving: From neuropsychology to functional neuroimaging. *J. Physiol. Paris* 99:308-317, 2006

Vanderbroucke ARE, Sligte IG, Barrett AB, Seth AK, Fahrenfort JJ, Lamme VAF. Accurate metacognition for visual sensory memory representations. *Psychological Science* 25(4):861-873, 2014

Verstynen T, Sabes PN. How each movement changes the next: An experimental and theoretical study of fast adaptive priors in reaching. *J. Neurosci.* 31(27):10050-10059, 2011

Wang XJ. Attractor Network Models. In: *Squire LR (ed.) Encyclopedia of Neuroscience*, Vol1 pp667-679. Oxford Academic Press, 2009

White JM, Sparks DL. Saccades to remembered targets: A behavioral and neurophysiological study. *Inv. Ophthalm. Vis. Sci. (Suppl.)* 27:155, 1986

White JM, Sparks DL, Stanford TR. Saccades to remembered target locations: An analysis of systematic and variable errors. *Vision Research* 34(1):79-92, 1994

Wickens DD, Born DG, Allen CK. Proactive inhibition and item similarity in short-term memory. J. *Verb. Learning & Verb. Behav.* 2:440-445, 1963

Wright AA, Katz JS, Ma WJ. How to be proactive about interference: Lessons from animal memory. *Psychological Science* 23(5):453-458, 2012

Chapter 3

Ghosts in the Machine II: Neural Correlates of Memory Interference from the Previous Trial

3.1 Abstract

Previous memoranda interfere with working memory. For example, spatial memories are biased toward locations memorized on the previous trial. We predicted, based on attractor network models of memory, that activity in the frontal eye fields (FEF) encoding a previous target location can persist into the subsequent trial and that this ghost will then bias the readout of the current target. Contrary to this prediction, we find that FEF memory representations appear biased away from (not towards) the previous target location. The behavioral and neural data can be reconciled by a model in which receptive fields of memory neurons converge toward remembered locations, much as receptive fields convergence towards attended locations. Convergence

increases the resources available to encode the relevant memoranda and decreases overall error in the network, but the residual convergence from the previous trial can give rise to an attractive behavioral bias on the next trial.

3.2 Introduction

Working memory, the ability to actively maintain and transform information, is necessary for performing a wide range of cognitive tasks. Spatial working memory is of particular interest. In a spatial working memory task the memoranda and responses are locations in space and are naturally continuous, not categorical, allowing investigators to finely probe the properties of the working memory circuits. Spatial working memory tasks can be easily performed by animals, allowing animal neurophysiology to be compared with human fMRI data.

Neurophysiology studies in monkeys identify the prefrontal cortex (PFC) as one of the key regions involved in maintenance of spatial information. Dorsolateral PFC and frontal eye fields (FEF) show a sustained increase in firing rates during spatial working memory tasks (Bruce and Goldberg, 1985; Chafee and Goldman-Rakic, 1998; Constantinidis et al., 2001; Ferrera et al., 1999; Funahashi et al., 1989, 1993; Fuster and Alexander, 1971; Kojima and Goldman-Rakic, 1982; di Pellegrino and Wise, 1993; Sommer and Wurtz, 2001; Takeda and Funahashi, 2002, 2004; Umeno and Goldberg, 2001). FEF is also involved in transforming visual signals into saccadic commands (Bruce and Goldberg, 1985; Schall, 1991; Sommer and Wurtz, 2000). During memory

tasks that allow for saccade plan generation early in the memory period, fMRI signals in human FEF show increased coherence with a network of oculomotor areas (supplementary eye fields, dorsal anterior cingulate) involved in maintaining saccade goals. In contrast, during memory tasks that prevent saccade planning until late in the trial, FEF instead shows increased coherence with a different network of areas (dorsolateral PFC, superior frontal sulcus, posterior parietal cortex). These areas are thought to be involved in sustaining covert attention at a particular spatial location (Corbetta et al., 2002; Curtis et al., 2005). These results suggest that FEF plays an important role both in maintaining the perceived position of a stimulus and in transforming that information into a saccade plan that can be maintained over time.

Working memory is susceptible to interference from previous stimuli (proactive interference: Dunnett and Martel, 1990; Edhouse and White, 1988; Jarvik et al., 1969; Moise, 1976; see also Jonides and Nee, 2006). Papadimitriou et al. (2015) identified proactive interference in a memory-guided saccade paradigm. The interference produces a bias with well-defined spatial and temporal characteristics. In this study we look for neural correlates of this bias in the spiking patterns of FEF neurons. We find two possible candidates: residual activity encoding the previous target, and a shift in the activity encoding the current target. The shift in activity provides a better temporal match to the behavioral bias, yet is in a direction that is opposite to that which we would have predicted.

To resolve this inconsistency we suggest that the shift in target-encoding activity may arise from a shift in receptive field positions. Previous reports suggest that receptive

fields in FEF move toward the target of an upcoming saccade (Zirnsak et al., 2014). Receptive field changes have also been shown in spatial attention tasks in V4 (Connor et al., 1997; Tolias et al., 2001). We present a model in which convergence of mnemonic fields toward memory targets can reconcile our neuronal and behavioral data.

3.3 Materials and Methods

Subjects

Two *Macaca mulatta* (M1, M3) and one *Macaca fascicularis* (M2) were used as subjects. Monkeys were fitted with a prosthetic device to stabilize the head, a single scleral search coil for eye movement recording (Judge et al., 1980; Robinson, 1963), and a recording chamber over either the left or right arcuate sulcus. Sterile surgery was performed under inhalation anesthesia (isoflurane, 0.5-2.0%). Post-operative analgesics were provided as necessary. All surgical and behavioral procedures conformed to National Institutes of Health guidelines and were approved by the Washington University Institutional Animal Care and Use Committee.

Recording Procedures

During experiments, the monkey was seated in a Lexan box (Crist Instruments). Eye movements were monitored using earth-mounted 4' rectangular field coils (CNC Engineering). Visual stimuli were projected (Electrohome, Model ECP 4100) onto a 100 x 80 cm screen placed 58 cm from the animal. The room was otherwise completely dark,

as confirmed by a dark-adapted human observer. All aspects of the experiment were computer-controlled (custom software). Eye position was logged every 2 ms. Visual stimulus presentation times were accurate to within one video refresh (17 ms).

Electrophysiological recording and stimulation were performed with tungsten microelectrodes (FHC or Alpha Omega; 0.2–2.0 M-ohms). Extracellular potentials were amplified (FHC) and filtered (band pass 400–5000 Hz; Krohn-Hite). Single units were isolated with a dual time-amplitude window discriminator (BAK Electronics).

Memory-guided saccade task

We trained three macaque monkeys to perform a memory-guided saccade task in which each animal first fixated a central fixation point. A peripheral target was then presented for 150 ms, followed by a 1.4, 2.8 or 5.6 s (randomly interleaved) memory period. Each animal was required to maintain fixation within 1.5° until the fixation point was extinguished, cueing the animal to saccade to within 1.5°–3.5° of the location of the target, depending on its eccentricity. Targets were presented at a single eccentricity, adjusted to the preferred eccentricity of the unit (range 5°–20°, mean 13°). Targets were presented at up to 16 possible locations along the circumference of a virtual circle (angular separation of 22.5°). On average, we presented 12 targets per unit and collected 8 trials of each memory period length per target.

Memory screening task

This task was used to screen single units for further study during an experimental session. The task was identical to the memory-guided saccade task, except that the delay could vary randomly from 1000–2000 ms. Targets were presented at up to 16 possible locations at either 10° or 20° eccentricity (angular separation of 45°).

FEF screening task

FEF sites were defined as those at which electrical microstimulation with current less than 50 μA evoked consistent saccadic eye movements (bipolar stimulation pulses, negative leading, 250 μs/phase, 333 Hz, 70 ms duration, applied with a software-controlled stimulus isolation unit [FHC] (Bruce et al., 1985)). In the screening task, animals began by fixating a central target for 400 ms. The target was extinguished, and in half of trials, stimulation began 100 ms later. The fixation point reappeared 300 ms after the initial offset. The animal was rewarded on all stimulation trials and also on control trials in which the eyes remained at the fixation target.

Behavior analysis

Analysis of behavioral error was similar to (Papadimitriou et al., 2015). Briefly, within each memory-guided saccade trial, we projected saccade response vectors to the unit circle. That is, we removed the radial component of each response and considered only the angular component. Using the interval 100 ms to 300 after the saccade, we calculated saccade error as the difference between the saccade angular direction and the

target direction. For each target location we then subtracted the mean error across all trials (systematic error) from the error in each individual trial. Response error as a function of relative target location was then fitted to a Gabor function:

$$y(x) = height * \sin(width * x)\, e^{-(width*x)^2} \qquad (Eqn.\ 1)$$

where y is the response error and x is relative direction of the previous trial's target (previous minus current target angles).

Population-averaged tuning curves

We wished to know how neural tuning curves might be altered as a function of the previous trial's target. A tuning curve describes how a cell responds to a range of target locations. In the population-averaged tuning curve receptive field centers of all cells are moved to a common location (0 deg) and cell responses are averaged. The receptive field center of each neuron was determined by fitting a Von Mises function to firing rate as a function of target direction in the interval 50 ms to 300 ms after target onset. Next, target directions were expressed relative to each cell's receptive field, that is, each cell's receptive field center was subtracted from each target direction. We then generated population-averaged tuning curves as a function of the current target (Fig. 3.2b), or population-averaged tuning surfaces as a function of both previous and current target directions (Fig. 3.5b).

Population response curves

For some analyses (Fig. 3.6a and 3.6b) we determine the "*population response curve*". Whereas a tuning curve describes how one cell responds to a range of different target locations, the population response curve describes how each cell in the population responds to one particular target location. Rather than constructing these curves based on the location of the target in two or even three spatial dimensions, we reduced the dimensionality of the problem by considering only targets arrayed in a circle about the fixation point. Cells are ordered by the location of their receptive field centers. As with the target location, we considered only a single dimension of receptive field centers, arrayed in a circle about the fixation point. The simplest form of a population response is a curve composed of points (x,y), in which x is the receptive field center of a cell, and y is the firing rate of that cell in response to a target at location C. C is held fixed for all cells in a given response curve.

Rather than trying to record from a population of cells that cover all possible receptive field locations, we record from 88 cells and apply a simplifying assumption. We assume that receptive fields of all cells in the memory circuit have the same shape, and construct a single (population-averaged) tuning curve from all 88 cells. In other words, we assume that any cell's response depends only on the distance of the current target from that cell's receptive field center. Next, we find the response of a cell with a receptive field center at location D to a target at location C. This is just the firing rate of the population-averaged tuning curve at x = C-D. To generate the population response curve for a target C, we repeat this last step, iterating through all possible values of D.

In order to take into account not just the current target C but also the previous target P, we extended this method to an additional dimension. We first build the population-averaged tuning surface (Fig. 3.5b) over the domain of previous and current target locations (P and C, respectively). To relate the population-averaged tuning surface to the population response curve, we again make the simplifying assumption that each cell in the memory circuit has a tuning surface for previous and current target locations that is identical to the population-averaged tuning surface. Different receptive field center locations differ relative to both current and previous target locations. Therefore, on average, cells with a receptive field center at D degrees in visual space would respond in trials with a current target at C and previous target at P with a firing rate determined by the distance between their receptive field center and both the previous and current target locations. This point is defined by the coordinates x = (C – D), y = (P – D) on the population-averaged tuning surface. To construct the population response curve for a current target C and a previous target P, we repeat this last step, iterating through all possible values of C. This set of responses defines a slice with a slope of +1 through the surface of Fig. 3.5b. Figures 3.6a and 3.6b show such slices, representing population response curves for particular combinations of current and previous targets. See Supplementary Fig. S3.1 for additional details.

Neural effects of previous target

We investigate two possible (and non-exclusive) effects of the previous target on the current trial – residual memory activity, which we call a ghost, and a shift in the center of neuronal activity representing the current target, which we call a shift. To compute

the normalized ghost amplitude at any particular point in time, we calculate the firing rate on trials in which the previous target was close to (within 22.5 deg of) the receptive field, and subtract as a baseline the firing rate on trials in which the previous target was far away (greater than 112.5 deg) from the receptive field. In addition, to avoid contamination from the shift effect, only trials in which the current target was more than 90 deg from the previous target were used. The ghost response at each point in time was normalized to the maximum response to a current target (i.e., the response to a target centered in the receptive field) at that same time interval (e.g., Fig. 3.7a). We do not show normalized data prior to 500 ms after target presentation, since normalization of cells with late responses results in unstable results in this period.

We also calculated the shift in the response to the current target. When the population response curve shifts *away* from a previous target location, neuronal tuning curves shift *toward* that location. For a clockwise (counterclockwise) tuning curve shift, firing rate clockwise (counterclockwise) from the preferred direction will be elevated compared to firing rate counterclockwise (clockwise) from the preferred direction. To compute the amount of shift at a particular point in time we computed the firing rate difference 20 to 70 deg from the receptive field center on the same side as and opposite side of the previous target. The firing rate difference was then converted into a shift amount in degrees. To accomplish this, we shift all trials by +S and -S deg relative to the preferred direction. We then calculate the firing rate difference on the same and opposite sides of the tuning curve flanks for each amount of shift, 2S (the distance between +S and -S). This gives the expected firing rate difference for a shift of 2S deg. We can then use this procedure to map firing rate differences to corresponding amounts of shift in degrees.

Positive values indicate a population response curve shift away from the previous target location. We computed the shift over time by calculating the shift quantity at 1 ms intervals from 500 ms after target onset until the end of the memory delay. This is shown for trials in which previous and current targets were close together (Fig. 3.7b) or far apart (Fig. 3.7c).

Population vector readout

To decode neural memory activity, we used a population vector readout (Georgopoulos, 1988) of population activity bumps, as described by the equation:

$$P = \sum_i a_i R_i$$

Where a_i is the normalized activity of a cell with receptive field center at R_i and P is the decoded remembered location.

Converging receptive fields model

We modeled a network of neurons that uniformly cover a visual space 100 x 100 units. Receptive fields were modeled as two dimensional Gaussians of the form

$$R(x, y) = e^{\frac{-[(x-\mu_x)^2+(y-\mu_y)^2]}{2\sigma^2}} \qquad (Eqn. 2)$$

where μ_x and μ_y are the coordinates of the receptive field center and σ is the standard deviation. A sigma of 8 was used for our simulations.

Zirnsak et al. (2014) showed that receptive fields of neurons in FEF converge toward saccade targets. To simulate this, receptive fields in the model converged toward target locations by a fraction of their distance, *c*, from the target location. This quantity was scaled so that cells with receptive fields close to the target location converged by a larger proportion of their distance than cells that were far away. We defined the sigmoid function by which *c* was scaled as:

$$s(d) = \frac{1}{1 + e^{-a(d-b)}}$$

where a defines the slope of the sigmoid, b is the value of x at the function's half-height, and d is the distance between the receptive field center at coordinates (R_x, R_y) and the target location at coordinates (T_x, T_y):

$$d = \sqrt{(T_x - R_x)^2 + \left(T_y - R_y\right)^2}$$

Therefore, the total movement of each receptive field R toward a target location is:

$$\vec{M} = M_x + M_y \qquad (Eqn. 3)$$

where the x-component M_x is defined as:

$$M_x = \frac{c(T_x - R_x)}{1 + e^{-a(d-b)}} \qquad (Eqn. 4)$$

and the y-component is similarly defined.

Qualitatively, receptive fields near the target move closer to it while receptive fields far from the target do not move (Figure 3.8a). In our simulations we set a = - *0.05* and b = *20*.

We found that convergence to the current target alone did not explain our neuronal data. However, when we also include a small amount of persistent convergence toward the previous trial's target, we find that the model precisely replicates our behavioral and neuronal data. With both previous and current target convergence the total receptive field movement is given by:

$$\vec{M}_T = \vec{M}_P + \vec{M}_C \qquad (Eqn. 5)$$

where $\vec{M}_P$ and $\vec{M}_C$, are the movement vectors toward the previous and current target (respectively) and determined from Eqns 3 and 4. In particular, we set the convergence amount c to 0.6 for convergence toward current target and 0.2 for convergence toward the previous target.

In our model, receptive fields in both the memory circuit and the readout circuit converge toward the target location in the same way.

To simulate the task we presented previous and current target combinations on a circle with radius of 15 units in our visual space. We then used a population vector readout that accounted for receptive field changes to determine the behavioral response predicted by the activity of the circuit. Finally, we calculated behavioral bias as a function of the relative distance between the previous and current target locations

predicted by the model (Figure 3.8d) using the same analysis as the actual behavioral data (Figure 3.1c).

3.4 Results

3.4.1 Response bias toward the previous target

Spatial memory responses are biased toward previously memorized locations in a memory-guided saccade task (Papadimitriou et al., 2015). In this study we look for neural correlates of this bias in frontal memory circuits. We first replicated the basic behavioral finding. Three macaques made saccades to memorized locations after delays of 1.4, 2.8 or 5.6 s (Figure 3.1a). Saccade endpoint error increases with delay (Fig. 3.1b). We defined trial-by-trial response error as the total error minus the mean of the error obtained for that particular target position (see *Experimental Procedures*). A plot of response error as a function of the distance between the previous and current target location reveals a systematic behavioral bias toward the memory location of the previous trial (Fig. 3.1c). The bias can be well fit by a Gabor function (peak-to-peak height = 1.13 deg, fit $p < 0.005$) and was significant in each of the three individual animals (peak-to-peak height = 1.8, 1.3, and 0.95 deg for monkey H, J, and P respectively; $p < 0.005$ for all fits).

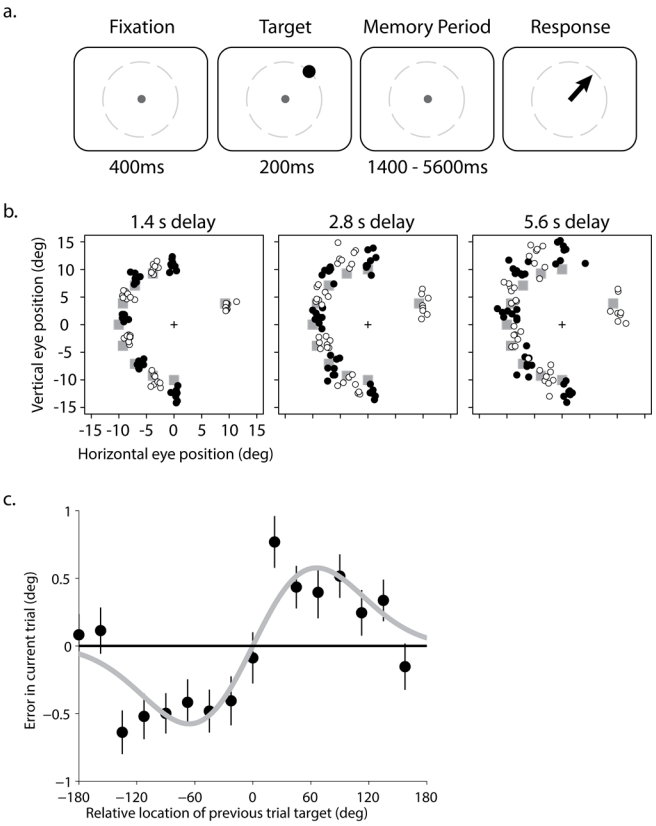


Figure 3.1. Behavioral task and responses.

a. Memory-guided saccade task. Subjects fixated on a central target presented at the center of the screen. After a fixation period of 400ms a memory target was displayed for 200ms at one of 16 possible peripheral locations at a fixed eccentricity. Target presentation was followed by a memory period between 1.4s to 5.6s in duration during which the subject continued to fixate. After the memory the fixation target disappeared and the subject responded by making a saccade to the remembered location. Dashed gray line indicates where targets might appear; it was not visible to the animal. **b.** Saccade endpoints in a subset of representative trials. Grey squares represent target locations. Responses have been colored black or white to more easily identify the associated memory target. **c.** Error in current trial response as a function of previous target location relative to current target location. When the previous target was clockwise from the current target (negative x-axis) the saccadic response was biased clockwise from the current target (negative y-axis) and vice versa. The gray line is the Gabor fit to the raw data (peak-to-peak height = 1.13, fit $p < 0.005$).

3.4.2 Neuronal responses in frontal eye fields

We looked for neural correlates of the behavioral response bias in the frontal eye fields (FEF). We recorded from 88 neurons in FEF while monkeys performed the memory-guided saccade task. Cells with sustained memory period activity were selected using a memory screening task (see *Experimental Procedures*). Figure 3.2a shows population-averaged firing rate as a function of time. Activity was higher when a memory target was presented at the center (red trace) or flank (orange) of each receptive field, as compared to when the target was presented outside the receptive fields (green). The elevated activity persisted for the duration of the memory period and is well-fit by a Von Mises function (Fig. 3.2b).

