Cognitive transforms in perception and memory

Matthew F. Panichello

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the
Princeton Neuroscience Institute
Advisers: Timothy J. Buschman and Nicholas B. Turk-Browne

September 2020

Abstract

Visual perception seems to provide a direct and immediate view onto the outside world. In reality, it is an active and adaptive process. Cognitive factors, such as our prior knowledge and goals, transform the information streaming in from the retina to create a reconstruction of our environment tailored to our needs. The study of these cognitive influences on perception and their underlying mechanisms falls within the purview of visual cognition. If the principles gleaned from these studies can inform our understanding of cognition more generally, however, it is necessary to test if these principles generalize to domains beyond visual perception (and, if not, to understand if these principles can at least provide a useful basis for comparison and understanding). Towards that end, the work in this thesis examines how expectation and attention influence visual perception and additionally interrogates these same processes in the context of working memory. In the realm of expectation, we show that percepts reflect a weighted average of sensory information and prior knowledge, biasing percepts towards expected values. We find that these biases persist in working memory, accumulating over time to counteract memory noise. In the realm of attention, we find that, once attended, both percepts and memories are represented using radically different (i.e. orthogonal) patterns of neural activity relative to their unattended state. Furthermore, in this new post-attentional subspace, perceptual and mnemonic codes are reorganized in a way that allows task-relevant features to be decoded and task-irrelevant features to be abstracted away. This transform may selectively gate the influence of perceptual and mnemonic representations on other cognitive processes. In both the case of expectation and attention, these common principles uniting perception and memory coexist with key differences. For instance, learning modifies the influence of expectations on memory faster than on perception, and attention biases competition between perceptual but not mnemonic representations. Together, these results suggest that while the cognitive transforms observed in perception do generalize to other domains, they may be actualized by distinct mechanisms.

Acknowledgements

The work described in this thesis has benefited enormously from the support and talents of a large group of colleagues, mentors, and friends. First and foremost, I'm grateful for my advisers, Tim Buschman and Nick Turk-Browne. Their clarity of thought and insight have broadened my horizons and allowed me to produce this body of work to the best of my ability. And their encouragement and kindness made even challenging times manageable. Because of them I've grown both as a researcher and as a person. It is a privilege to have been so fortunate twice in one go.

The members of the Buschman and Turk-Browne labs were an enormous source of help of all kinds. Hannah Weinberg-Wolf and Britney Morea taught me to work with the animals and kept the lab running. Nikola Markov showed me that manual work in the lab demands as much skill and creativity as analysis. Sina Tafazoli provided crucial help with microstimulation. Pavlos Kollias was a great source of help as we figured out how to train animals together. Megan deBettencourt and Ghootae Kim volunteered to spend many, many hours in the scanner room as I collected fMRI data. Mariam Aly, Nick Hindy, Ben Hutchinson, and Vik Rao Bejjanki kept an open door and were happy to answer analysis questions. Judy Fan and Cameron Ellis provided extensive, insightful feedback. I'm grateful to all of the people above, as well as Flora Bouchacourt, Natalia Cordova, Norbert Cruz-Lebron, Victoria Jackson-Hanen, Caroline Jahn, Peter Kok, Alex Libby, Camden MacDowell, Neeraja Rajagopalan, Greg Nowak, Anna Schapiro, and Motoaki Uchimura for their insights, feedback, and camaraderie.

I also owe a great deal to the broader neuroscience and psychology community at Princeton. Brian DePasquale and Jonathan Pillow were fantastic collaborators on the work described in Chapter 3. I learned much from them and the work benefited tremendously from their input. Sabine Kastner provided insightful feedback throughout my graduate career, first as a rotation adviser and then as a member of my generals and thesis committees. Her questions and comments are routinely the sort that stick with you for months afterwards and change your view of your data. I also benefited greatly from the frequent exchange of expertise and equipment across the porous boarder between Buschman and Kastner lab. Thank you to Ruja Chen, Manoj Eradath, Ian Fiebelkorn, Na Yeon Kim, Ryan Ly, Anne Martin, Mark Pinsk, Tara van Viegan, and Xiofang Yang for helpful discussions. Similarly, the members of the Ghazanfar lab have been great neighbors. I'm grateful to Jeremy Borjon, Asif Ghazanfar, Lauren Kelly, Diana Liao, Daniel Takahashi, Yayoi Teramoto Kimura, Thiago Varella, and Yisi Zhang for their contagious enthusiasm for science and good humor. Beyond the second floor corridor, Alex Piet taught me maximum likelihood estimation and pushed me to fit models

to behavioral data. Jon Cohen, Sebastian Musslick, Ken Norman, and the entire PDP contingent provided feedback at key junctions. Carlos Brody communicated formative ideas during my first year coursework and my generals exam. Olga Lositsky was always willing to share her clear thinking. Garrett McGrath skillfully answered many computing questions, Alex Michaud kept me on track to graduation, and the staff at laboratory animal resources ensured the animals were well cared for.

For their care and support over the years, I'm grateful for Ryan Ly, Kelsey Ockert, Alex Piet, Diana Liao, Alex Song, Gecia/Lee/Ashlyn Gundermsdorff, Kevin Miller, Greg Nowak, Talmo Pereira, Kim Stachenfeld, Abby Hoskin, Ben Hoskin, Bas van Opheusden, Sam Ritter, Luis Piloto, Sarah Cutter, the Lositsky family, and many others.