FEF reflects saccade endpoints as well as target locations. This can be seen by contrasting the population-averaged activity across all recorded cells when the memory-guided saccade lands either clockwise (> 12.5 deg; mean = 16.2 deg) or counterclockwise (< -12.5 deg; mean = -16.1 deg) of the memory target (Fig. 3.3a, red versus blue trace). The tuning curves are constructed based on activity recorded in the 500 ms immediately prior to the go cue. If FEF encoded only target at that time interval, the two curves would perfectly overlap. If FEF encoded only saccade endpoint, the curves would be separated from one another by 32.3 deg, the difference in the means of the saccade endpoints used to construct each curve. In fact, they are separated by 25.3 deg. Fig. 3.3b shows similar data from 7 different bins of saccade error. In each case, the vector sum readout of the data (a prediction of saccade error if saccade endpoint is encoded; see *Experimental Procedures*) is plotted as a function of the actual error in the saccade endpoints. A linear fit with a slope of 0 would indicate that FEF encodes only the target location. A slope of 1 (dashed line) would indicate coding of only the saccade endpoint. The actual

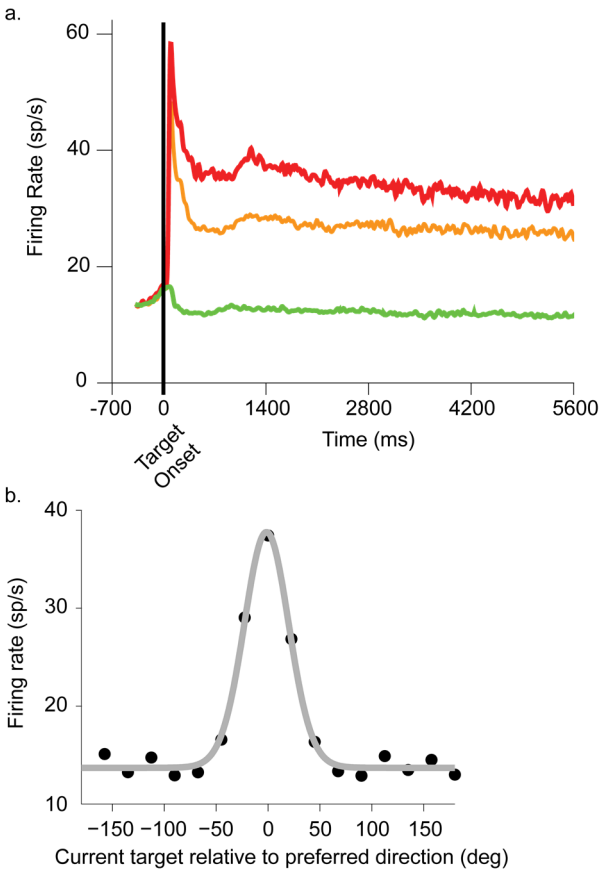


Figure 3.2. Tuned and sustained memory responses in FEF neurons.

a. FEF population response when the memory location was presented at the center of cells' receptive fields (red trace), 22.5 degrees from the receptive fields (orange trace) or 180 degrees away from the receptive fields (green trace). Firing rates when the target was presented in the receptive field stay high after target offset and for the duration of the memory period. **b.** Firing rate as a function of target location relative to the receptive field interval is well fit by a von Mises function (500 to 1500 ms, $p<0.005$).

slope (solid line) is intermediate (0.70 +/- 0.24 deg/deg, $p < 0.005$), indicating that, in the final 500 ms prior to the go cue, FEF neurons encode a location that is closer to the saccade endpoint than to the target.

Figure 3.3c shows how this measure changes over time. At the start of the trial (50 ms to 300 ms after target onset), the neural activity encodes target location independent of saccade error (slope = 0.01 +/- 0.17 deg/deg, $p = 0.97$). Early memory period activity (350 ms to 750 ms after target onset) is influenced by (or influences) the saccade

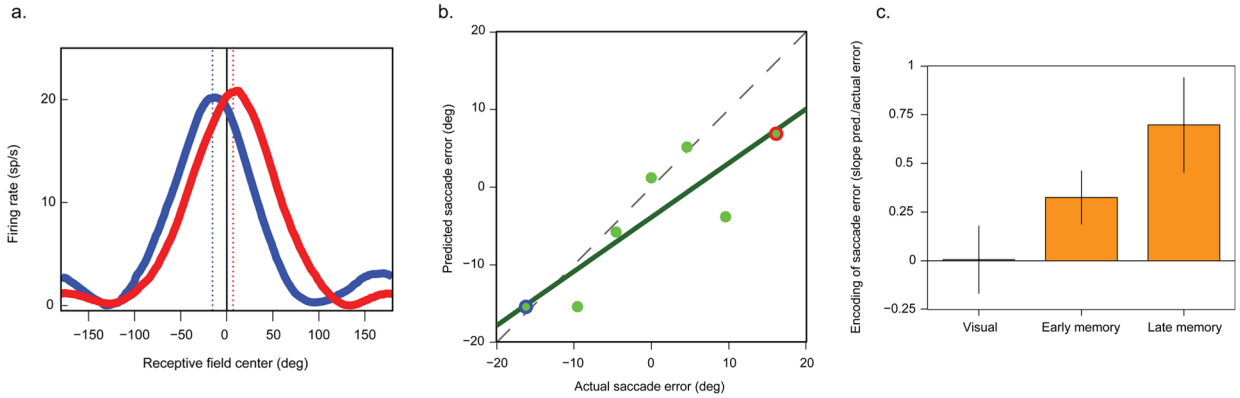


Figure 3.3. Neural activity reflects behavioral responses

a. Population activity for trials when response error was greater than 12.5 deg. (mean = 16.2; red trace) and less than -12.5 deg. (mean = -16.1 deg; blue trace). The dotted lines show the encoded location determined with population vector decoding (red trace, 6.9 deg; blue trace, -15.4 deg). **b.** Linear regression of saccade error predicted by neural activity and response error observed behaviorally for the time interval 500 ms to 0 ms prior to the go-cue. Trials are binned by observed response error (-10 deg to 10 deg, steps of 5 deg, bin width of 5 deg). We also included two bins with response error greater than 12.5 deg (mean = 16.2) and less than -12.5 deg (mean = -16.1 deg). The data points outlined in blue and red correspond to the curves in panel a. The regression line (green) has a slope of 0.70 deg / deg (regression $p < 0.015$). The dashed line shows a slope of one. **c.** Linear regression slopes for the visual period (50 ms to 300 ms after target onset; slope = 0.006 deg / deg; $p = 0.97$), early memory (350 ms to 750 ms after target onset; slope = 0.33 deg/ deg; $p < 0.02$), and late memory (-500 ms to 0 ms prior to go-cue; slope = 0.70 deg / deg; $p < 0.005$).

endpoint (slope = 0.30 +/- 0.14 deg / deg, $p < 0.02$), and in the final 500 ms, the effect is twice as strong (slope = 0.70 +/- 0.24 deg / deg, $p < 0.005$). The early and late slopes are significantly different from each another ($p < 0.05$). Thus, while FEF activity initially encodes the location of the memory target, it becomes more closely linked to the endpoint of the upcoming memory-guided saccade as the trial progresses. This suggests

that if the bias related to the target position from the previous trial (Fig. 3.1c) is produced either within or upstream of FEF, then this bias will be able to be read out from the FEF neurons (see also Wimmer et al., 2014).

3.4.3 Residual memory trace in the ITI and subsequent trial

Papadimitriou et al. (2015) modeled the bias from the previous target using a combination of a long-term and a short-term store. Only the long-term store, modeled as a bump attractor, is biased by previous target position. A simple way to produce this bias is for a remnant of activity encoding the previous target to persist into the subsequent trial. The attractor dynamics may then merge this remnant from the previous trial's target with the "bump" encoding the current target. The merger would result in a single bump of activity, encoding a location intermediate between the current and previous target. This would manifest in the behavior as a bias toward the location encoded in the previous trial. To test this hypothesis, we looked for evidence of a remnant or "ghost" of the previous target during the fixation period, after the animal had successfully completed the previous memory trial and returned to the fixation point, but before the target for the current trial had appeared.

We plotted firing rate during the fixation period, prior to target onset, as a function of the previous trial's target position relative to the receptive field. The elevation in firing rate seen on the previous trial when the target was in the receptive field (Fig. 3.2a) persisted, in an attenuated form, into the subsequent trial's fixation interval. Figure 3.4a shows data from an example cell. The firing rate is normalized to the activity recorded 50-300 ms after target presentation. The ghost activity in the fixation period

is about one-quarter as large as the previous visually-evoked activity from that same target. Even though the previous trial has ended, the cell shows clear tuning to the previous target location and is well fit by a Von Mises function ($p < 0.0001$). Of 88 cells, 49 showed significant ($p < 0.05$) tuning to the previous target location during the fixation interval and only 8 showed significant tuning for a location opposite to the previous target (Figure 3.4b). Figure 3.4c shows population-averaged firing rates, similar to Fig. 3.2c but sorted by the target location from the *previous* trial. The figure confirms that, within the fixation period that separates the end of one trial from the start of the next, there is a ghost of the previous trial's memory activity. The population-averaged effect across all cells is 5.41 +/- 0.95 sp/s ($p < 0.0001$), which is 32% of the activity 500 ms to 0 ms before the end of the previous memory period (Fig. 3.2c, far right). The ghost disappears abruptly once the next target is presented. The response to target onset shows no tuning, that is, the red, orange and green traces overlay one another shortly after the vertical line at time zero in Fig. 3.4c. To some extent this is to be expected, since the traces are sorted on the previous target position, and previous and current target positions are completely independent of one another. However, it is contrary to the model of Papadimitriou et al. (2015). This model predicted that the residual ghost would merge with and shift the current trial's bump, preserving a small bias in firing related to the previous trial's target position such that the red trace would remain slightly higher than the green trace. There is no evidence for this in Fig. 3.4c;

the firing rate difference between the red and green trace is -0.21 +/- 0.41 sp/s (p = 0.613) in the interval from 200 ms to 1400 ms after target presentation.

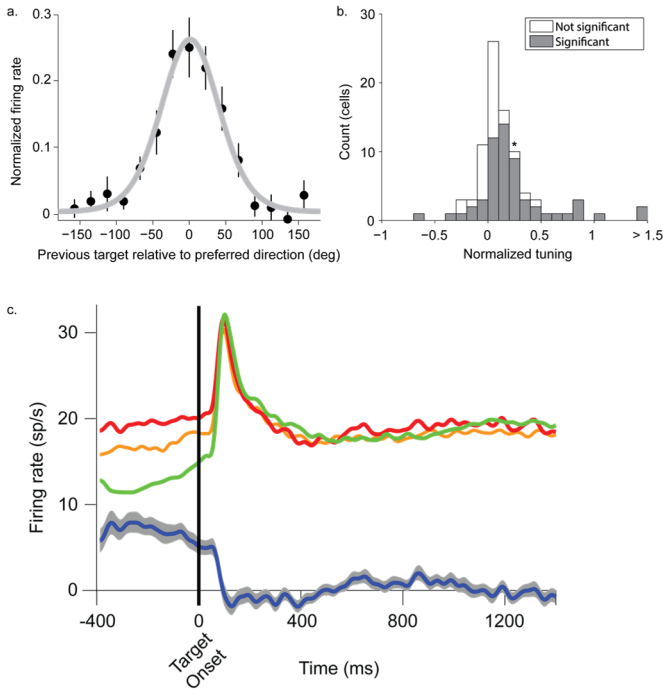


Figure 3.4. Residual memory tuning from the previous trial.

a. Firing rate of an example cell during fixation as a function of target location on the previous trial. Firing rate is scaled to the tuning amplitude 50ms to 300ms after target onset and the baseline is removed. **b.** Histogram of normalized tuning to the previous target in the current trial fixation period (as in panel a) for the cell population. Grey bars indicate cells that show a significant difference in firing rates for previous targets in or out of their receptive fields ($p < 0.05$) and white bars indicate cells that did not show a significant difference. The asterisk indicates the bin that includes the example cell in panel a. **c.** Population firing rate when the previous target was presented at the center (red trace), 22.5 degrees away (orange trace) or 180 degrees away from the cells' receptive fields (green trace). The blue trace shows the difference between the red and green traces.

3.4.4 Influence of previous target on neural activity

The influence of the previous trial on *behavior* is weak for previous targets far from the current target, and strongest when the previous and current memory targets are about 60 deg from one another (Fig. 3.1c). This is consistent with attractor models, in which broad inhibition quickly quashes activity that is far from the dominant bump, with little effect on the dominant bump itself. Only residual activity near the dominant bump would be expected to exert an influence. The manifestation of the ghost of the previous trial might therefore depend on how far away it is from the current target. To test this idea, we constructed a population-averaged tuning surface as a function of current and

previous target locations. In order to combine data across all recorded cells, we expressed target locations relative to the center of each cell's receptive field.

Figure 3.5 shows the resulting population tuning surfaces for the fixation and early memory periods. The activity ghost appears in the fixation period (panel a) as a horizontal band at y=0, that is, on trials in which the previous target is aligned with the receptive field. During the memory period (panel b), activity is dominated by the current target. This is indicated by the vertical band at x=0, reflecting trials in which the current target is aligned with the receptive field and therefore evokes a large response. There is also a faint but persistent ghost of the previous target in the memory period. The ghost is smaller in amplitude than in the fixation period, and appears only when the previous and current targets are more than about 100 deg apart, that is, points defined by the locus of $y = 0$ deg and $x < 100$ and $x > -100$ deg. These loci are indicated by the green circles. The pattern is precisely the opposite of what we predicted from the behavioral data. Instead of the ghost being most obvious in cases in which the previous and current targets are close together ($x = \pm 60$ deg, magenta ovals), the ghost is instead most obvious when the previous and current targets are far apart (green circles).

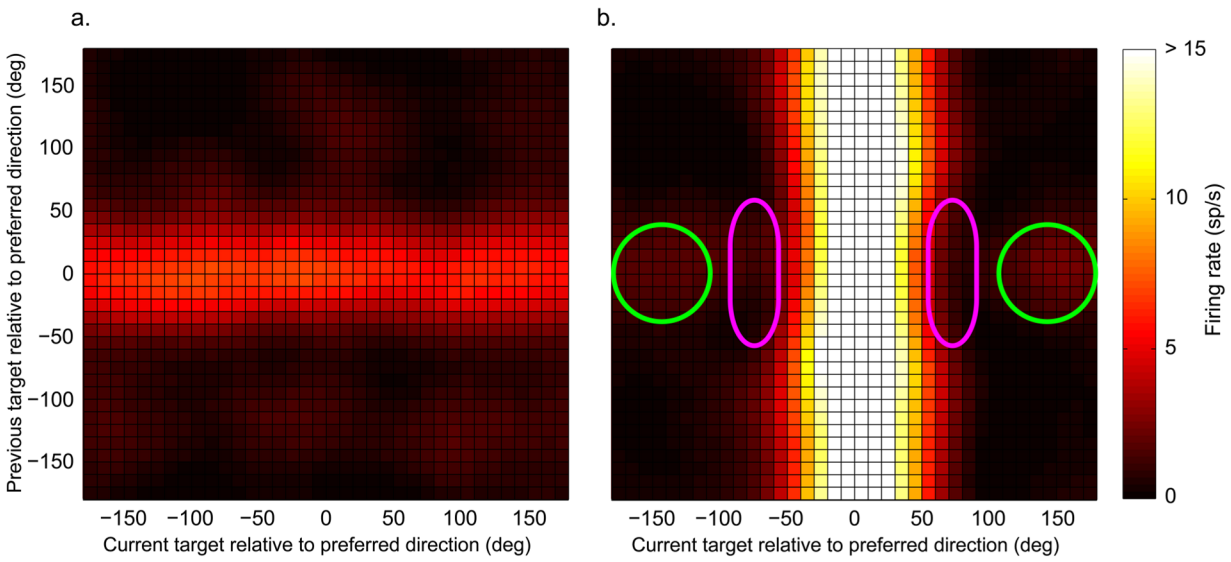


Figure 3.5. Two dimensional population tuning curve of firing rate as a function of previous and current target location.

In both panels, preferred direction of each unit has been rotated to 0 degrees. **a.** Neural activity as a function of previous and current target location during the fixation period of the current trial -375ms to -175ms prior to current target onset. Fixation period activity is elevated when the previous target was in the preferred direction (y = 0 degrees). **b.** Activity during the memory period 1000ms to 1500ms after target onset. Activity is high when the current target is in the receptive field (x = 0 degrees). Smaller but clear activity elevation is evident when the previous target was in the preferred direction (y = 0 degrees) and the current target is away from the preferred direction (x > 90 degrees).

We can use the tuning surface of Figure 3.5b to determine how each cell in the population will respond for any combination of previous and current target locations. To capture cells with receptive field centers at all possible locations for a specific combination of current and previous target position, we must take a slice through the surface with slope of +1. As an example, consider a trial in which the (current) target is 130 degrees counterclockwise to the previous target. The relevant points on the surface are those for which the previous target direction (expressed relative to the direction of the receptive field center of each cell, or preferred direction, which ranges from -180 to +180 deg) equals the current target direction (expressed relative to the preferred direction) minus 130 deg, or y = x - 130. Thus the population response to any combination of current and previous target locations is described by a line with a slope

of +1. When the current and previous targets coincide, this line runs from the bottom left to the top right. For all other cases, the line starts on the far left, ascends to the top of the plot, wraps around to the bottom, and then continues on up again. This results in two parallel line segments. For the particular example of a current target 130 deg counterclockwise to the previous target, the locus of points forms two line segments, one from (x=-180, y=-50) to (x=+50, y=180) and the other from (x=+50, y=-180) to (x=180, y=-50). See *Experimental Procedures* and Supplementary Figure S3.1 for details.

Figure 3.6a contrasts two slices through the tuning surface of Fig. 3.5b. The slices represent the conditions when the previous target was 130 deg away from the current target in either a clockwise (blue) or counterclockwise (red) direction. As in Fig. 3.5b, the most prominent feature is a large bump at 0 deg, representing the response to the current target. A much smaller bump is present in each curve at the location of the previous target, corresponding to the slightly elevated firing previously noted in the tuning surface of Fig. 3.5b (green circles), close to the y=0 line – the ghost of the previous trial's bump. To quantify this effect we took all trials in which the previous and current targets were separated by 90 to 170 degrees, clockwise or counterclockwise, and measured the difference in firing at the previous target location, that is, the separation between the blue and red lines at the dashed lines. (For target separations less than 90 deg, the effect is different – see next paragraph. For separations approaching 180 deg, the red and blue lines must converge.) The mean separation, that is, the height of the ghost, was 3.04 +/- 1.2 sp/s ($p < 0.01$).

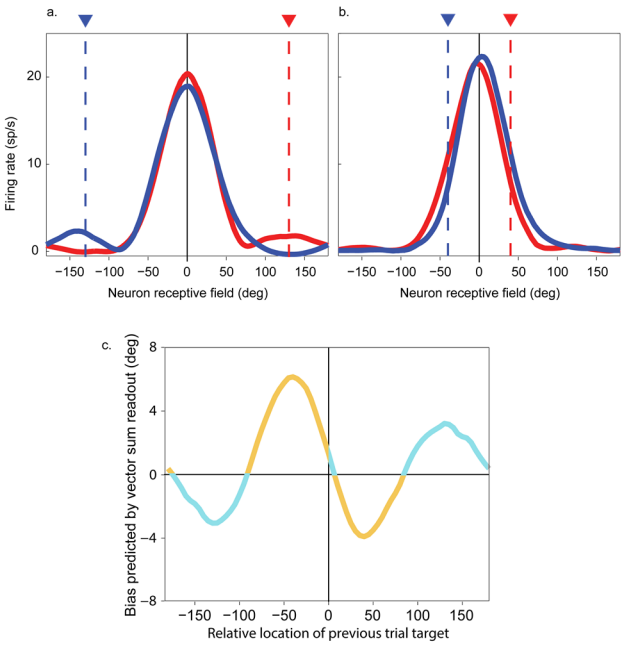


Figure 3.6. Population response curves and behavioral readout.

a. When the previous target was at 130 or -130 degrees (red and blue triangle, respectively) activity in neurons with preferred direction near 130 or -130 degrees, (red and blue traces respectively), is elevated in the current trial. **b.** When the previous target was at 40 or -40 degrees (red and blue triangle, respectively) activity in neurons with preferred direction near 40 or -40 degrees, (red and blue traces respectively), is reduced in the current trial. **c.** Population vector readout of FEF activity in the interval 1000 ms to 1500 ms after target onset. When the previous and current targets are close together (e.g., panel b), the readout predicts a repulsive bias (yellow) away from the previous target location in the behavioral response. When the previous and current targets are far apart (e.g., panel a), the readout predicts an attractive bias (blue) toward the previous target location.

In Figure 3.6b, we show a similar plot as in panel A, but now representing the condition in which the previous target was close to the current target — 40 deg clockwise (blue) or counterclockwise (red). Once again, the prominent feature is a large bump at 0 deg, representing the current target position. The behavioral data and the attractor model both suggest that attractive bias will be strongest when the previous and current targets are close together. This leads to a prediction of a large ghost. Instead, at the location corresponding to the previous target (the dashed vertical lines), the blue trace is above the red on the right and below the red on the left. This is exactly the opposite of the pattern in Fig. 3.6a. We quantified this by taking all trials in which the previous and current targets were separated by up to +/- 80 degrees and measured the difference in firing. The mean difference was -6.05 +/- 1.0 sp/s ($p < 0.0001$). The negative sign means that the effect of a nearby previous target was to *lower* firing rate in the subsequent trial.

This effect can either be a suppression at the location of the previous target (a negative ghost), or a shift of the current target representation in a direction away from the previous target location. A negative ghost would be a localized phenomenon with no other associated changes in firing. A shift, in contrast, would be associated with an increase in firing on the other side of the current target activity bump. We analyzed this and found that the effect is best described as a shift away from the previous target location (Supplemental Figure S3.2). Yet both the behavior and the model attractor network predict a shift *towards* the previous target location. Thus these results are inconsistent with our predictions.

3.4.5 Readout

We generalized these results across a full range of target positions. We generated a family of population activity curves like those of Fig. 3.6a and 3.6b, for all relative target positions and for several different time points, and used a population vector method to read out the location encoded by the activity. We hypothesized that a systematic error or bias in the readout would match the behavioral bias seen in Fig. 3.1c. In order to test this hypothesis, we plotted the predicted bias in saccade endpoint (actual target location minus the neuronal readout of activity 1000 to 1500 ms into the memory period) as a function of the distance between the previous and current target locations (Figure 3.6c). The plot provides a prediction of the behavioral bias we might expect to see in memory-guided saccades after a 1000 to 1500 ms memory period, based only on FEF activity. (Given that the influence of FEF activity on saccade endpoint is only 70% complete (Fig.

3.3b), this plot may overestimate the magnitude of the effect, but correctly captures the sign – attraction versus repulsion.)

The results do not match our expectations. When the previous and current target are far apart (greater than 90 deg, as in Fig. 3.6a), the ghost from the previous trial biases the population readout in an attractive direction, such that the predicted bias has the same sign as the relative location of the previous target (Fig. 3.6c, blue sections of the trace). This attractive bias matches the attractive bias that is observed in the behavior (Fig. 3.1c). However, the predicted peak attraction occurs at 130 deg, whereas the peak attraction in the behavior occurs at 60 deg. When the previous and current targets are close together (less than 90 deg, as in Fig. 3.6b), the shift in FEF activity away from the location of the previous target biases the population readout in a repulsive direction, such that the bias and previous target locations have opposite signs (Fig. 3.6c, orange section of the trace). This repulsion is opposite to the attractive bias that is observed in the behavior (Fig. 3.1c). Thus, although FEF memory circuits show clear previous trial effects, a straightforward readout of the activity does not match the observed behavior.

3.4.6 Previous target effects over time

We now turn from the spatial pattern of the previous target effect to the temporal pattern. The magnitude of the attractive behavioral bias grows over the first several seconds of the delay period (Papadimitriou et al., 2015). Papadimitriou et al. modeled this by proposing that the behavior is driven by two independent stores working in parallel: a rapidly decaying but veridical visual sensory store and a sustained but distorted working memory store. The sustained store has a constant bias, present from

the very start of the trial. This store has no information about the veridical target location and therefore has no way to correct its bias. The behavior relies on a weighted average of the two stores. Initially the unbiased visual store has a high amplitude and so early responses are nearly veridical. However, the visual store decays rapidly. After several seconds the output is driven almost entirely by the sustained store, and so becomes biased. Thus the model predicts that FEF, the putative sustained store, will be biased from the very start of each trial and that this bias will persist over time.

Figure 3.7a shows the time course of the normalized height of the residual ghost. The ghost is present at the start of trials in which the previous and current targets are far apart, with a normalized amplitude (relative to the response to a visual target) of ~20%. However, the ghost disappears rapidly. This does not match the time course of the behavioral bias, which persists for over 5 s without attenuation (Papadimitriou et al., 2015; Supplementary Figure S3.3). Thus neither the temporal nor the spatial aspects of the ghost match the observed behavioral bias.

Attractor network models predict that a residual ghost of activity from the previous trial will merge with the representation of the current trial's target, shifting the current target representation towards the previous target location. This could conceivably explain the rapid disappearance of the ghost. In this case, the bias would manifest as an attractive shift of the current target representation, starting as the ghost disappears and persisting to the end of the trial. Figure 3.7b shows that this was not the case; the early

disappearance of the ghost was not accompanied by an attractive shift of the target representation.

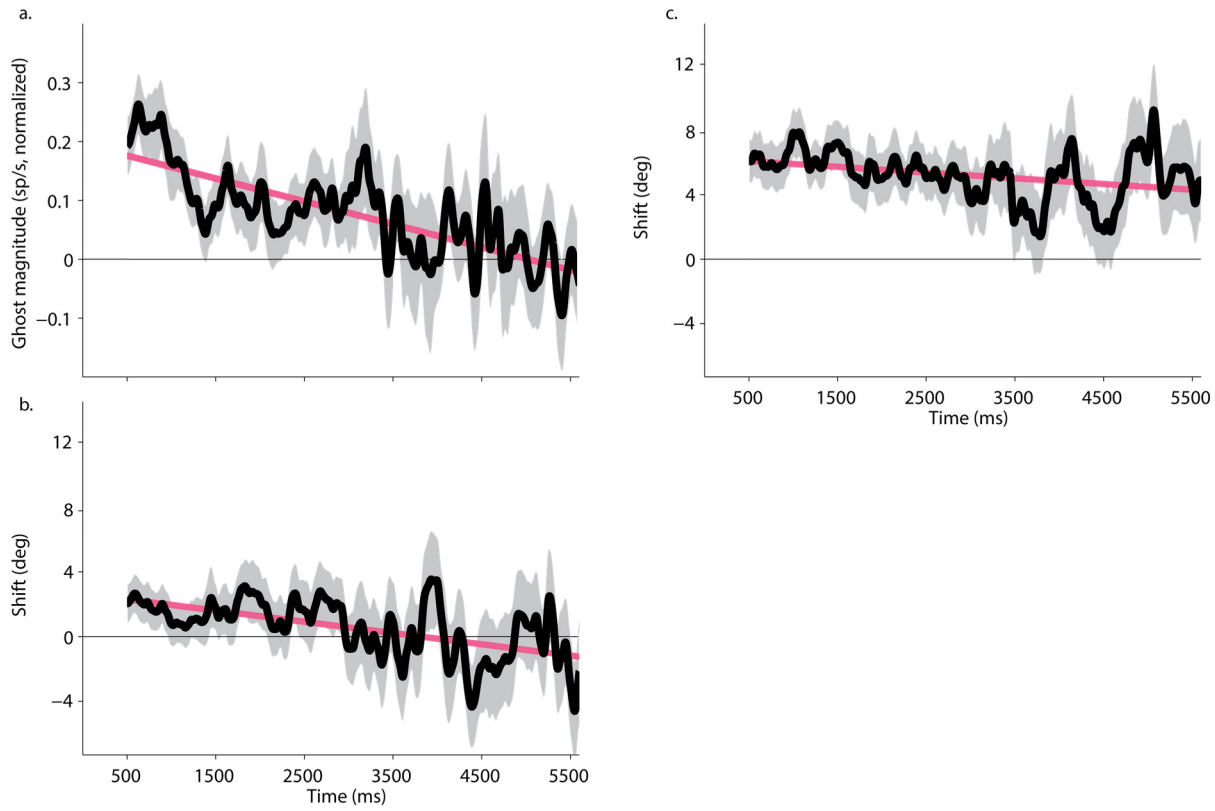


Figure 3.7. Previous trial effects on current trial over time.

a. The ghost (the residual activity encoding the previous target) is not sustained through the delay period. Ghost amplitude, measured for previous and current targets that are far from one another (more than 90 deg apart), is initially ~20% as large as the visually evoked response, but decreases with a slope of 4 % per second. It disappears entirely after 3.5 seconds (mean effect 3.5 to 5.6 s after target onset = 0.02 +/- 0.05 %, p = 0.736). **b.** The disappearance of the ghost in panel 'a' is not accompanied by a shift of the tuning curve center (mean shift = -0.57 +/-1.8 deg, p = 0.737). **c.** The shift in the current target representation, measured for previous and current targets that are close together (less than 90 deg apart), is largely sustained (slope = -0.24 deg / sec). It remains highly significant even at the end of the delay period (mean effect 3.5 to 5.6 s after target onset = 4.38 +/- 1.26 deg, p < 0.002). In all panels, red lines are linear fits.

In contrast, Figure 3.7c shows that, when the previous and current targets were close together, a strong repulsive shift was present throughout the entire delay period. As previously noted, this shift appears to be in the wrong direction to produce an attractive bias. While the sign of the effect is reversed, the spatial profile (the relative locations of

previous and current target at which the maximum effect occurs) and the temporal aspects of the shift are consistent with the observed behavior.

In summary, we observe two distinct neuronal effects of the previous target. Neither effect provides a good match to the behavior. Ghost activity has the right sign to produce an attractive bias. However, neither its spatial nor temporal properties match those of the behavior. In particular, when the current and previous targets are separated by 60 deg we find the maximum behavioral effect, but under these conditions the ghost disappears as soon as the target appears. When the current and previous targets are far apart, we find no behavioral effect, yet the ghost persists for 3.5 s after the target appearance. Thus the ghost activity does not match the spatial and temporal patterns of the behavioral bias. In contrast, the neuronal shift effect shows a better, but still incomplete, match to the behavior. In particular, both the behavioral bias and the neuronal shift occur only when a target appears close to the location of the previous target, and both persist for the entire delay period. However, the neuronal shift predicts a repulsive bias, while the behavior shows an attractive bias. Thus, neither the ghost nor the shift can explain the behavioral bias.

One way to reconcile the repulsive shift predicted by neural activity with the attractive bias observed behaviorally would be if subjects fixate at a location biased toward the previous target. FEF encodes the relative change in eye position (the saccade vector) and not the absolute saccade endpoint. A large enough displacement in initial fixation position could result in a saccade vector that is biased away from the previous target, even while the saccade endpoint is biased toward the previous target (Supplementary

Figure S3.4a). To address this possibility we recomputed the behavioral effect, taking into account the subject's eye position immediately prior to saccade onset (-200ms to 0ms). We found that saccade vectors calculated in this way still were still biased toward the previous target (Supplementary Fig. S3.4b), indicating that small differences in fixation location cannot account for the discrepancy between neural activity and behavior.

3.4.7 Proposed model to resolve neuronal and behavioral manifestations of bias

We next consider whether the phenomenon of shifting receptive fields might help explain the neuronal-behavioral discrepancy. Neurons in some areas involved in visual, oculomotor and mnemonic processing appear to temporarily shift their receptive fields towards the goal of an upcoming saccade or attended location (Connor et al., 1997; Tolias et al., 2001; Zirnsak et al., 2014). The shifts we observe could reflect temporary shifts in receptive field locations. Previous studies have also revealed systematic mislocalizations of stimuli that occur under the same circumstances as receptive field shifts (Hamker et al., 2008; Ross et al., 1997, 2001). These earlier findings led us to hypothesize that the two effects that we observed – shifts in tuning curves during a memory period and behavioral mislocalizations of remembered targets – might be explained if receptive fields converge toward remembered locations and some fraction of that convergence persists across trials.

To test this idea, we simulated a network of memory neurons with receptive fields uniformly tiling visual space. When a memory target is presented to the network, neurons shift their receptive fields toward the target with an amplitude that is

proportional to their distance from the target, multiplied by a sigmoid. The multiplication by a sigmoid confines the shifting to the vicinity of the target; receptive fields far from the target do not shift. Figure 3.8a shows the resulting shifts. The starting points of the depicted vectors represent the original receptive field centers, and the endpoints represent the final shifted position due to stimulus presentation. By construction, there is strong convergence towards the current target (red; both left and right panels) and a weak convergence towards the previous target (right panel; blue).

These receptive field shifts affect the network readout. Imagine a neuron with a receptive field centered 16 deg to the left of the fovea. If this field shifts 10 deg to the right, its new center will be 6 deg to the left of the fovea. In a vector sum readout, this cell would "vote" for a position 16 deg to the left. After the shift, the cell would respond most strongly to a target appearing at 6 deg left, not 16 deg left. Yet this strong response to a target at 6 deg left would be mistaken as a "vote" for the 16 deg leftward location; the cell would now bias the vector sum readout to the left. In general, a receptive field shift in one direction would shift a vector sum readout in the opposite direction. Note, however, that with just one target, the shifts of the receptive fields across the population are symmetric (Fig. 3.8a, left). As a result, the biases produced by individual cells will exactly cancel one another, producing no net bias in the vector sum readout. The addition of even a small residual shift from the previous trial will break the symmetry and result in a distorted readout (Fig. 3.8a, right). Since shifts bias the readout in the opposite direction from the shift, the distortion will result in a repulsion away from the location of the previous target.

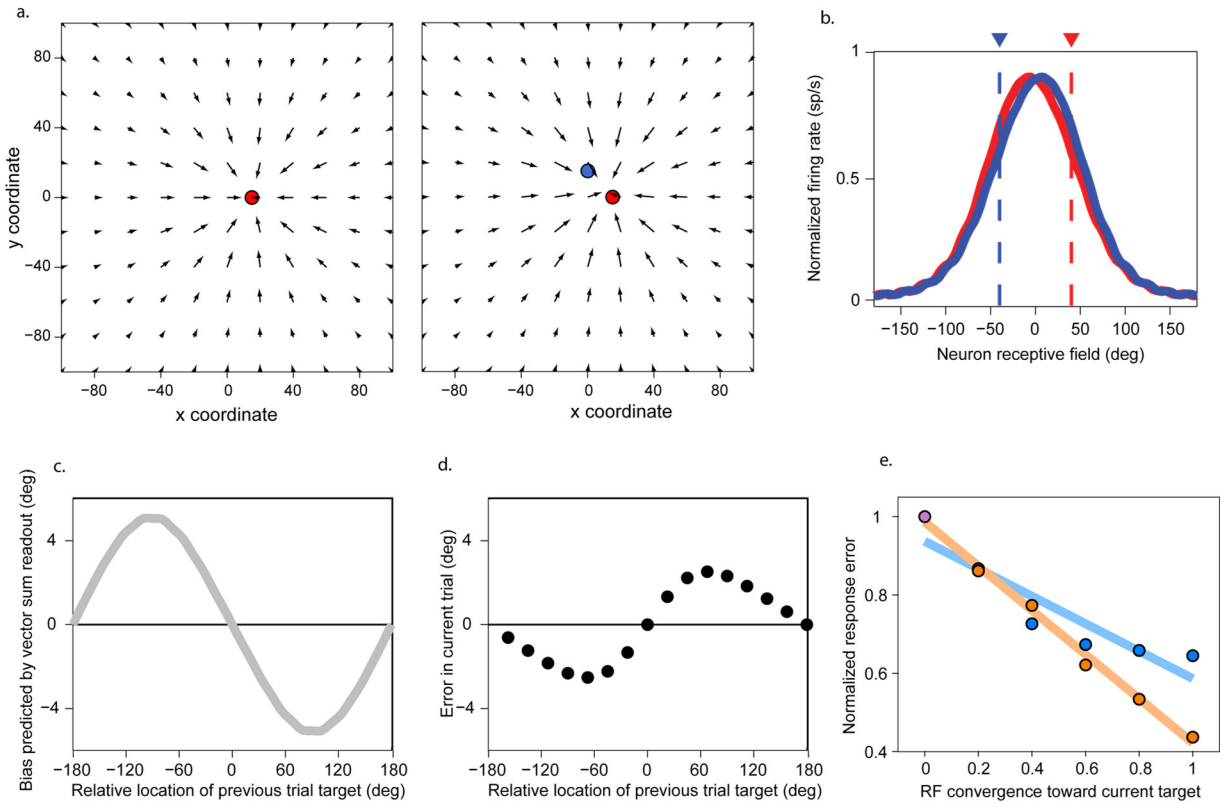


Figure 3.8. Convergence of receptive fields can explain both neuronal activity patterns and behavioral bias. **a.** Left - Example of receptive fields convergence toward the memory target (c = 0.6). Convergence amount is comparable to Zirnsak et al. (2014). Right - Receptive fields converge toward the current memory target (c = 0.6) with some residual convergence toward the previous target (c = 0.2). b. Population response curves appear to move away from the previous target location when receptive fields shift toward both the current and previous target (compare Fig. 6b). c. The population vector sum of population response curves like those in panel b predict repulsive bias (compare Fig. 6c). d. When the readout takes receptive field shifts into account (see text) the distorted distribution of receptive fields produces attractive bias (compare Fig. 1c). e. Response error as a function of RF convergence toward current target (x-axis) in the presence (orange) or absence (blue) of convergence toward the previous target. When noise is added to cell firing rates convergence of receptive fields improves performance, even when some convergence persists into the subsequent trial. See text for additional details.

Our quantitative simulations confirm this qualitative description. We replicated the structure of our task, presenting memory targets along a circle of radius 15 deg from the fixation point. The simulated neurons shifted their receptive field locations during the memory period, as described for area FEF (Zirnsak et al., 2014). In the model, there

was no residual ghost activity from the previous trial. We asked what the effect of the receptive field shifts would be on the neuronal data. Population response curves from the simulated neurons shows repulsive neuronal shifts, just as in the actual data (compare Fig. 3.8b with 3.6b). Vector sum readouts of the simulated network show repulsive biases, matching the repulsive bias in the recorded neuronal actual data (compare Fig. 3.8c with 3.6c). (For simplicity, the simulation did not include residual ghosts when previous and current targets were far apart [Fig. 3.6a]; had this been included, then the small portions of the actual behavioral readout shown in blue in Fig. 3.6c would also have been replicated.) Critically, although the readouts of the simulated network generally match the vector sum readouts of the actual neuronal recordings, neither of these two readouts consistently match the behavioral results, which show an attractive rather than a repulsive bias (Fig. 3.1c).

Next, we asked what the vector sum readout of the simulated network would look like if we assume that the network "knows" about the receptive fields shifts and takes them into account when generating the readout. In this case, a cell whose RF is normally at 16 deg left but which shifts over 10 deg to the right would "vote" for the 6 deg left position, not the 16 deg position. As a result, individual cells show no bias. However, because the population of cells no longer uniformly tiles space, the vector sum becomes biased. In particular, the vector sum is biased towards the location with the most dense accumulation of receptive fields, that is, the point in space towards which the receptive fields are converging. This results in a strong attractive bias towards the current target location, and a weak attractive bias towards the previous target location (Fig. 3.8d). (Note that this mechanism is different from the template-matching algorithm described

by Abott (Abbott, 1994) and used in the model by Zirnsak et al. (2014) The template-matching algorithm produces an attractive bias only if the saccade endpoint is at the center of the array of receptive field centers; for saccades to peripheral locations, the mechanism will produce a repulsive bias.)

A shift in the overall distribution of receptive fields, such that they cluster about a particular stimulus location, can be viewed as a shift in computational or representational resources to that location. For example, increasing the density of neurons that encode a memorized location could serve to make the memory trace more robust and resistant to noise compared to a network that lacked such a shifting mechanism. Our simulation confirms this intuition. We added random noise to each cell's response, prior to computing a vector sum readout. We then calculated the response error for each trial under various conditions of receptive field convergence (Figure 3.8e). Response error was normalized to the error when convergence to both the previous and current target was 0 (purple point in the top left corner). As we varied the degree of convergence towards the current target from no convergence (x=0; no change in receptive field locations) to complete convergence (x=1; all receptive fields are aligned with the target position), the error in the behavioral response decreased linearly (orange points and trace; 0.6% reduction in error per 1% of RF convergence, $p < 0.0001$). This improvement in the behavioral response was reduced but still present even when a fraction of the convergence (0.2) from the previous trial persists into the current trial, as in our model (blue trace, 0.4 % reduction in error per 1 % RF convergence, $p < 0.03$).

3.5 Discussion

Behavioral responses in spatial memory tasks are biased toward the memoranda of the previous trial (Fig. 3.1c). To identify neural correlates of this bias, we recorded from FEF during a memory-guided saccade task. We selected spatially-tuned cells with sustained responses during a memory period (Fig. 3.2). These cells code target location early in the trial and saccade endpoint late in the trial (Fig. 3.3). A small amount of activity persisted after the end of each trial and could be seen in the subsequent fixation period prior to the appearance of the next target (Fig. 3.4). When averaged across all conditions, this residual or ghost activity disappeared as soon as a new target appeared. Given that the behavioral bias depends on the distance between the previous and current targets, we examined the ghost activity as a function of this distance (Fig. 3.5). We found that ghost activity persists during the memory period only when the current and previous targets are separated by more than 90 deg (Fig. 3.6). However, this activity cannot explain the behavioral bias, since the behavioral bias is strongest when the previous target is 60 deg from the current target (spatial mismatch). In addition, even when the target separation was large and the ghost did persist, the ghost lasted only about 3 s, whereas the behavioral bias persisted indefinitely (Fig. 3.7a).

When the previous target appeared within 90 deg of the current target, there was no ghost, but the population activity encoding the current target was shifted in position. However, this shift was directed *away* from the location of the previous target, that is, in a direction opposite that which would be predicted by the behavioral bias (Fig. 3.6b). Unlike the ghost but like the behavioral bias, this shift persisted throughout the

duration of the trial (Fig. 3.7c). In summary, our data show that neural activity in FEF is influenced by prior memoranda, but a conventional readout of this activity (Fig. 3.6c) is not congruent with the observed behavior (Fig. 3.1c). Specifically, when the previous and current targets are close together, the neural activity (the shift in the representation of the current target) predicts that saccades will be repulsed away from the previous target, whereas the observed behavior shows a strong attractive bias. When the previous and current targets are far apart, the neural activity predicts a large attractive bias, whereas the observed behavior shows minimal bias.

To reconcile the neuronal data with the behavioral responses, we propose that receptive fields in FEF shift in response to memory targets, and that the fields do not completely revert back to their original locations after the end of a trial (Fig. 3.8). In a model, a small amount of residual shift exactly reproduces both behavioral and neuronal effects: the memory-guided saccades read out of the model show an attractive bias towards the location of the memoranda of the previous trials, and the activity in the simulated neurons show a repulsive shift.

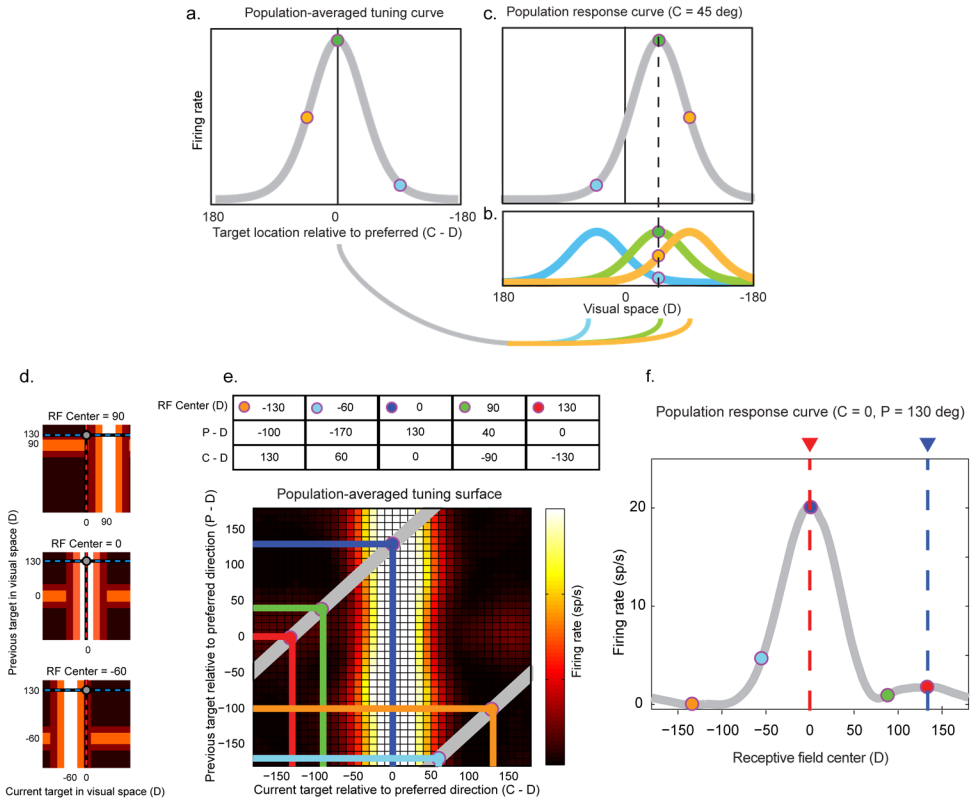
Receptive fields may converge to over-represent a location in order to increase processing of that location. More specifically, in the case of spatial working memory, receptive field convergence may make the memory trace more robust to noise, since the effect of stochastic fluctuations in activity will drop as the number of neurons involved increases (Fig. 3.8e, orange trace). If a fraction of this convergence persists into the subsequent trial, then this will introduce a bias toward the previous trial's target location (Fig. 3.8d). However, as long as the residual convergence from the previous

trial is small compared to the convergence towards the current trial's target, the total behavioral error will still be reduced as compared to the case of no convergence (Fig. 3.8e, blue trace).

Adaptation versus receptive field changes

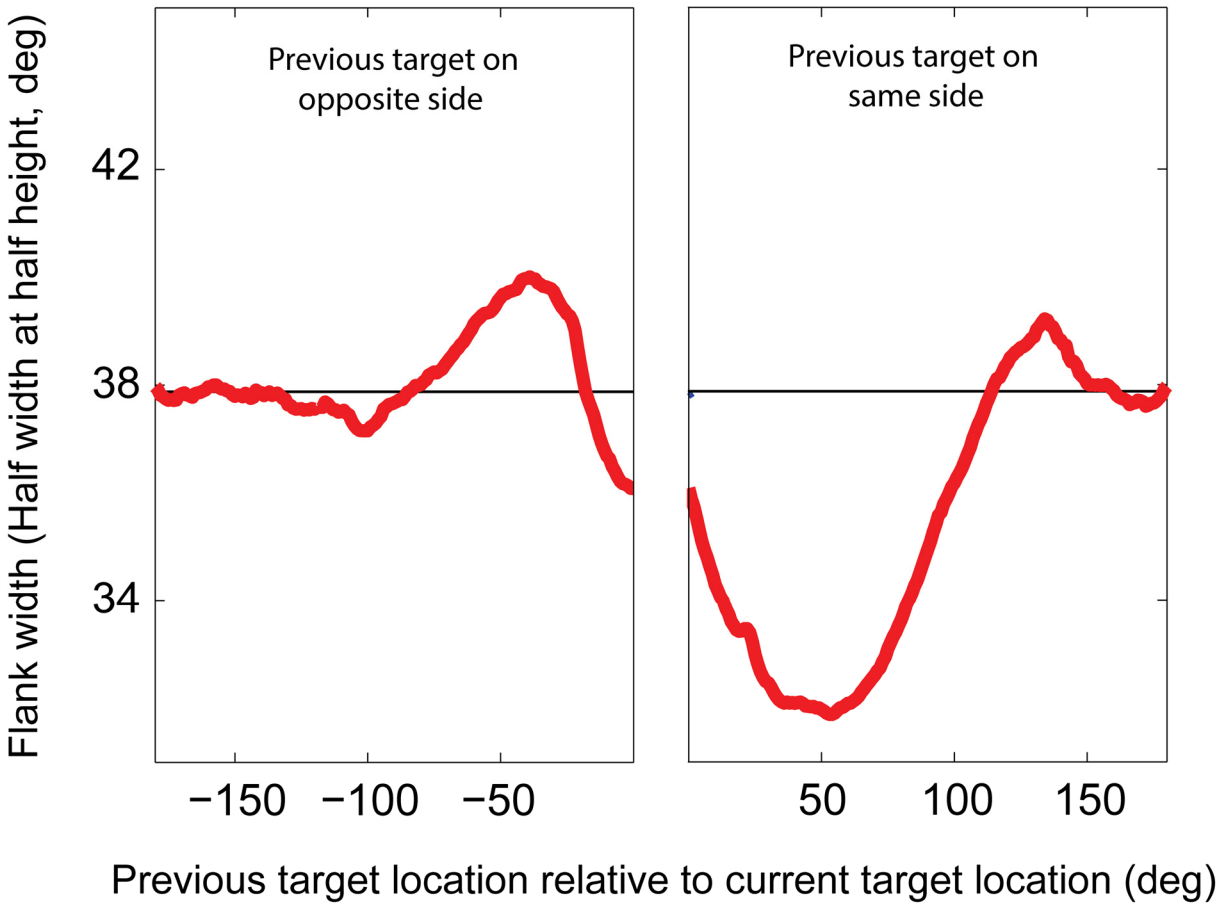
An alternative explanation for the neuronal-behavioral discrepancy that we observe is a form of firing rate adaptation. Strong activation on the one trial might lead to short term plasticity, such that neurons driven by the target on the previous trial fire at reduced levels on the next trial, compared to the level they would have fired at if they had not been active on the previous trial. This would produce a pattern of results similar to what we have shown, with a readout of neural activity that would be biased away from the previous target location. However, this explanation would account for the neural results but not the behavioral findings. In order for adaptation to account for the behavioral findings, downstream circuits closer to the motor output would have to show facilitation to counteract the suppression in FEF circuits, and this facilitation would need to overcompensate for the FEF adaptation in order to convert the repulsive FEF bias into an attractive behavioral bias. In addition, the fact that neurons show clear shifts in activity is further evidence against the adaptation model (e.g., Supplemental Figure S3.2).

Supplementary Figures



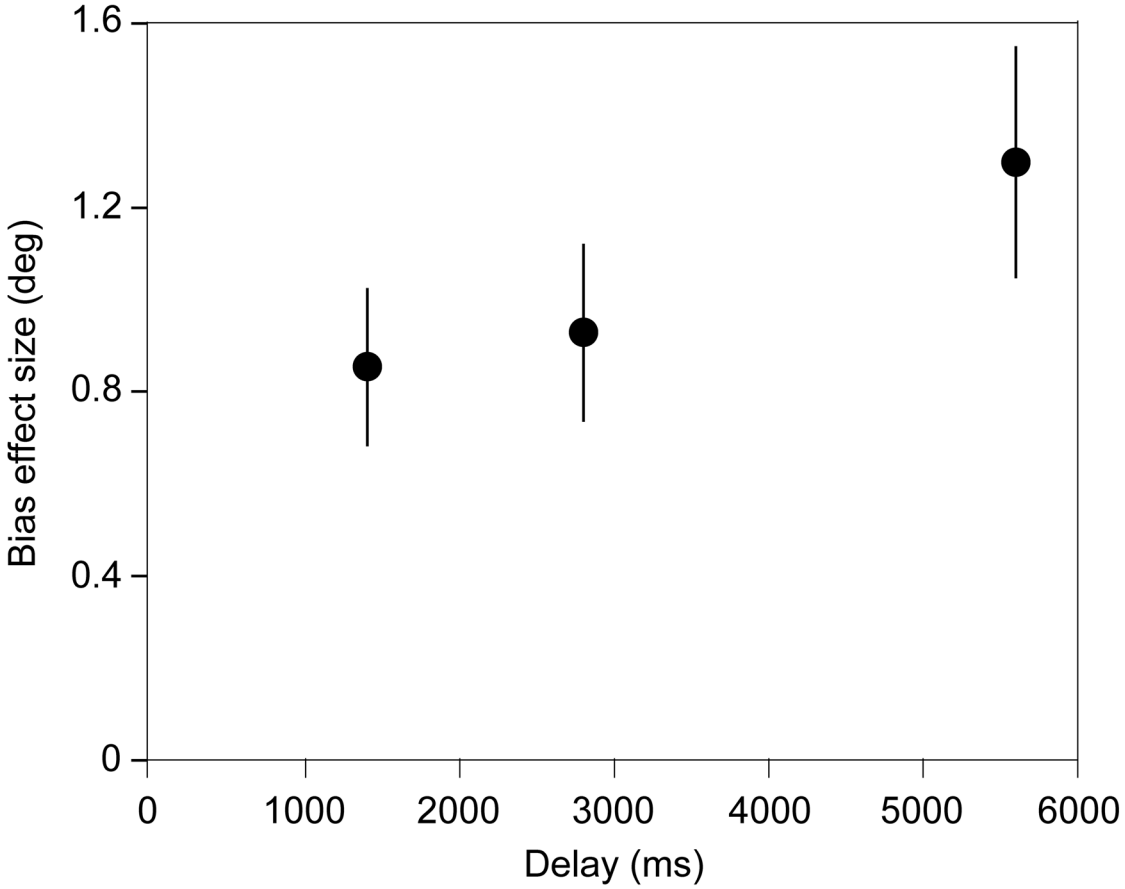
Supplementary Figure S3.1. Construction of population activity response curves from population-averaged tuning curves.

a. Schematic population-averaged tuning curve where all receptive fields have been rotated to 0 degrees and cell responses have been averaged. The y-axis shows how neurons fire to a range of targets presented at positions relative to the receptive field. The X axis is 'C – D' where 'C' is current target location and 'D' is the location in visual space of the preferred direction. **b.** Tuning curves of cells with receptive field centers at three different locations in visual space (D). Each cell's tuning curve is assumed to be identical to the population averaged tuning curve. **c.** Population response curve when a target is presented at 45 deg in visual space (dashed line). The gray curve shows how cells with receptive fields covering the entire visual space fire in response to the target presented at 45 deg. On average, the firing rate of cells with receptive field centers at location 'D' can be determined by taking the firing rate at 45 deg from the population-averaged tuning curve centered at 'D'. This is highlighted for 3 cells (orange, green, blue) with different receptive field centers in panels a, b, and c. **d.** Schematic two-dimensional population-averaged tuning surfaces (as in Fig. 3.5b) as a function of previous and current target locations for cells with 3 different preferred directions (top – 90 deg, middle – 0 deg, bottom – negative 60 deg). These surfaces are two-dimensional analogs to the curves in panel b. Brighter areas indicate higher firing rates. The dashed lines indicate previous (blue) and current (red) target locations, presented at 130 deg and 0 deg, respectively. The gray points indicate the response, on average, of cells with these receptive field centers in trials where the previous target was presented at 130 deg and the current target was presented at 0 deg. **e.** To determine the population response curve for trials with a previous target at 130 deg and a current target at 0 deg we can read the population-averaged tuning curve in Fig. 3.5b as shown in panel d. That is, to determine the firing rate of a cell with receptive field center at 'D', the surface in Fig. 3.5b is shifted so that the receptive field center is at location 'D' in visual space. Firing rate is then determined from the location x = 0, y = 130 (equivalently, point [x = 0 – D, y = 130 – D] on the surface with the receptive field center rotated to 0 deg in Fig 3.5b.) The table (left) shows the appropriate coordinates from which to read firing rates for cells with receptive fields in five different locations in space. The color plot (right), replicated from Fig. 3.5b, shows the coordinates that define the population activity curve for these trials (gray line with slope 1) and each of the five points from the table. **f.** Population response curve when the current target was presented at 0 deg and the previous target was presented at 130 deg.



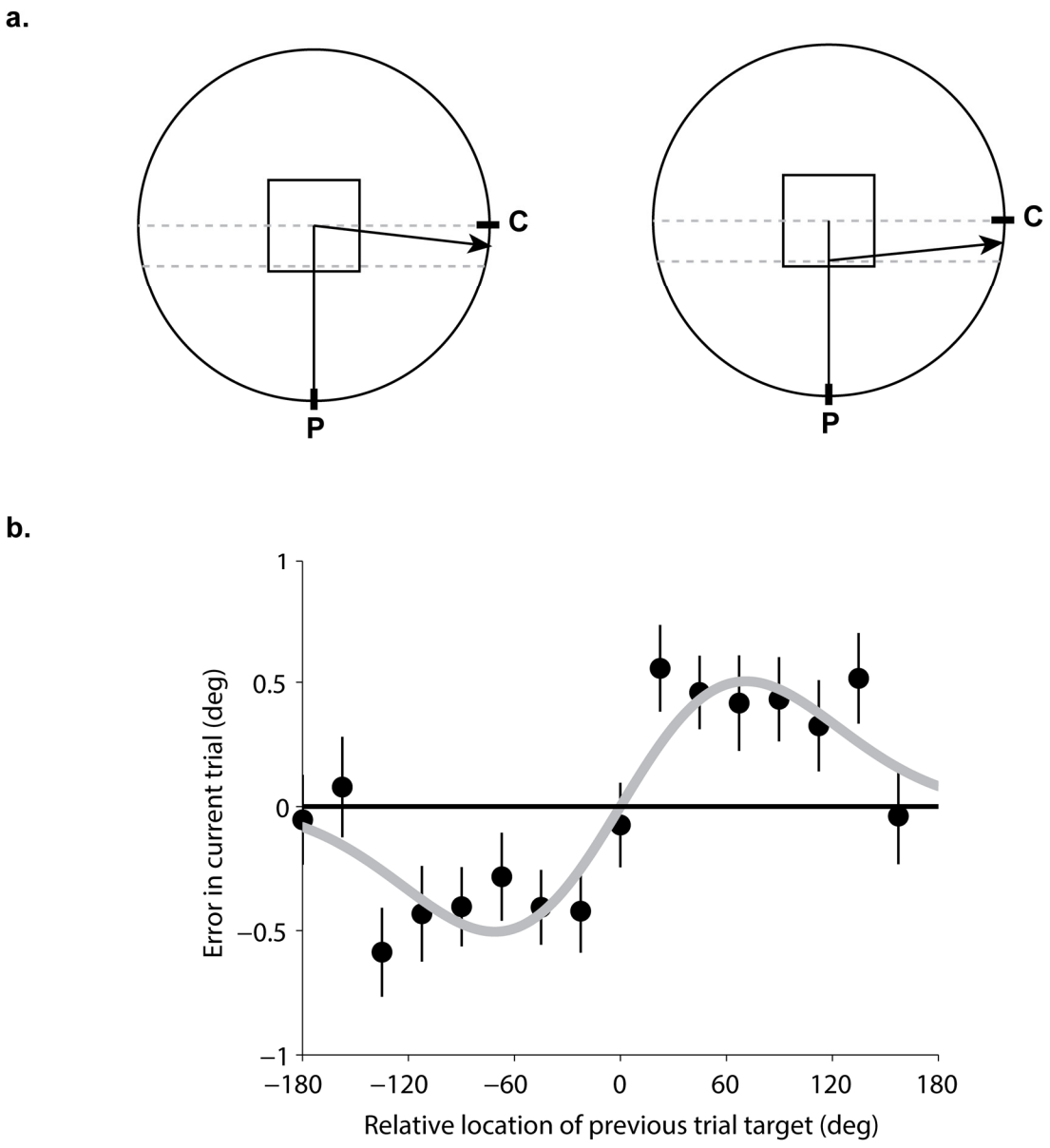
Supplemental Figure S3.2. FEF neurons show a shift in firing that is directed away from the previous target location.

We wished to determine whether the effect of a previous target, close to the location of the current target, on the population response curve (e.g., Figs.3.6a and b) is better described as a suppression (a negative ghost) or a shift. Suppression will decrease the width of the population response curve ("flank width") only on the side where the previous target was presented, and only for previous target locations close to (e.g., less than 100 deg from) the current target. A shift will have a similar effect on the side where the previous target was presented, but will in addition cause an increase in flank width on the opposite side. The baseline flank width (38 deg) was defined as the mean width when previous and current targets were more than 100 degrees apart and is indicated by the black horizontal line. We observe a clear increase in firing rate on the side opposite to that on which the target was presented. This is consistent with a shift in activity away from the side on which the previous target was presented. The effect is asymmetric, consistent with the addition of some degree of suppression in addition to the shift. Critically, however, both effects are in the opposite direction from those which would be predicted by an attractor model.



Supplementary Figure S3.3. Behavioral bias persists throughout the delay.

Each point shows the peak-to-peak height for the gabor fit for trials with delay length 1.4 seconds (peak-to-peak height = 0.85 +/-0.17 deg, $p < 0.0001$), 2.8 seconds (0.92 +/- 0.19 deg, $p < 0.0001$), and 5.6 seconds (1.3 +/-0.25 deg, $p < 0.0001$). This is consistent with previously published findings from an entirely different set of animals (Papadimitriou et al. 2015).



Supplementary Figure S3.4

a. Saccade directions calculated using two different methods for trials in which the previous target was presented at P, the current target was presented at C, and the saccadic response was made to a spatial location biased toward the previous target. Left - When determining the saccade direction in most analyses, we measured the behavioral bias based on using the fixation point as the starting point for each saccade. Right – We can instead determine saccade direction using the eye position immediately preceding saccade onset. Although the saccades shown on the left and right have the same endpoints, the saccade angles are in opposite directions. Thus, differences in initial fixation position could conceivably reconcile our neuronal and behavioral data. **b.** Error in current trial response as a function of previous target location relative to current target location. In this figure saccade directions were calculated based on the actual starting eye position (-200 to 0 ms prior to saccade onset) rather than the fixation target at the center of the circle. Behavioral responses still show an attractive bias toward the previous target and well fit by a Gabor function (peak-to-peak height = 1.07, fit $p < 0.005$). Thus systematic effects of previous target location on fixation location cannot explain the discrepancy between the neuronal and behavioral data.

References

Abbott, L.F. (1994). Decoding neuronal firing and modelling neural networks. Q. Rev. Biophys. *27*, 291–331.

Bruce, C.J., and Goldberg, M.E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. *53*, 603–635.

Bruce, C.J., Goldberg, M.E., Bushnell, M.C., and Stanton, G.B. (1985). Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. J Neurophysiol *54*, 714–734.

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

Connor, C.E., Preddie, D.C., Gallant, J.L., and Van Essen, D.C. (1997). Spatial attention effects in macaque area V4. J. Neurosci. *17*, 3201–3214.

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

Corbetta, M., Kincade, J.M., and Shulman, G.L. (2002). Neural systems for visual orienting and their relationships to spatial working memory. J. Cogn. Neurosci. *14*, 508–523.

Curtis, C.E., Sun, F.T., Miller, L.M., and D'Esposito, M. (2005). Coherence between fMRI time-series distinguishes two spatial working memory networks. Neuroimage *26*, 177–183.

Dunnett, S.B., and Martel, F.L. (1990). Proactive interference effects on short-term memory in rats: I. Basic parameters and drug effects. Behav. Neurosci. *104*, 655–665.

Edhouse, W. V., and White, K.G. (1988). Cumulative proactive interference in animal memory. Anim. Learn. Behav. *16*, 461–467.

Ferrera, V.P., Cohen, J.K., and Lee, B.B. (1999). Activity of prefrontal neurons during location and color delayed matching tasks. Neuroreport *10*, 1315–1322.

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

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

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

Georgopoulos, A. (1988). Neural integration of movement: role of motor cortex in reaching. FASEB J *2*, 2849–2857.

Hamker, F.H., Zirnsak, M., Calow, D., and Lappe, M. (2008). The Peri-Saccadic Perception of Objects and Space. PLoS Comput. Biol. *4*, e31.

Jarvik, M.E., Goldfarb, T.L., and Carley, J.L. (1969). Influence of interference on delayed matching in monkeys. J. Exp. Psychol. *81*, 1.

Jonides, J., and Nee, D.E. (2006). Brain mechanisms of proactive interference in working memory. Neuroscience *139*, 181–193.

Judge, S.J., Richmond, B.J., and Chu, F.C. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. Vision Res. *20*, 535–538.

Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–50.

Moise, S.L. (1976). Proactive effects of stimuli, delays, and response position during delayed matching from sample. Anim. Learn. Behav. *4*, 37–40.

Papadimitriou, C., Ferdoash, A., and Snyder, L.H. (2015). Ghosts in the machine: memory interference from the previous trial. J. Neurophysiol. *113*, 567–577.

Di Pellegrino, G., and Wise, S.P. (1993). Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J. Neurosci. *13*, 1227–1243.

Robinson, D.A. (1963). A Method of Measuring Eye Movemnent Using a Scieral Search Coil in a Magnetic Field. Ire Trans. Biomed. Electron. *10*, 137–145.

Ross, J., Morrone, M.C., and Burr, D.C. (1997). Compression of visual space before saccades. Nature *386*, 598–601.

Ross, J., Morrone, M.C., Goldberg, M.E., and Burr, D.C. (2001). Changes in visual perception at the time of saccades. Trends Neurosci. *24*, 113–121.

Schall, J.D. (1991). Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields. J Neurophysiol *66*, 559–579.

Sommer, M.A., and Wurtz, R.H. (2000). Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. J Neurophysiol *83*, 1979–2001.

Sommer, M.A., and Wurtz, R.H. (2001). Frontal eye field sends delay activity related to movement, memory, and vision to the superior colliculus. J Neurophysiol *85*, 1673–1685.

Takeda, K., and Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. J Neurophysiol *87*, 567–588.

Takeda, K., and Funahashi, S. (2004). Population vector analysis of primate prefrontal activity during spatial working memory. Cereb Cortex *14*, 1328–1339.

Tolias, A.S., Moore, T., Smirnakis, S.M., Tehovnik, E.J., Siapas, A.G., and Schiller, P.H. (2001). Eye Movements Modulate Visual Receptive Fields of V4 Neurons. Neuron *29*, 757–767.

Umeno, M.M., and Goldberg, M.E. (2001). Spatial processing in the monkey frontal eye field. II. Mem. Responses. J Neurophysiol *86*, 2344–2352.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. *17*, 431–439.

Zirnsak, M., Steinmetz, N.A., Noudoost, B., Xu, K.Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. Nature *507*, 504–507.

Chapter 4

Evolution of Working Memory Neural Activity Over Long Memory Periods

4.1 Abstract

Working memory, the ability to maintain and transform information, is critical for cognition. Neuronal activity in dorsolateral prefrontal cortex (dlPFC) and frontal eye fields (FEF) is elevated while monkeys hold a spatial location in memory for 1-3 s (Bruce and Goldberg, 1985; Funahashi et al., 1989). The premier model for spatial memory is the continuous attractor network (Compte et al., 2000; Wang, 2009). Memoranda are represented as localized "bumps" of activity in a topographic map of nodes. Due to a balance between local excitatory and global inhibitory recurrent feedback, the bumps are maintained indefinitely, even after the original stimulus is removed. The amplitude

and profile of the bump does not change over time, although random drift leads to inaccuracy that tends to increase with time.

Other studies have argued for more complex dynamics in memory activity. For example, single cell responses may ramp up or down over the memory period, or turn on for only a limited time within the memory period (Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010). These results have inspired alternatives to the attractor model.

We recorded 161 neurons in dlPFC and FEF as monkeys held spatial memories for up to 15 s. Single cells were carefully isolated and optimally driven, using a continuous distribution of targets. Cells with significant memory activity became tuned early (within 1 or 2 s) and most (80%) lost their tuning (turned off) before the end of the 15 s memory period. Across trials and cells, the time at which cells lost their tuning was exponentially distributed. Surprisingly, however, each cell had a relatively fixed turn-off time. Once off, cells did not turn back on and cells without memory in the first 4 seconds did not show memory later in the memory period. These results are not compatible with most previous models of memory.

4.2 Introduction

Working memory is the ability to maintain and transform information. Many cognitive tasks rely on working memory. Many studies focus on how neural circuits support

memory, and how information in these circuits may decay. These studies have implicated prefrontal cortex (PFC) as a key locus of working memory. Firing rates in dorsolateral PFC (dlPFC) and frontal eye fields (FEF) neurons are elevated while subjects hold a spatial location in memory (Bruce and Goldberg, 1985; Chafee and Goldman-Rakic, 1998; Constantinidis et al., 2001; Ferrera et al., 1999; Funahashi et al., 1989, 1993; Fuster and Alexander, 1971; Kojima and Goldman-Rakic, 1982; di Pellegrino and Wise, 1993; Sommer and Wurtz, 2001; Takeda and Funahashi, 2002, 2004; Umeno and Goldberg, 2001).

Much of the electrophysiology literature on spatial working memory is based on tasks with short memory periods, often between 1 and 3 seconds. Elevated responses observed in these tasks are sustained, often without appreciable decay, for the entirety of the memory period. These results have inspired neural attractor networks, the premiere framework for working memory circuits (Amit, 1992; Amit and Brunel, 1997; Brunel, 1996; Compte et al., 2000; Wang, 2009). Attractor networks model memory as a topographic map of nodes, with memoranda represented by a "bump" of elevated activity. Due to a balance of recurrent excitatory and inhibitory connections, the circuit can maintain the bump indefinitely, even after the original stimulus is removed. The amplitude and profile of the bump does not change over time. However, the bump can drift slowly and randomly over time, mimicking a slow decay in memory accuracy over time.

Other studies have argued for more complex dynamics in memory cells (Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010). Rather than the well-

behaved, sustained responses of attractors that are homogeneous from one cell to the next, these studies argue that the activity of single cells waxes and wanes over the course of a single memory period. At any given time the activity of some cells is elevated and others are at baseline, but population activity across all cells still provides a continuous and robust memory trace. However, many of the studies reporting heterogeneous responses do not optimize their target locations to produce a maximum response. Furthermore, many of these studies use array recordings and offline sorting. These factors raise the possibility that response heterogeneity may be due to stimuli that sub-optimally drive cell response or sub-optimal single-unit isolation.

In the current study we use a simple memory task with long memory periods of up to 15 seconds to systematically examine the time course of spatial working memory activity in prefrontal memory circuits. We use a spatial memory task with target locations continuously distributed in space, allowing us to optimally drive the cells we record. One to four electrodes are actively adjusted and monitored to ensure good isolation. We find that the time course of memory activity over long (5-15 s) memory periods is not well predicted by attractor or decaying attractor models. Most memory cells lose tuned memory activity before the end of the 15 second memory period. Surprisingly, turn-off times are not stochastic, but rather are specific to each individual cell. Once off, the cells do not turn back on, and cells without early memory responses do not turn on late in the delay.

4.3 Materials and Methods

Two macaques participated in the experiment and were trained on a center-out memory-guided saccade task. During the task subjects sat in a dark room in a primate chair. Subjects were head-fixed in a straight-ahead position facing a white screen located 30 cm away. Visual stimuli were projected onto the screen during the experiment. Stimulus presentation was controlled with custom software. Eye-position was recorded using an ISCAN infrared video eye tracking system.

Behavioral task

In the memory-guided saccade task monkeys were required to remember peripheral spatial locations. Each trial began with the presentation of a fixation point on which monkeys had to maintain fixation within 3.3 deg of visual angle. After 1.5 s, a peripheral memory target was flashed for 300 ms at a random location on a circle with a radius of 10 deg centered on the fovea. Stimulus presentation was followed by a memory period that lasted for 5.1-5.6, 7.6-8.1 or 15.6-16.1 s, during which the subject maintained fixation while remembering the location of the flashed stimulus. Subjects received up to four small rewards during the memory period. At the end of the memory period, the fixation point disappeared, cuing the subject to make a saccade to the remembered location. All behavioral measures used in our analyses are based on this initial saccade. If the initial saccade landed within 5.5 deg of the target, subjects received an immediate reward. Three hundred ms after the initial memory-guided saccade, the memory target reappeared. The subject was then required to make a corrective saccade to within 3.5 deg of the visible target in order to receive a large reward.

Behavioral performance

We computed several measures of performance. We computed the percentage of trials in which a fixation break occurred during the memory period ("fixation breaks"), ignoring those trials in which the animal broke fixation before the target appeared (Supplementary Figure S4.1). After an initial drop in fixation breaks early in the memory period, the percentage of fixation breaks remains relatively constant throughout the memory period. We also observe an increase of fixation breaks following rewards, perhaps due to the animal being distracted by the reward delivery. We also tallied the number of grossly inaccurate memory-guided saccades – those landing more than 80 deg away from the memory target – and expressed this as a fraction of all memory-guided saccades, i.e., trials in which fixation was maintained until the end of the memory period (Figure 4.1b).

To compute the mean angular error of the memory-guided saccades, we reduced the data to a single dimension by projecting all saccade endpoints onto the unit circle. Next, we computed the mean angular difference between the target and saccadic response across all trials within each experimental session, and subtracted that mean from each individual saccade angle. We then averaged the absolute angular error across all trials for each subject and for each memory period duration to obtain the mean angular (one-dimensional) error. We used a similar process to compute the mean Euclidean error, that is, the mean error in two dimensions. In each case, grossly inaccurate trials (see previous paragraph) were excluded from these computations.

Recording

In each experimental session we lowered between 1 and 4 electrodes in frontal eye fields (FEF) and dorsolateral prefrontal cortex (DLPFC) and manually isolated single units. Cells were selected for recording using a memory task with a 1.5 s memory period. Those cells that appeared to have tuned memory responses at any time during this period were selected for recording, along with a fraction of those that did not show tuning. In total, 161 cells were recorded.

Memory tuning

We classified our cells into those that showed memory tuning in the early memory period and those that did not. We first binned target locations into 22.5 deg bins and computed mean firing rates for each bin. For cells that appeared to show a tuned response in the search task, we identified the preferred direction as the bin with the highest firing rate in the search task data. We then used that direction to test for tuning in the memory task, combining the preferred direction data with the data from the two immediately flanking bins, and comparing the pooled responses from those three bins (spanning 67.5 deg) to the pooled responses from the three diametrically opposite bins. Cells that showed a significant difference ($p < 0.05$) in mean firing rate between the preferred and opposite bins early in the memory period were classified as memory cells. Cells that showed signs of tuning in the search task may either lose tuning quickly or take some time to reach tuning significance. To account for this we looked for tuning in two early memory intervals (0.5 to 1.5 s or the 2 to 4 s). Of the 161 recorded cells 70 cells showed significant tuning in the first interval and 23 additional cells became

significantly tuned in the second interval. Cells that either did not show a tuned response in the search task or were not tuned in the memory task were classified as untuned. Of the 68 cells that were untuned in the early memory intervals, 15 were tuned only during stimulus presentation and 53 showed neither visual nor memory tuning.

To generate population averaged tuning curves we binned target locations into 45 degree bins and then calculated mean firing rates for each bin. A Von Mises function was fit to the binned firing rates at multiple points in time.

Tuning changes and decay

We modeled how the random drift of a bump in an attractor circuit that does not decay would affect firing rates in single unit recordings. We assume that all of the observed behavioral error comes from random drift, so that the amount of error on any one trial indicates the amount of drift on that trial. We simulated 10,000 trials in which early memory activity randomly drifts. For each trial we shifted the early memory tuning curve (Fig. 4.2b) by an amount randomly selected from the distribution of behavioral error in 15 s trials and then sampled firing rate at a random target location. The amplitude of the tuning curve we obtain by fitting a Von Mises function to firing rate as a function of target location is the amount of tuning predicted by 15 seconds of random drift. We performed similar analyses in both one and two dimensions.

To generate a histogram of cell offset times, we first computed tuning amplitude in each cell for 0.5-1.5 s and 2-4 s memory period intervals. The larger of the two values was taken as the maximal tuning of the cell. The offset time was then defined as the time

when a cell first dropped below 25% of its maximal tuning and remained below that level for at least 1.5 s.

Sustained cells

We would like to know whether there are cells that sustain memory activity indefinitely, but we cannot test an indefinitely long memory period. Instead, we fit a curve to those cells that shut off between 2.5 s and 15 s after target offset, and then extrapolated this curve to estimate the number of cells from that distribution that we would expect to maintain memory beyond 15 s. An exponential decay fit the data better than a linear decay. Extrapolating the fit beyond 15 s and taking its area (yellow region in Fig. 4.6a) provides an estimate of the number of cells expected to sustain memory activity exceeding 15 s in our sample, assuming that there is no separate population of indefinitely-tuned cells. We tested this assumption by subtracting the estimate from the actual (observed) number of cells still holding a memory after 15 s, and asking if the result was greater than that which would be predicted by uncertainties in the fit and the data.

Untuned cells

We wished to know if cells that are initially not tuned become tuned later in a memory period. We collected 68 cells without tuning in the first 4s of the memory period. We tested whether these cells became tuned later in the memory period. To accomplish this, we split the data into 200 ms time bins and fit a cosine function to the firing rates in each bin. We computed the proportion tuned intervals (number of intervals tuned /

total number intervals) for 3 different p-criteria (0.01, 0.025, 0.05), shown in Fig. 4.4 (top). To determine the result expected by chance we randomly shuffled trials of cells with replacement, while maintaining the number of trials and trial types (5 s, 7.5 s, 15 s) in each cell. That is, for a cell with ten trials with three 5 second trials three 7.5 second trials, and four 15 second trials, we randomly sampled the equivalent number of trials for each delay length. We performed 1,000 shuffles and calculated the proportion of tuned intervals for each shuffled cell population. We used the distribution of proportion tuned intervals from the shuffles to compute the proportion of tuned intervals expected by chance, shown as the gray shaded area in Fig. 4.4 (top). This analysis was repeated for time bins of 2000 ms, shown in Fig. 4.4 (bottom).

4.4 Results

We use long memory periods of up to 15 seconds to systematically examine the time course of spatial working memory activity in prefrontal memory circuits. Two macaques participated in a memory guided saccade task (Figure 4.1a). In this task subjects are required to remember the location in space at which a stimulus was presented for a period of time and then make a saccadic response to the remembered location. We first examined behavioral error as a function of memory period length. Errors due to subjects' failure to maintain fixation are confounded by subjects' motivation and cannot be attributed to memory degradation and were not included in these analyses. As expected, behavioral performance in the task decreases over the course of the memory period. We find that the proportion of failed trials increases from 0.01 to 0.05 in

monkey C (red, $p < 0.0001$) and from 0.05 to 0.12 in monkey W (blue, $p < 0.0001$) (Fig 4.1b). Error in saccadic responses also increases for longer delay periods (Fig. 4.1c – angular error increases by 4.7 deg and 2.7 deg for monkeys C and D respectively, Fig. 4.1d – error in Euclidean space increases by 0.75 and 0.70 deg. vis. angle for monkeys C and D respectively, both measures significant to $p < 0.001$ for each monkey). While subjects performed the memory task we recorded from dorsolateral prefrontal cortex (dlPFC) and frontal eye fields (FEF).

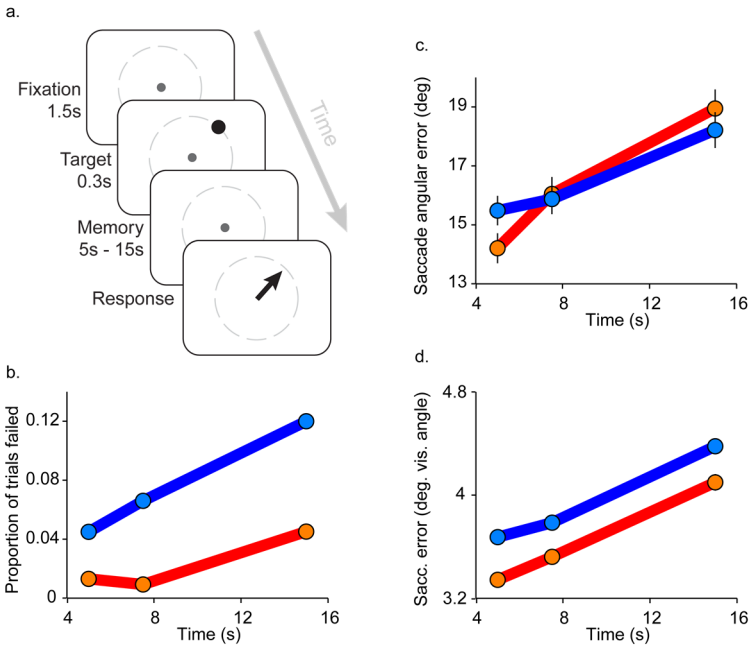


Figure 4.1. Memory task and performance.

a. The task begins with 1.5 s of fixation to a central location. A peripheral stimulus turns on for 300 ms and is then extinguished. After a memory period of 5 s to 15 s the subject makes a saccadic response to the remembered location. **b.** Proportion of trials failed as a function of memory period length for each monkey (monkey C - red, monkey W - blue). Only trials in which the subject fixated until the go-cue are included. **c.** The angular error of saccadic responses as a function of memory period length. **d.** The Euclidean error of saccadic responses as a function of memory period length. Error bars in c and d represent standard error and are smaller than the data point markers where not shown.

4.4.1 Decay of spatial working memory activity

We first asked whether memory responses were sustained throughout the 15 second memory period as predicted by attractor networks. Figure 4.2 shows that sustained memory responses become less robust over the course of the memory period. Figure 4.2a shows population firing rate over time. In the top panel traces are color-coded by their distance from the receptive field center. When the memory target was presented at the center of cells' receptive fields (red trace) population activity is elevated compared to

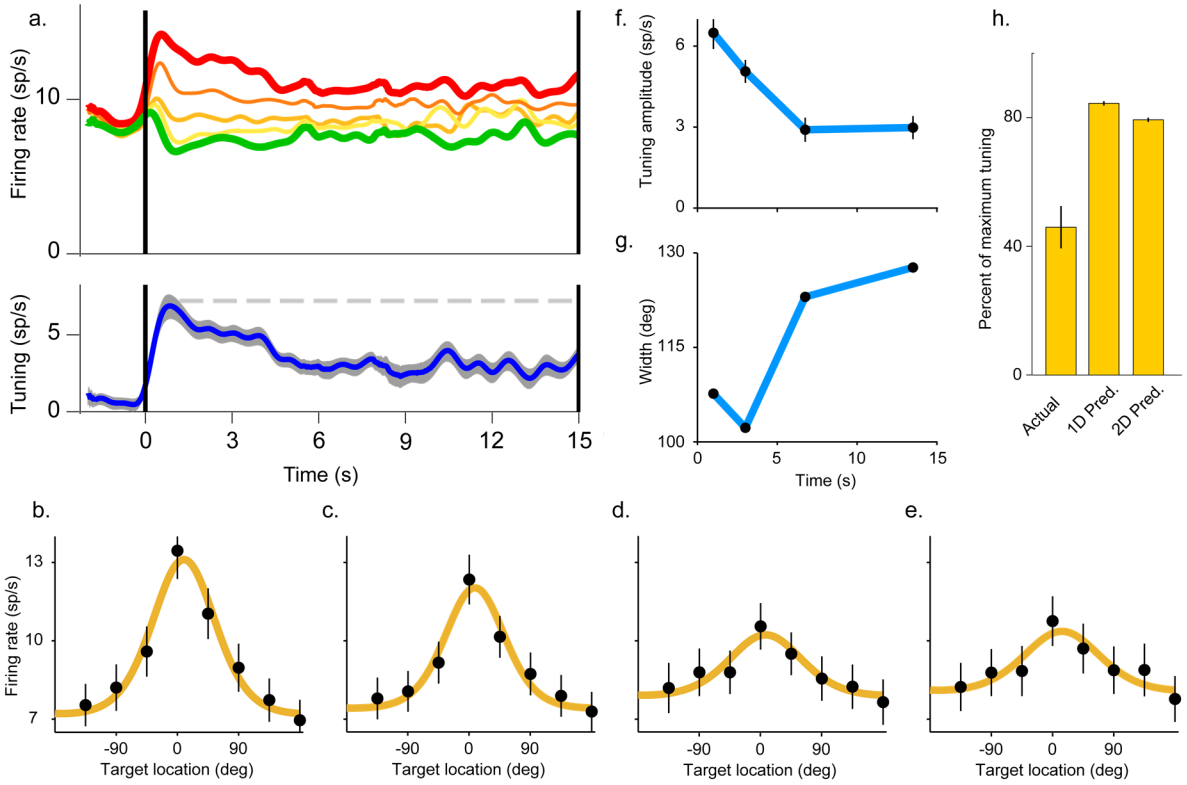


Figure 4.2. Memory tuning throughout the memory period.

a. Top - Population neural activity when the memory target was in cells' receptive fields (red trace), outside of the receptive field (green trace), or at various points in the receptive field flanks (orange and yellow traces). Bottom - Memory tuning (difference between the red and green traces) throughout the memory period. **b - e.** Neural activity as a function of memory target location at different parts in the memory period (b - 500 ms to 1500 ms, c - 2000 ms to 4000 ms, d - 6000 ms to 7500 ms, e - 12000 ms to 15000 ms). **f.** Tuning amplitude of the data in b, c, d, and e, computed as the difference between a target at the center of the receptive field (0 deg) and on the opposite side of the receptive field (180 deg). **g.** Full width at half height of the tuning curves shown in b, c, d, and e. **h.** Percent of tuning amplitude remaining after a 15 s memory period compared to early memory tuning shown in b. 'Observed' is the value from actual data (shown in e). '1D' and '2D' bars show predictions of random drift in one and two dimensions.

when it was presented outside of cells' receptive fields (green trace). The blue trace in the bottom panel shows the difference between the red and green traces, that is, the difference in response for memory targets at the center and outside of cells' receptive fields. Although memory tuning is sustained for the duration of the trial we see a clear decrease in tuning from early to late in the memory period.

A number of studies have identified similar elevated neural memory responses in prefrontal circuits (Bruce and Goldberg, 1985; Chafee and Goldman-Rakic, 1998; Constantinidis et al., 2001; Ferrera et al., 1999; Funahashi et al., 1989, 1993; Fuster and Alexander, 1971; Kojima and Goldman-Rakic, 1982; di Pellegrino and Wise, 1993; Sommer and Wurtz, 2001; Takeda and Funahashi, 2002, 2004; Umeno and Goldberg, 2001) that are sustained during the memory period of a memory task. Much of the working memory literature examining neural correlates of memory decay in these circuits have focused on memory periods of 1 to 3 seconds. These studies find cells with elevated memory responses many of which are sustained for the duration of the memory period (e.g. Chafee and Goldman-Rakic, 1998; Funahashi et al., 1989; Wimmer et al., 2014). These findings have given rise to the neural attractor framework, a class of neural networks whose dynamics include stable attractor states (Amit, 1992; Amit and Brunel, 1997; Brunel, 1996; Compte et al., 2000; Wang, 2009). Input from a briefly presented stimulus can drive the network from the baseline state to the attractor state, in which elevated firing rates initially driven by stimulus presentation are then preserved indefinitely by recurrent connections. Dynamics in attractor networks that hold memory for continuous memoranda (e.g. remembering locations in space) contain line attractor states. A line attractor network can indefinitely maintain elevated firing

rates, but during the memory period the elevated activity can randomly drift from cells responsive to the memory stimulus to cells that are responsive to nearby stimuli (e.g. Compte et al., 2000). The random drift of elevated activity during the memory period leads to behavioral responses that are increasingly inaccurate over time. Thus these networks predict decay in memory encoding without a decrease of the population firing rates.

Contrary to this we find that memory tuning in our population of cells decays over time. To quantify memory decay over time we computed population tuning curves for different intervals in the memory period. We first aligned cells' receptive field centers to 0 degrees. We then fit a Von Mises function to mean neural activity across cells as a function of target location (Figures 4.2b – 500 ms to 1500 ms, 2c – 2000 ms to 4000 ms, 2d – 6000 ms to 7500 ms, and 2e – 12000 ms to 15000 ms). Tuning curves become progressively shallower (Fig. 4.2f) and broader (Fig. 4.2g) from early to late memory intervals. The amplitude decays by 3.5 +/- 0.7 sp/s ($p < 0.0001$) while the width (Fig. 4.2g) increases by 20 deg although this increase is not significant ($p = 0.2$).

To determine how information decays in the memory network we computed Fisher information in four time intervals in the memory period (Supplementary Figure S4.2). This is a measure that describes the amount of information the firing rate carries about the encoded memory targets. We find that target information in late memory is reduced by 86% of information present in the early memory interval. This is a much larger change than the increase in behavioral error, which only increases by 25% (mean error across two monkeys). This seems to suggest that the neural activity may initially

contain excess information compared to what is needed to drive the behavior. However, there may be a complex relationship between Fisher information and saccade endpoint error. Alternatively, information degradation in neural activity may be accompanied by transfer of information to other forms of encoding such as synaptic storage.

4.4.2 Apparent tuning decay due to random drift

We looked at whether the memory tuning decay we observed is consistent with that expected from a continuous attractor. Although the peak firing rate in an attractor does not decay, from the perspective of single unit recordings random drift across many trials can appear as a decay in memory tuning (see Supplemental Figure S4.3). The amount of random drift in a trial is proportional to the memory period. As activity drifts randomly away from its initial location, cell responses to stimuli presented at different parts of space become less distinguishable from one another. That is, the firing rate difference between a stimulus in the receptive field and one outside of the receptive field become smaller.

To determine whether the decay in tuning is consistent with an attractor circuit, we simulated tuning decay due to random drift along the circle of memory targets. That is, we used the amount of response error (Fig. 4.2c) at the end of the 15 second memory period to predict the expected tuning decay based on the random drift hypothesis. The amount of error in saccadic responses is proportional to the drift in activity because the population activity bump center represents the encoded location. We simulated 10,000 trials with a random memory target location and a drift amount randomly selected from the distribution of errors in saccadic response angles in 15 second trials. For each trial,

we shifted the early memory tuning curve in Fig. 4.2b by the drift amount. The firing rate of the shifted tuning curve at the memory target location is the firing rate predicted by the random drift hypothesis for that trial. We binned target locations in 22.5 degree intervals and averaged firing rates across all trials in each bin, generating the tuning curve predicted by the random drift hypothesis for a 15 second memory period. We then compared this prediction with the actual tuning curve shown in Fig. 4.2e.

We find that apparent decay due to random drift greatly underestimates the amount of tuning decay observed in actual neural responses. Memory tuning amplitude 15 seconds after memory target presentation drops to 45 +/- 7 percent of early memory tuning, from 6.49 +/- 0.60 to 2.98 +/- 0.43 sp/s (Fig. 4.2f, 4.2h). In comparison, the random drift along the circle of targets predicts a drop in tuning amplitude to 84 +/-0.7 percent. We also simulated random drift in two dimensions, allowing activity to drift on the two dimensional plane on which targets were presented. Two dimensional drift predicts that tuning drops to 79 +/- 0.7 percent of early memory tuning. Both the one dimensional and two dimensional drift simulations significantly underestimate tuning decay observed in the recorded cells ($p < 0.0001$ in both cases).

4.4.3 Distribution of memory tuning offset times

We next asked whether the tuning decay we observed was consistent with an imperfect, decaying attractor. A decaying attractor network uses the same architecture and dynamics as a standard attractor model but excitation and inhibition is not balanced so that neural activity is indefinitely sustained. Instead, the 'attractor' state is not a truly stable state but a state that slowly decays back to baseline over time (e.g. Wimmer et al.,

2014). Cells in this type of network lose tuning with a similar time constant to each other and to that of the population. To determine whether decay of neural activity is consistent with a decaying attractor we looked at memory activity offset times.

We first looked at the distribution of offset times across cells. We calculated tuning over time for each cell and defined the offset time as the time when a cell dropped to 25% of its maximum tuning and remained below that level for at least 1.5 seconds. Because the population activity only drops to 45% of tuning by 15 seconds, few if any cells are expected to drop to 25% of maximum tuning if the recorded cells are part of a decaying attractor. Contrary to this, we found that cells had offset times that were distributed broadly throughout the entire memory period (Figure 4.3a – histogram, Figure 4.3b – survival curve). Only 19 of the 93 cells had offset times greater than 15 seconds (Fig. 4.3a – red bar).

We next asked whether offset times within a cell were the same in each trial. If the memory circuit is noisy such that cells in the circuit can randomly shut off at a random time in each trial, the distribution of cell offset times may appear broad when a limited number of trials is collected per cell. We therefore wished to determine whether the wide spread of offset times observed in cells was due to random noise in trial-by-trial offset times. Figure 4.3c shows responses of five example cells with early offset times, late offset times, and sustained (no offset time) for trials with memory targets presented in their receptive fields. We found that cells followed a repeatable time course, and offset times were similar in each trial. We quantified this result by computing the mean offset time for the trials that decayed in each cell, and then computing absolute distance

in time of each offset from that mean (Fig. 4.3d). The mean distance in time of each trial's offset to the mean offset is significantly smaller than that expected by chance (Left set of bars, 1.36 s, $p < 0.002$). This result was also true when looking at only 15 second trials, which allow for a larger spread in offset times (Right set of bars, 1.63 s, $p < 0.006$).

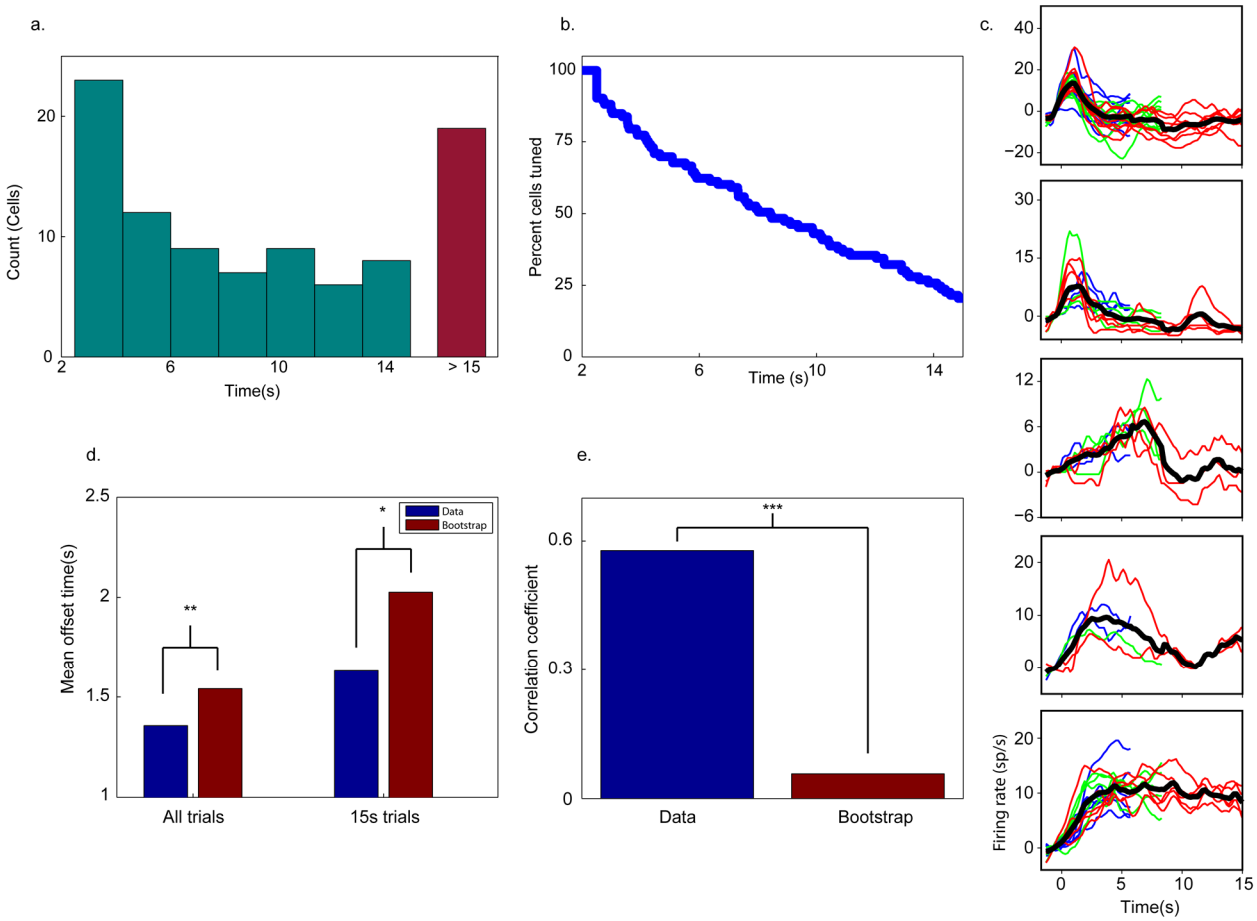


Figure 4.3. Tuning properties of individual cells.

a. Histogram of offset times (when cells reach 25% of their maximum tuning). **b.** Survival curve of offset times showing the percentage of cells that are tuned throughout the memory period. Of the 93 cells, 19 (20%) do not turn off 15 s into the memory period. **c.** Firing rates from five example cells in 5 s (blue), 7.5 s (green), and 15 s (red) trials when the memory target was at the center of each cell's receptive field. The black trace is the mean of trials shown. **d.** Mean absolute difference of each trial's offset from the mean offset across all trials. The mean absolute difference is significantly smaller than that expected by chance for all trials (1.36 s, $p < 0.002$). This result was also true when looking at only 15 second trials, which allow for a larger spread in offset times (1.63 s, $p < 0.006$). **e.** Correlation of mean offset time of trials that decayed with the proportion of trials that persisted for the entire memory period ($r = 0.57$, $p < 0.0001$).

Because delay times ranged from 5 to 15 seconds, many decaying cells included trials that were sustained. For example, 5 second trials should not decay prior to the end of their memory period if a cell has a mean offset at 7 seconds, while most 15 second trials should decay. If cells decay repeatably in each trial, we expect that the later the mean offset time of the trials that decay, the larger the proportion of trials that should not decay prior to the end of their memory period. That is, the mean offset time of decaying trials should be strongly correlated with the proportion of trials that don't decay for each cell. Alternatively, if decay is random in each trial, the mean offset time of decaying trials and the proportion of sustained trials should be uncorrelated. We find that cells whose decaying trials have a late mean offset also have a much larger proportion of trials that are sustained until the end of the memory period (Fig. 4.3e, $r = 0.57$, $p < 0.0001$). This result indicates that memory activity in each cell is sustained for a common length of time in every trial. These findings, taken together, suggest that cell responses do not reflect an attractor circuit with indefinitely sustained activity with memory degradation due to random drift or a decaying attractor in which individual cells turn off at the same rate as the population average. Instead, memory circuits are composed of cells with repeatable offset times that broadly span the entire memory period.

4.4.4 Heterogeneous responses

The variety of offset times across memory cells suggest more complex memory dynamics than predicted by attractor networks. These results are consistent with a number of studies that report great heterogeneity in the memory activity traces of each cell but repeatable trial-by-trial responses within each cell (Baeg et al., 2003; Brody et al., 2003;

Harvey et al., 2012; Jun et al., 2010). These studies show single cell responses that may ramp up or down over the memory period, or turn on for only a limited time within the memory period. Although no single cell allows for memory decoding at all times during the memory period, memory can be retrieved by looking at the combined population of cells.

We therefore investigated whether cells we record show similar complexity in onset and offset patterns. We first asked whether cells whose tuning decayed showed significant tuning again at a later time in the memory period. We found that only one of the 49 cells that decayed before 9 seconds showed significant memory tuning later in the memory period (p-criterion for significance = 0.05). We then asked whether cells typically become active immediately after target presentation or can become active at any point in the memory period. The 93 memory cells we recorded were selected to show memory early in the memory period. We therefore recorded from 68 cells located in the same FEF or DLPFC tracks as the memory cells but that did not show early memory responses in the first 4 seconds of the memory period. Of the 68 cells, 19 cells showed significantly tuned visual responses ($p < 0.05$) during the 300 ms of stimulus presentation. We asked if this population of 68 cells showed significant tuning at any point in memory period. We divided the memory period into 500 ms bins. We then fit a cosine function to firing rate as a function of memory target location to each bin. A significant fit indicates that a cell showed significant tuning during that part of the memory period. We computed the number of significantly tuned bins using 3 different p-criteria (Figure 4.4 - top, red points) and evaluated whether or not that number was higher than what is expected by chance (grey shaded areas) using a shuffle procedure

(see *Methods*). This analysis was also repeated for 2000 ms bins (Fig. 4.4 - bottom). We found that the number of significantly tuned bins was not greater than that expected by chance for any of the time bin lengths we tested. These results indicate that cells with no early memory response do not turn on later in the memory period.

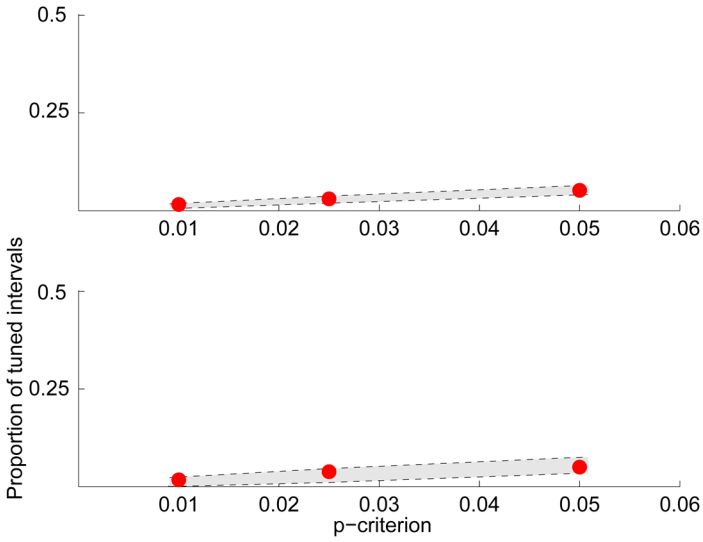


Figure 4.4. Cells untuned in the early memory period show no tuning in later memory.
Proportion of time intervals that show significant tuning at a p-criterion of 0.01, 0.025, and 0.05. Gray shaded region indicates chance at a p > 0.05 based on repeating the analysis on 1000 cells populations in which trials have been randomly shuffled. The analysis was done with 500 ms (top) and 2000 ms (bottom) time intervals.

Although some previous studies found that memory cells can alternate between on and off at any time during the memory period, many of these studies only used two memoranda – one inside the receptive field and one outside – to determine memory responses. It is therefore likely that memory stimuli were not optimized to maximally drive memory cells in these studies. Figure 4.5 shows tuning amplitude for two cells in our dataset evaluated at the flank rather than at the center of their respective receptive fields (red traces). The cell in Fig. 4.5a shows a delayed onset, and turns on and off multiple times throughout the memory period. The cell Fig. 4.5b shows an early onset, turns off at 2.5 seconds, and comes back on for the last half of the memory period. These tuning patterns generated when cells are driven sub-optimally are consistent with heterogeneous responses described in previous literature. However, when the example cells are optimally driven by stimuli at the center of their receptive fields (blue traces) both cells show continually sustained

responses for the entire duration of the memory period. In addition, a number of studies use array recordings and offline sorting algorithms to isolate memory cells. Sub-optimal isolation and classification of single units may also produce firing rate tuning patterns that appear to be heterogeneous even though cells may be continually active. It is therefore possible that heterogeneous responses of the type described in these studies could be due to sub-optimal driving of memory cells or sub-optimal isolation and classification. We find no evidence of repeated onsets and offsets in the cell population we record when driving cells with optimal stimuli and using manual online isolation methods. Instead, our findings show that memory cells turn on early after memory stimulus presentation, sustain activity for distinct and fixed lengths of time, then turn of and stay off for the remainder of the memory period.

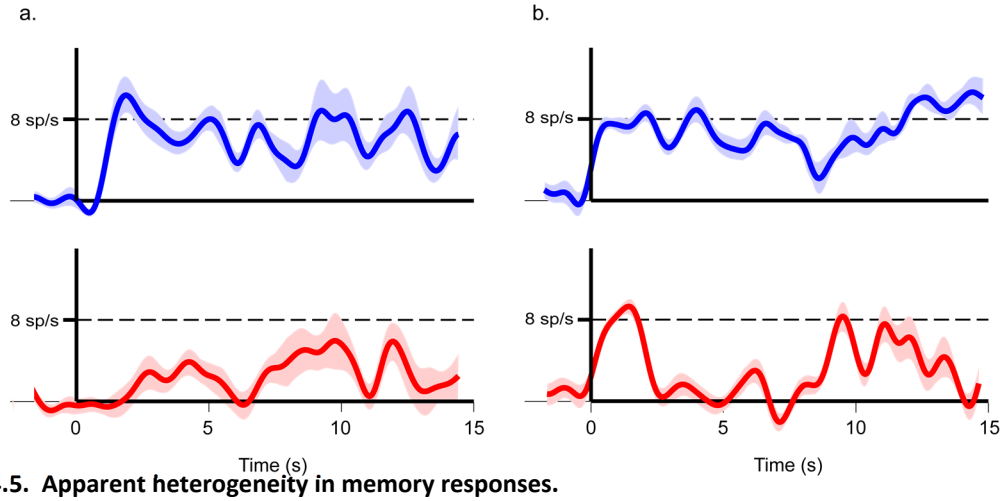


Figure 4.5. Apparent heterogeneity in memory responses.

Two cells with well-behaved sustained memory tuned responses. Blue traces show the difference in firing when memory targets are either at the center of cells' receptive fields or at the opposite location on the circle. Red traces show difference in firing when the memory targets are either at a flank 45 deg away from the receptive field center, or at the opposite location on the circle.

4.4.5 Population of sustained cells

We asked whether the 19 cells that sustained tuning beyond 15 seconds were part of the same population of cells as those that decayed, or a different population that is perhaps part of an attractor circuit. We first determined the probability distribution that describes cell offset times. Because cells were not required to show tuning until 2s in the memory period, sampling of offset times prior to 2s when cells may still be building up tuning is different from sampling later in the memory period. We therefore excluded 9 cells with offset times less than 2.5 s or less from this analysis. We fit a number of common probability distributions to the distribution of offset times for the 74 cells that decayed after 2.5 s but before 15 s (see *Methods*). We found that distribution of offset times was best described by an exponentially decaying distribution with a time constant of 5.4 s, shown in Figure 4.6a. Offset times of up to 15 seconds account for 90 percent of the area under the distribution curve, shaded in green, and marked in the cumulative distribution (Fig. 4.6b) with a dashed line. The remaining 10 percent of the area, shaded in yellow, describes the expected proportion of offset times in our sample that are part of this distribution but greater than 15 seconds. Based on this distribution we expect 8.5 cells (95% CI 5 to 14 cells) to be sustained, lower than the number of sustained cells we observe. This result suggests that at least some of the sustained cells may be part of a different population. However, the proportion of memory cells that may be part of a circuit with longer sustained memory responses is small. Repeating the analysis with the excluded cells did not change this result.

If some of the sustained cells are part of a separate population from the decaying cells they may be anatomically clustered. To determine if this is the case we plotted the anatomical location of recorded cells for monkey C (Fig. 4.6c) and monkey W (Fig. 4.6d). Points plotted in grey show recording tracks with cells that decayed prior to the end of the memory period. Points plotted indicate tracks that contain at least one cell from the sustained population. We find that sustained cells are not clustered but are distributed across both principal and arcuate sulci.

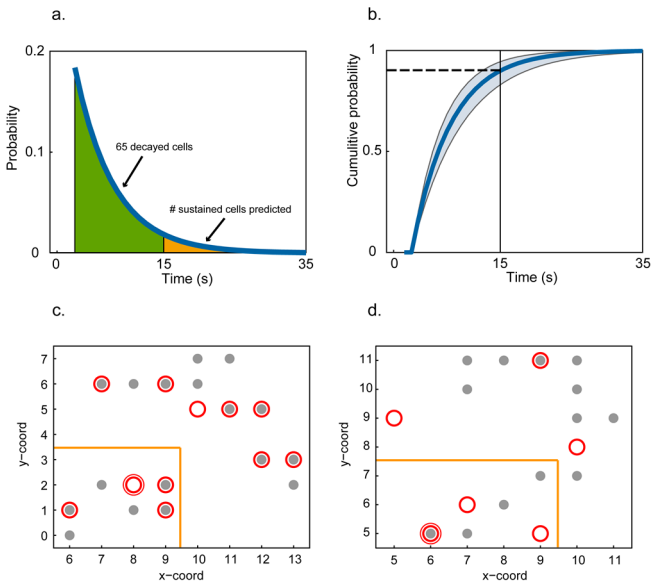


Figure 4.6. Properties of sustained cells.

a. Fitted probability density exponential distribution to offset times. The green area represents the cells that decayed. The yellow area is the proportion of cells predicted to sustain activity for 15 s or more. **b.** Cumulative distribution of the distribution in a. At 15 s 90% of the cells have decayed, indicating a prediction that 10% sampled cells should decay later than 15 s. **c,d.** Anatomical distribution of sustained cells for monkey C (c) and monkey W (d). Each grey circle represents a location in which one or more decaying memory cells were recorded. Red open circles indicate the locations of sustained cells, one for each cell recorded at those coordinates. Orange lines separate the principle sulculs (bottom left) from the arcuate (top right).

4.5 Discussion

In this study we systematically examine the evolution of frontal neural activity during a spatial working memory task with longer memory periods (up to 15s) than used typically. Because memoranda and behavioral responses in our task are continuous we can optimally drive the cells we record. We actively monitor recorded cells to ensure good isolation. Our recordings indicate that memory cells typically turn on early in the

memory period and stay active for a fixed length of time specific to each cell. Once cells turn off they do not come back on. Cells in the same recording tracks that do not show early memory activity do not become active later in the memory period. These findings indicate that working memory dynamics are more complex than predicted by attractor networks but contain more structure than suggested by competing heterogeneous models.

Previous studies of working memory with short (e.g. 1 to 3 s) memory periods found that many memory cells show elevated neural activity that is sustained for the entire memory period (e.g. Chafee and Goldman-Rakic, 1998; Funahashi et al., 1989; Wimmer et al., 2014). These findings inspired attractor networks with stable, non-decaying memory states (Amit, 1992; Amit and Brunel, 1997; Brunel, 1996; Compte et al., 2000; Wang, 2009). Continuous attractors, a class of attractor networks used to model memory circuits that maintain continuous memoranda such as spatial locations, represent memory as a 'bump' of elevated neural with its peak centered at neurons representing the memorized part of space. The amplitude of the bump can be sustained indefinitely without decay. To account for behavioral performance degradation that increases as a function of memory period length the activity bump is allowed to drift randomly over time. Even though the bump does not decay, single unit mean tuning across many trials with random drift will still decrease (Supplementary Fig. S4.1). We find that tuning of neural activity during spatial working memory maintenance decreases by 45% over the course of the 15 s memory period. This decrease is much higher than that predicted from random drift alone (Fig. 4.2h).

If inhibition and excitation in an attractor circuit is not perfectly balanced then the stable attractor state may decay after some time. We asked whether the decay we see is consistent with a decaying attractor network. To answer this question we generated the distribution of cell offset times, defined as the time when each cell dropped to 25% of its maximal tuning. Cells in an imperfectly balanced attractor network will lose memory tuning at the same rate to each other and to that of the population. In addition, because the population only drops to 45% of its initial tuning by the end of the memory period, few if any of the cells we record are expected to turn off. Contrary to these predictions we find that cell offset times are broadly distributed throughout the course of the entire memory period. Only 19 of the 93 cells remain tuned for the entire 15 seconds.

Cells in a noisy network can appear to have broadly distributed offset times due to a limited number of recorded trials. To determine if this is the case we examined the trial-by-trial responses of each cell. We found that cell responses, whether sustained or decaying, are consistent and repeatable from one trial to the next. The offset time differences between cells is not due to random noise but instead offset times are consistent from one trial to the next. This could be due to internal mechanisms within each cell or to large-scale network level effects such as heterogeneity in the recurrent connectivity between cells.

The heterogeneity in cell offset times is consistent with a number of studies that report memory activity traces with highly complex dynamics. These studies show single cells that may ramp up or down and turn on or off at any time in the memory period (Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010). Single cell responses

are repeatable from one trial to the next. However, these studies often use stimuli not optimized to maximally drive memory cells. Tasks in some studies are more complex than our task and involve a decision making component. Finally, cell isolations may be suboptimal due to use of electrode array recording and offline sorting techniques. We show that recording from the flanks of memory cells can produce responses that appear to have a complex time-course with multiple onsets and offsets, but actually come from cells that sustain memory activity for the duration of the memory period. We also find that after our memory cells turn off they do not show significant memory activity later in the memory period. Finally we record from 68 cells in the same recording tracks that do not show memory activity in the first 4 seconds and find that these cells do not acquire memory tuning later. The time-courses of memory activity in cells we record show more structure than heterogeneous models of working memory circuits.

4.6 Conclusions

In summary, most memory cells turn on early in the memory period, stay active for a variable but cell-specific amount of time, then shut off and stay off for the remainder of the memory period. These findings challenge the attractor network model of working memory, but also show that memory responses are much more structured than suggested by competing heterogeneous models.

While our results are quite novel, they nonetheless replicate the main findings of earlier work that tested memory for only a few seconds. Like the earlier studies (e.g. Chafee

and Goldman-Rakic, 1998; Funahashi et al., 1989; Wimmer et al., 2014), we find that many cells are continuously active during the first two seconds of memory, with activity in some cells ramping up and others ramping down. Transient activations, multiple on/off cycles, and other complex dynamics similar to those reported previously period (Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010) also appear in our data, but only when non-optimal stimuli are used. Our data opens up the possibility that these complex dynamics do not reflect the operation of memory circuits. Instead, they may reflect responses from the flanks of the tuning curves.

4.7 Future Directions

We show clear and novel findings regarding the time-course of working memory activity in frontal memory circuits. We are working to add to these findings by using our dataset to address a number of additional and related questions.

4.7.1 Relation of neural and behavioral responses

We have shown that both behavioral performance and neural encoding of target location degrades with time. We asked whether behavioral degradation is driven by attractor dynamics. That is, we used the observed behavioral error to simulate the amount of decay in neural activity we expect to see in an attractor circuit. We find that the decay in neural activity is too large to be accounted for by continuous attractor random drift. We plan to expand this analysis by asking how the amount of memory degradation observed

in neural and behavioral responses are related. This will allow us to determine how neural activity in memory circuits is processed by readout mechanisms to drive behavior.

4.7.2 Simultaneously recorded pairs

Our dataset contains a number of simultaneously recorded cell pairs. We will analyze trial-by-trial correlations in the neural activity time course in these pairs to determine network properties. To determine whether the controlling factor in causing cells to turn off is a network property or a cell-autonomous property, we will test whether trial-by-trial offset times are correlated across cell pairs**.** The strength of correlation will also reveal to what extent recurrent connections influence neural activity dynamics between cells. Stronger correlation between cells with similar offset times may indicate that these cells have stronger connectivity to one another and implicate models with differential connectivity strength across multiple memory sub-populations.

4.7.3 Sustained cell properties

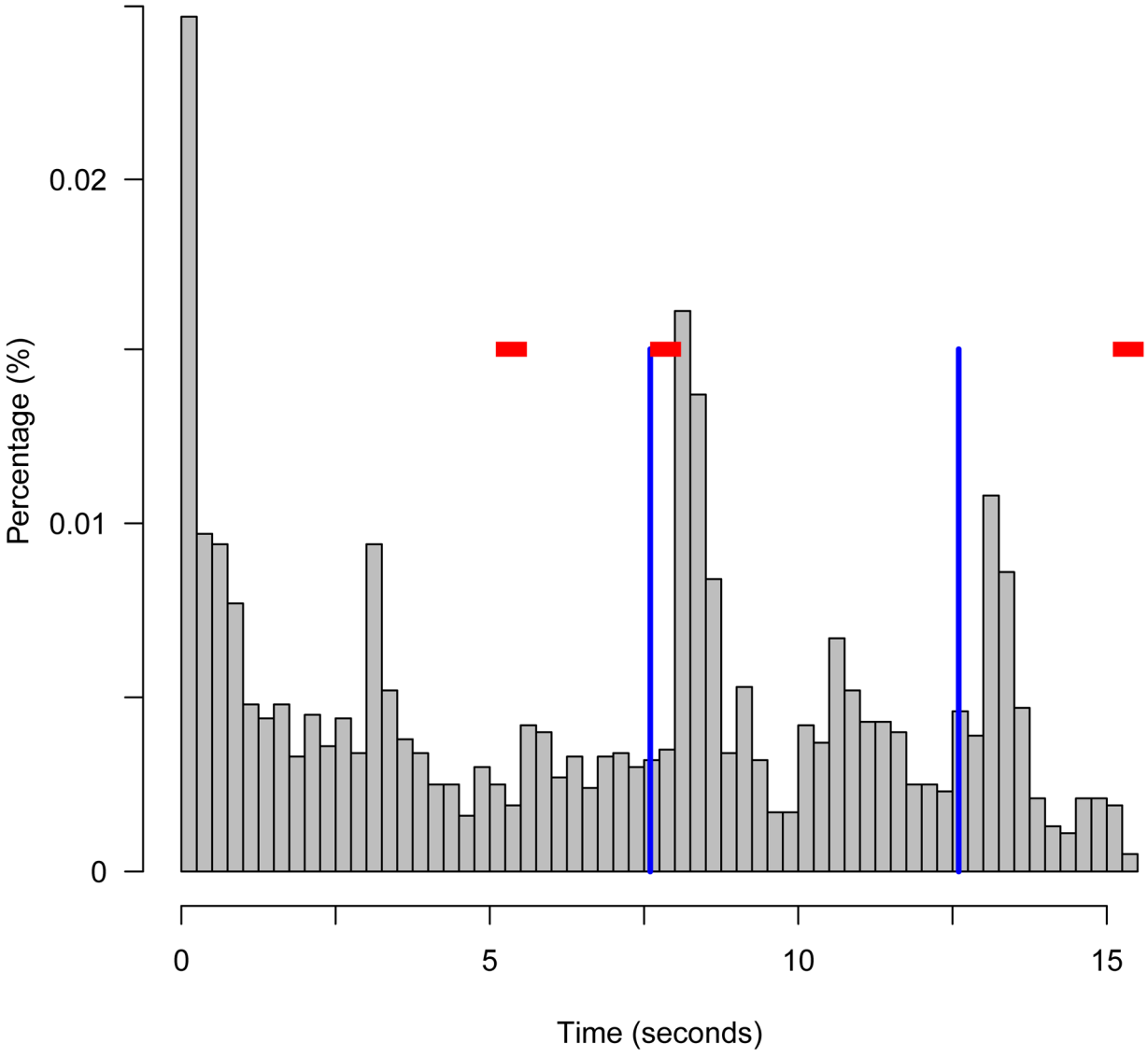
A number of cells showed sustained memory activity for the entire memory period. We asked whether these cells were a separate population of attractor cells, assuming a continuous distribution of offset times, and found that about 8 of these cells could come from a different population with a substantially longer time constant of decay. We have begun to look at the sustained cells in more detail and preliminary findings suggest that they have other qualitative differences from the decaying cells. Of note is the fact that, compared to decaying cells, a large proportion of sustained cells (~50%) have

contralateral visual tuning responses but ipsilateral memory tuning. This pattern of response in which visual and memory periods have opposite tuning and memory responses are ipsilateral suggest that these cells may be inhibitory interneurons (Gabbott and Bacon, 1996; Wang et al., 2004; Zhou et al., 2012). In future analyses we can analyze their spike wave forms and firing properties in more detail to determine whether our dataset contains a population of inhibitory neurons with longer decay times than pyramidal cells.

4.7.4 Computational models

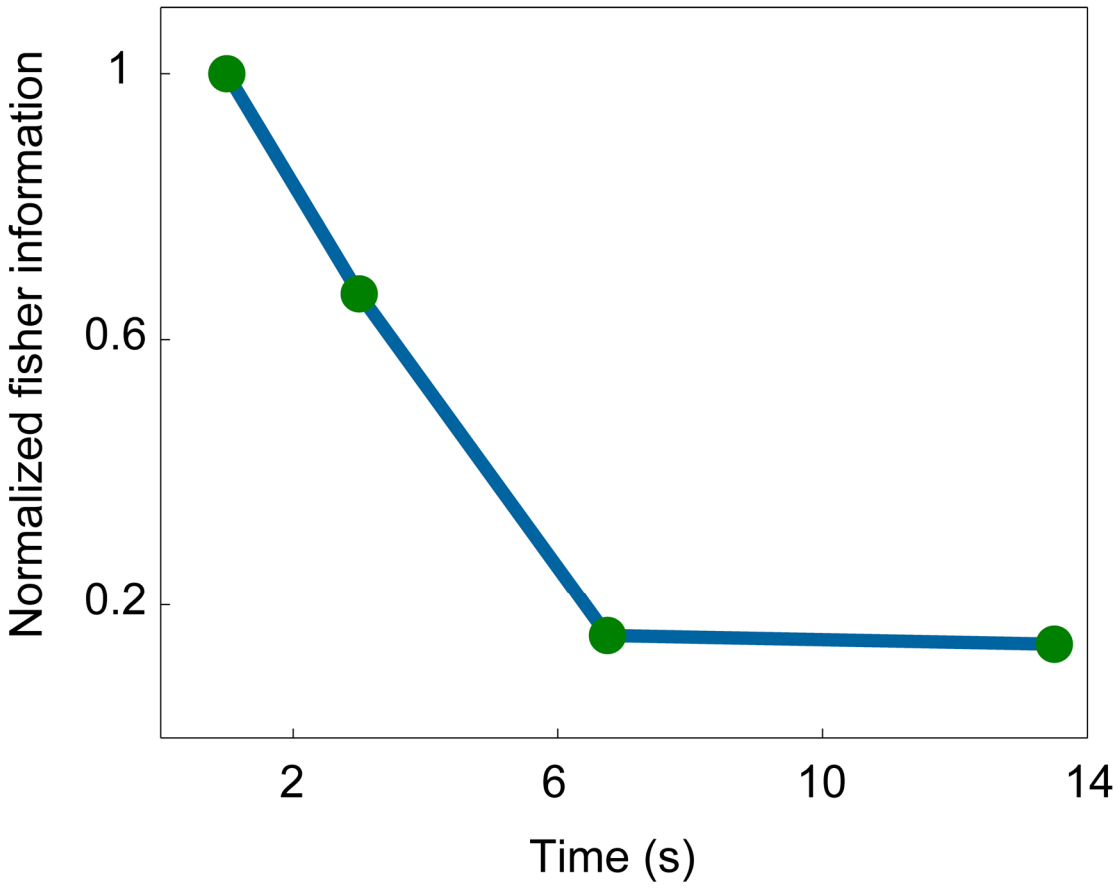
The results we present in this study are not consistent with any of the leading models of working memory circuits. Qualitatively, the dynamics we report are more complex than attractor network models. Yet the dynamics are much more constrained than the proposed alternatives such as reservoir network computational models (Appeltant et al., 2011; Bernacchia et al., 2011; Maass et al., 2002; Verstraeten et al., 2007) or feedforward networks such as Goldman et al. (2009) that have a great variety of heterogeneous responses in network nodes. Building computational networks that reproduce the dynamics we observe is an important step for understanding working memory cortical networks and generating hypotheses for future experiments.

Supplementary Figures



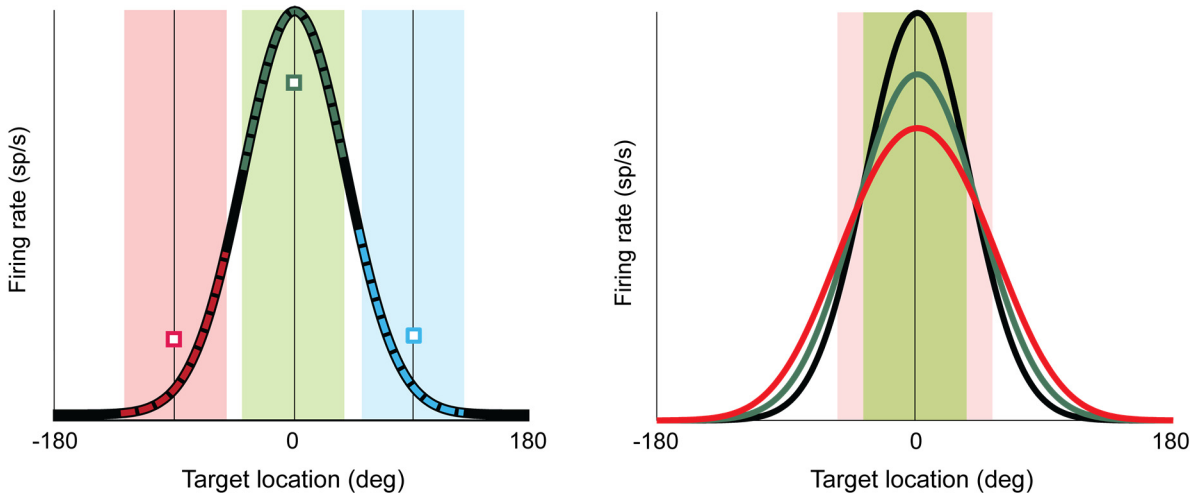
Supplementary Figure S4.1. Fixation breaks during the memory period.

Each bin shows the percentage of trials animals broke fixation at that time interval in the memory period. After a sharp initial decrease, percentage of trials that fail remains constant throughout the memory period, although we see a notable modulation at the time of reward delivery (blue lines). Red rectangles indicate the end of each of 3 memory periods (5s, 7.5s, and 15s).



Supplementary Figure S4.2. Fisher information drops over the memory period.

Fisher information in neural tuning curves (see Fig. 4.2b-c) at 4 different times in the memory period normalized to the earliest time interval. Fisher information drops to 14% by the end of the memory period.



Supplementary Figure S4.3. Appearance of the tuning curve of a single neuron as a function of delay using single-unit recording methods.

As delay increases the range of random drift also increases. Left - The tuning curve (black trace) shown represents the neuron's preferred direction without random drift. The three points represent how the tuning curve appears to single-unit recording methods after a delay during which random drift can occur. Green rectangles depict the range of random drift around their center. Over many trials with random drift, the neuron's activity may take any values across the red, green, and blue parts of it's tuning curve. The three points show the effect of averaging firing rates over trials for three different target locations (-90, 0, and 90 deg), showing a reduced amplitude and wider flanks. Right - Tuning curve without drift (black trace) and how tuning curves appear when averaging firing rates over many trials with random drift (green and red traces, drift amount shown by the green and red rectangles respectively). The tuning curve appears broader with increasing drift.

References

Amit, D. (1992). Modeling brain function: The world of attractor neural networks (Cambridge: Cambridge University Press).

Amit, D.J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex *7*, 237–252.

Appeltant, L., Soriano, M.C., Van der Sande, G., Danckaert, J., Massar, S., Dambre, J., Schrauwen, B., Mirasso, C.R., and Fischer, I. (2011). Information processing using a single dynamical node as complex system. Nat. Commun. *2*, 468.

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

Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. *14*, 366–372.

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

Bruce, C.J., and Goldberg, M.E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. *53*, 603–635.

Brunel, N. (1996). Hebbian learning of context in recurrent neural networks. Neural Comput. *8*, 1677–1710.

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

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

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

Ferrera, V.P., Cohen, J.K., and Lee, B.B. (1999). Activity of prefrontal neurons during location and color delayed matching tasks. Neuroreport *10*, 1315–1322.

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

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

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

Gabbott, P.L. a, and Bacon, S.J. (1996). Local circuit neurons in the medial prefrontal cortex (areas 24a, b, c, 25 and 32) in the monkey: I Cell morphology and morphometrics. J Comp Neurol *364*, 567–608.

Goldman, M.S. (2009). Memory without Feedback in a Neural Network. Neuron *61*, 621–634.

Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C.D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. J. Neurosci. *30*, 916–929.

Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–50.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput. *14*, 2531–2560.

Di Pellegrino, G., and Wise, S.P. (1993). Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J. Neurosci. *13*, 1227–1243.

Sommer, M.A., and Wurtz, R.H. (2001). Frontal eye field sends delay activity related to movement, memory, and vision to the superior colliculus. J Neurophysiol *85*, 1673–1685.

Takeda, K., and Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. J Neurophysiol *87*, 567–588.

Takeda, K., and Funahashi, S. (2004). Population vector analysis of primate prefrontal activity during spatial working memory. Cereb Cortex *14*, 1328–1339.

Umeno, M.M., and Goldberg, M.E. (2001). Spatial processing in the monkey frontal eye field. II. Mem. Responses. J Neurophysiol *86*, 2344–2352.

Verstraeten, D., Schrauwen, B., D'Haene, M., and Stroobandt, D. (2007). An experimental unification of reservoir computing methods. Neural Networks *20*, 391–403.

Wang, X.-J. (2009). Attractor Network Models. In Encyclopedia of Neuroscience, L.R. Squire, ed. (Oxford: Oxford: Academic Press), pp. 667–679.

Wang, X.-J., Tegnér, J., Constantinidis, C., and Goldman-Rakic, P.S. (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*, 1368–1373.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. *17*, 431–439.

Zhou, X., Katsuki, F., Qi, X.-L., and Constantinidis, C. (2012). Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex. J. Neurophysiol. *108*, 31–38.

From error to error, one discovers the entire truth.

Sigmund Freud

Chapter 5

Conclusions

How does information in working memory degrade, and what does that tell us about the neural substrate that supports working memory function? To answer this question we investigate behavioral and neural correlates of information degradation in spatial working memory. We interpret our results in the context of computational models of memory networks. We begin by classifying information degradation as one of two broad types: 1) degradation due to interference from irrelevant information and 2) degradation due to imperfect maintenance mechanisms that lose information with the passage of time. The specific patterns of failure we observe in each type of memory degradation reveal new mechanisms employed in memory circuits to maintain information.

5.1 Degradation due to irrelevant information

Chapters 2 and 3 address memory degradation due to interference from irrelevant information. In Chapter 2 we look at how information that was previously (but no longer) relevant interferes with actively maintained memory information. The pattern of behavioral errors reveals that on trials in which previously relevant information causes interference, it does not do so by completely overwriting the current memory. Instead, memory of a currently maintained target location is biased toward the previously memorized location. The memory representations from the previously and currently memorized targets combine into an intermediate representation, with a mix of information from both stimuli. Bias as a function of the relative distance between a previous and a current target location is well-fit by a Gabor function, with a peak error at around 60 degrees. This spatial profile, with a strong bias when the previous and current targets are close together but no bias when they are far apart, is immediately reminiscent of continuous attractor models (Compte et al., 2000; Wang, 2009). In these models competing memory information for two stimuli merges into a single representation that contains some information about both stimuli. But merging only happens when two stimuli are similar enough that they recruit overlapping neuronal populations (e.g. in the spatial memory case, when the previous and current target locations are close together). When generally non-overlapping populations are recruited by two stimuli (e.g. previous and current target locations are far apart) the stronger of the two representations drives global inhibitory connections that quash the weaker representation. The stronger representation remains unbiased. We posit that neural activity encoding the previous target location persists into the current trial and

competes with a current target representation. Our simulations using a continuous attractor network show that activity from the previous trial can in fact persist for some time even after the network has received an end-of-trial reset signal. This information then biases the encoded location in the subsequent trial, successfully replicating the spatial profile of error we observe in behavioral responses.

In Chapter 3 we record from memory cells in frontal eye fields to test the prediction that persistent activity from the previous trial biases encoding of the current target location. We find that a strong and clear neural signal encoding the previous target does in fact persist after the end of the previous trial and into the fixation period of the subsequent trial. That is, cells that were activated by the stimulus in the previous trial maintain a fraction of that activity into the next trial. However, we also find that this residual ghost of neural activity does not persist until the end of the current trial's memory period when activity is read out. Therefore it is unlikely to directly drive behavioral bias. Notably, we also find when the previous and current targets are close together, the population activity bump encoding the current target location is shifted *away* from the previous target location, not toward it as we predicted. This repulsive shift persists until the end of the memory period and is therefore likely to drive the observed behavioral responses.

How is it that behavioral responses are biased toward the previous target location but the neural activity that drives behavior is biased away from that location? To reconcile these observations we propose a model in which receptive fields in memory circuits shift in response to memory targets, and in which the fields do not completely revert back to

their original locations after the end of a trial. If receptive field shifts are not accounted for, the small amount of residual shift toward the previous target location produces a population activity 'bump' of neural activity that appears to shift away from that location. This accounts for the repulsive pattern we observe in the neural data. The receptive field shift also leads to a relative increase in the number of neurons that 'vote' for a behavioral response close to the previous target. This results in behavioral responses that are biased toward the previous memory location. Receptive field shifts like those in our model have been observed in FEF and other circuits involved in visual or memory processing in a variety of tasks that require the execution of saccades or the use of spatial attention (Colby and Goldberg, 1992; Connor et al., 1997; Sommer and Wurtz, 2006; Tolias et al., 2001; Walker et al., 1995; Zirnsak et al., 2014). Evidence showing that receptive fields converge toward saccadic endpoints or attended locations have given rise to the hypothesis that these receptive field changes may reflect redistribution of attentional resources to increase processing of task-relevant parts of space. Researchers have begun to consider whether receptive field changes also take place in other cognitive tasks that may benefit from an increase in local processing of space such as spatial working memory (Merrikhi et al., 2014). We provide the first evidence that these changes occur in spatial memory tasks and show functional significance for overall error reduction in the memory network as a result of receptive field convergence. That is, receptive field convergence makes the memory network more resistant to noise, reducing overall error, but introduces systematic spatial bias effects as a function of previous memoranda.

We also looked at the temporal profile of behavioral bias due to interference from previous memoranda. We hypothesized that as the memory period of the previous trial gets longer, the memory trace of its stimulus would decay. The weakened memory from the previous trial would therefore have a smaller effect on the subsequent trial. Contrary to our prediction we found that the bias toward the previous target location increased asymptotically as the memory period length increased. Like the spatial profile of the bias, this temporal profile initially appeared to implicate attractor dynamics. The memory trace in an attractor network never decays, and there is an initial buildup of activity after stimulus presentation before the network reaches the steady memory state. Therefore, with increasing memory period length, the strength of the memory representation of the previous trial will asymptotically grow, increasing its effect on the subsequent trial.

If this prediction were true, behavioral bias should depend on the memory period length of the *previous* but not the current trial. Instead, we found that the increase in bias toward the previous trial's target was a function of the *current* trial's delay length. This finding suggests that increase in bias is caused by decay of the current trial memory. Furthermore, with just a single memory store, memory decay would affect both the previous and current memory. Since the previous memory is farther back in time the relative strength of the current memory to the previous memory would always increase. We successfully modeled the finding that bias increases with the memory period length of the current trial by positing that behavior depends on two memory stores – a quickly decaying unbiased store and a sustained store susceptible to bias. Evidence of short-lasting memory stores such as iconic memory have been previously identified in both

behavioral (Averbach and Coriell, 1961; Averbach and Sperling, 1961; Dick, 1974; Pinto et al., 2013; Sligte et al., 2008; Sperling, 1960, 1963; Vandenbroucke et al., 2014) and neurophysiological (Bisley et al., 2004; Pasternak and Greenlee, 2005; Sligte et al., 2009) studies. The behavioral response in our model initially depends on the short-lasting veridical store and is therefore unbiased. As the unbiased store decays, the behavioral response is driven primarily by the sustained but biased store, leading to a saturating increase in bias as a function of the length of the current trial's memory period. On the one hand, mechanisms that allow a memory store to sustain information over extended periods of time may make that store 'sticky', causing persistence of irrelevant information. On the other hand, a veridical store that can be readily cleared of all information may be unable to maintain information for extended periods. The tradeoff between information persistence and information accuracy in any single store may lead to the use of multiple stores to maximize performance in tasks that rely on working memory.

Future directions

Future experiments should aim to look for direct evidence of receptive field changes in memory tasks. Receptive fields of memory cells can be mapped using visual probes presented at locations tiling the visual space. We can compare receptive field centers of cells mapped during a memory task to those mapped during a fixation task without memory to determine whether they have shifted toward the memorized target location. By mapping receptive fields at different intervals in the memory period (early, middle, late) we can determine the time course of the shift. The functional role of receptive field

changes can also be evaluated. We can ask whether trials in which convergence was greater also show reduced error in memory performance. Finally, an important open question is how receptive field convergence toward a spatial location produces behavior that is attractive toward that location. Although a number of observations have tied receptive field changes to attractive behavioral and perceptual bias, no compelling readout has been proposed for how this comes to be without assuming that output connections are also remapped. In fact, standard readout methods predict behavioral responses and perception that is biased *away* from the location of receptive field convergence in the absence of downstream receptive field changes. We can therefore ask whether downstream readout circuits 'know' that upstream receptive fields have remapped – that is, do output connections of FEF neurons also remap to activate different sets of neurons in downstream areas such as the superior colliculus? By simultaneously mapping receptive field changes in memory circuits and in downstream circuits more closely tied to behavioral responses we can determine how the reallocation of neural resources in cognitive tasks is interpreted by the readout mechanisms that drive behavior.

5.2 Degradation due to accumulation of error over time

In Chapter 4 we address degradation of working memory information due to imperfect maintenance mechanisms that lose information with the passage of time. A number of studies have looked at neural correlates of spatial working memory maintenance and decay (Bruce and Goldberg, 1985; Chafee and Goldman-Rakic, 1998; Constantinidis et

al., 2001; Ferrera et al., 1999; Funahashi et al., 1989, 1993; Fuster and Alexander, 1971; Kojima and Goldman-Rakic, 1982; di Pellegrino and Wise, 1993; Sommer and Wurtz, 2001; Takeda and Funahashi, 2002, 2004; Umeno and Goldberg, 2001). Most studies that directly interrogate spatial working memory circuits use short memory periods, between 1 and 3 seconds. These studies typically find that single-unit activity does not appreciably decay during the memory period even though error in behavioral responses increases. The computational network models they have inspired reflect this, having stable, non-decaying memory states and error arising through other mechanisms such as random drift rather than decay of neural activity. However, due to visual responses early in the trial, anticipatory responses late in the trial, and short memory periods these studies are not able to resolve time constants of decay larger than a few seconds. In fact, no studies have systematically looked at single-unit neural activity over the course of long memory periods.

In Chapter 4 we look at how memory decays over long (15 second) memory periods and whether the pattern of memory degradation is consistent with current models of working memory circuits. We find that unlike attractor network predictions, memory activity does decay. In fact, the offset times of memory cells we record are broadly distributed throughout the memory period and 80% of cells we record lose memory tuning before the end of the 15 s memory period. Furthermore, in each cell stimulus-elicited elevated activity is lost at nearly the same time on each trial indicating that the offset time for each cell is hardwired.

The variety of time constants of memory decay across the cell population brings to mind heterogeneous models of working memory circuits. These models are inspired by studies showing PFC memory traces that appeared to greatly vary during the memory period and from one cell to another (e.g. Baeg et al., 2003; Brody et al., 2003; Harvey et al., 2012; Jun et al., 2010). Cells may turn on and off at any time during the memory period, and memory contents can only be decoded accurately by considering the entire cell population. Experimental evidence supporting heterogeneous models often comes from experiments with a small number (often just two) of possible memoranda, and those memoranda are frequently not optimized to cells' receptive fields. In addition, a number of studies use electrode array recordings and offline sorting algorithms to collect and isolate memory neurons. Firing rate responses that appear to be heterogeneous could therefore be due to sub-optimal driving of memory cells or sub-optimal isolation and classification. In the experiment described in Chapter 4, we use continuous memoranda covering all of visual space and online experimenter-monitored cell isolations to ask if the onsets and offsets of cell responses are heterogeneous. We find that once cells turn off, they are not reactivated later in the same memory period. We also find that cells without early memory activity do not show memory activity later in the memory period.

The findings presented in Chapter 4 indicate that memory cells in working memory circuits come on early in the memory period, maintain memory activity for a fixed amount of time, and then turn off and stay off for the remainder of the memory period. This is a drastically different picture than the well-behaved permanently sustained memory responses of attractor network cells. And although there is some degree of

heterogeneity in cell offset times, they behave more regularly than suggested by heterogeneous models.

Future directions

How can we model the findings of memory decay in our data? One possibility is that cells may have an internal timer, after which they shut off until the next external input signal activates the memory circuit. Another possibility is that the variety in decay time constants is due to network dynamics. Reservoir computing networks are one class of networks that can generate a variety of onset and offset times in cell responses (Appeltant et al., 2011; Bernacchia et al., 2011; Maass et al., 2002; Verstraeten et al., 2007). They accomplish this through random connectivity structure in recurrent connections and weights. However our data are more structured than predictions of standard reservoir computing networks. For example, we don't find evidence of cell onsets late in the memory period and cells do not turn back on once they have turned off. Network effects can also take the form of multiple decaying attractors with independent time constants. That is, cells may be grouped into separate populations, with each population having a different time constant of decay. Whether starting from the perspective of reservoir computing, attractor networks, or single-cell timers, specific implementations of these frameworks that reproduce our results are necessary. This will help make clear the computational network properties and constraints needed to model the data presented in Chapter 4. New and refined models will lead to the next round of hypothesis generation for future experiments.

Why do memory cells have repeatable offset times that span the memory period? One hypothesis is that memory networks may require more neural resources to encode a memory stimulus than to maintain it. This could be due to alternate stores (e.g. synaptic storage) that may increasingly store some fraction of the memory information over the course of the memory period, thereby reducing the demands on active firing for storage. It may also be the case that the memory network is designed to free neural resources from stimuli that have already been processed and are no longer novel. Neurons may be progressively released from memory maintenance after initial encoding of a stimulus, making them available to encode additional memory targets. To test this idea we can ask if cells whose activity has decayed during maintenance of a stimulus turn back on when a second memory stimulus is presented while the first stimulus is still being maintained. On one extreme, the second stimulus may only be encoded by cells that turned off, indicating that offset times serve to release cell in the memory circuit to make them available for maintenance of new information. On the other extreme, cells that turned off may not respond at all to the second stimulus. Finally, a more complicated pattern of responses may reveal new ideas for the function of cells' repeatable and distributed offset times.

References

Appeltant, L., Soriano, M.C., Van der Sande, G., Danckaert, J., Massar, S., Dambre, J., Schrauwen, B., Mirasso, C.R., and Fischer, I. (2011). Information processing using a single dynamical node as complex system. Nat. Commun. *2*, 468.

Averbach, E., and Coriell, A.S. (1961). Short-term memory in vision. Bell Syst. Tech. J. *40*, 309–328.

Averbach, E., and Sperling, G. (1961). Short-Term Storage of information in Vision. Inf. Theory 196–211.

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

Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. *14*, 366–372.

Bisley, J.W., Zaksas, D., Droll, J.A., and Pasternak, T. (2004). Activity of neurons in cortical area MT during a memory for motion task. J. Neurophysiol. *91*, 286–300.

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

Bruce, C.J., and Goldberg, M.E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. *53*, 603–635.

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

Colby, C., and Goldberg, M. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. Science *255*, 90–92.

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

Connor, C.E., Preddie, D.C., Gallant, J.L., and Van Essen, D.C. (1997). Spatial attention effects in macaque area V4. J. Neurosci. *17*, 3201–3214.

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

Dick, A.O. (1974). Iconic memory and its relation to perceptual processing and other memory mechanisms. Percept. Psychophys. *16*, 575–596.

Ferrera, V.P., Cohen, J.K., and Lee, B.B. (1999). Activity of prefrontal neurons during location and color delayed matching tasks. Neuroreport *10*, 1315–1322.

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

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

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

Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C.D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. J. Neurosci. *30*, 916–929.

Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–50.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput. *14*, 2531–2560.

Merrikhi, Y., Parsa, M., Albarran, E., and Noudoost, B. (2014). Maintenance of spatial information gates the processing of incoming visual information in area V4. In Society for Neuroscience,.

Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.

Di Pellegrino, G., and Wise, S.P. (1993). Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J. Neurosci. *13*, 1227–1243.

Pinto, Y., Sligte, I.G., Shapiro, K.L., and Lamme, V. a (2013). Fragile visual short-term memory is an object-based and location-specific store. Psychon. Bull. Rev. *20*, 732–739.

Sligte, I.G., Scholte, H.S., and Lamme, V. a F. (2008). Are there multiple visual short-term memory stores? PLoS One *3*, 2–10.

Sligte, I.G., Scholte, H.S., and Lamme, V. a F. (2009). V4 activity predicts the strength of visual short-term memory representations. J. Neurosci. *29*, 7432–7438.

Sommer, M. a, and Wurtz, R.H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. Nature *444*, 374–377.

Sommer, M.A., and Wurtz, R.H. (2001). Frontal eye field sends delay activity related to movement, memory, and vision to the superior colliculus. J Neurophysiol *85*, 1673–1685.

Sperling, G. (1960). The information available in brief visual presentations. Psychol. Monogr. Gen. Appl.

Sperling, G. (1963). A Model for Visual Memory Tasks. Hum. Factors J. Hum. Factors Ergon. Soc. *5*, 19–31.

Takeda, K., and Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. J Neurophysiol *87*, 567–588.

Takeda, K., and Funahashi, S. (2004). Population vector analysis of primate prefrontal activity during spatial working memory. Cereb Cortex *14*, 1328–1339.

Tolias, A.S., Moore, T., Smirnakis, S.M., Tehovnik, E.J., Siapas, A.G., and Schiller, P.H. (2001). Eye Movements Modulate Visual Receptive Fields of V4 Neurons. Neuron *29*, 757–767.

Umeno, M.M., and Goldberg, M.E. (2001). Spatial processing in the monkey frontal eye field. II. Mem. Responses. J Neurophysiol *86*, 2344–2352.

Vandenbroucke, A.R.E., Sligte, I.G., Barrett, A.B., Seth, A.K., Fahrenfort, J.J., and Lamme, V. a F. (2014). Accurate metacognition for visual sensory memory representations. Psychol. Sci. *25*, 861–873.

Verstraeten, D., Schrauwen, B., D'Haene, M., and Stroobandt, D. (2007). An experimental unification of reservoir computing methods. Neural Networks *20*, 391–403.

Walker, M.F., Fitzgibbon, E.J., and Goldberg, M.E. (1995). Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye movements. J Neurophysiol *73*, 1988–2003.

Wang, X.-J. (2009). Attractor Network Models. In Encyclopedia of Neuroscience, L.R. Squire, ed. (Oxford: Oxford: Academic Press), pp. 667–679.

Zirnsak, M., Steinmetz, N.A., Noudoost, B., Xu, K.Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. Nature *507*, 504–507.