I wouldn't have had the opportunity to interact with any of these people, or the ability to make the most of that time, if it weren't for my pre-graduate mentors Lisa Feldman Barrett and Moshe Bar. I can't overstate the impact of the training and inspiration that they provided. I'm also indebted to Olivia Cheung, Kestas Kveraga, Max Chaumon, Marilee Ogren, Elizabeth Kensinger, Neil Wolfman, Mark O'Connor, Tim Duket, Chris Constas, Avner Ash, and Shelly Chamness for formative instruction, advice, and encouragement.

Finally, I'm thankful for my family. Thank you to my parents for teaching me to be curious and to take risks. Most of all, thank you to my parents and also Stephanie, Dan, and Annabel for all of your love and support over the years.

Contents

Chapter 1

Introduction

1.1 Vision as a model system for studying cognition

The goal of cognitive neuroscience is to explain how the nervous system produces intelligent behavior. Over the past few decades, the study of visual cognition – or how humans and other animals perceive and make judgments about the visual world – has proven to be a particularly productive domain in which to study the relationship between brain and behavior.

The notion that vision is a productive domain in which to study cognition may be counter-intuitive. As a first-order approximation, visual processing is often described as an automatic, feedforward process displaying little of the flexibility associated with intelligent behavior. Visual input is understood to pass through a hierarchical sequence of cortical regions along the ventral surface of the brain and on to prefrontal cortex. At each level of the hierarchy, neural activity reflects increasingly abstract features of the visual input (DiCarlo et al., 2012), from simple edges to longer contours and surfaces to view-invariant high-level semantic information (Freiwald and Tsao, 2010; Yamins et al., 2014). From this perspective, knowledge of the visual input arriving from the retina and the sequence of feedforward transforms executed along the ventral stream are sufficient to describe visual processing.

Although this perspective provides a reasonable first pass description, it has been long known to psychology and neuroscience that visual processing critically depends on cognitive factors as well (Gregory, 1980; Von Helmholtz, 1866). The history and goals of the observer matter. Consider the painting by Joseph Turner shown in Fig. 1.1.

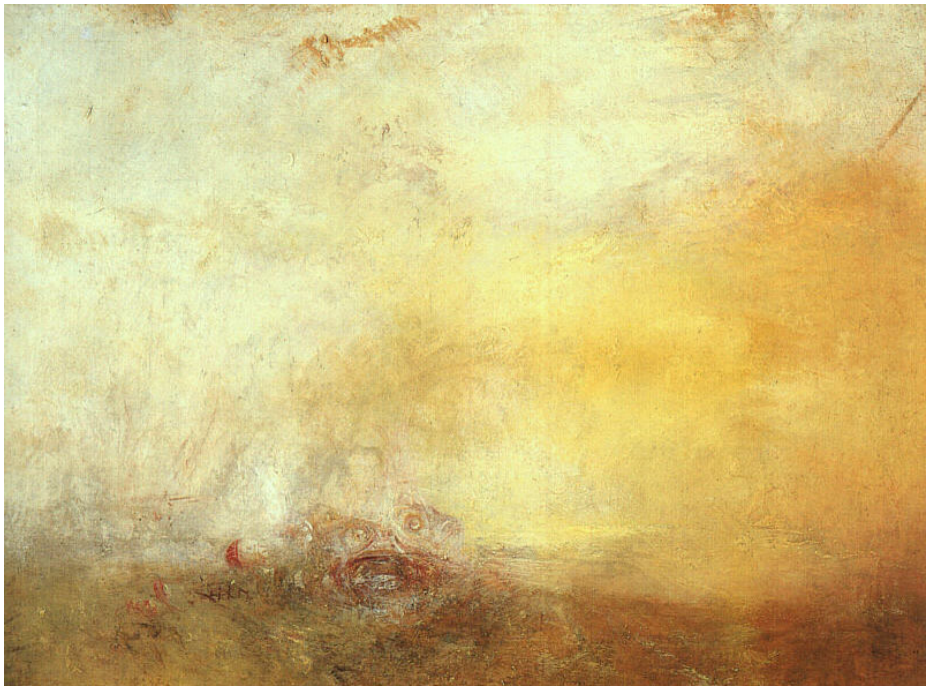


Figure 1.1: Painting by J. M. W. Turner, ca. 1845

Take another look after learning that the work is titled 'Sunrise with Sea Monsters' and pay particular attention to the lower central portion of the painting. For many, this additional context and the cue to focus on a particular aspect of the visual input substantially alters their perceptual interpretation.

Such cognitive influences allow vision to be adaptive - sensory input is transformed not just according to a set of predefined rules, but also according to the knowledge and goals of the observer. Two cognitive factors, expectation and attention, are each particularly crucial for overcoming a severe limitation of visual processing. First, visual sensory inputs are often ambiguous. The presence of unaccounted-for fluctuations ("noise") in patterns of light striking the retina as well as in neural activity, coupled with the projection of a 3D visual environment onto a 2D retinal sheet, means that many different visual scenes can evoke any particular pattern of sensory stimulation. The visual system must rely on expectations derived from prior experience to resolve this ambiguous input (de Lange et al., 2018; Panichello et al., 2013). Second, primates receive far more visual input through the retina (an estimated tens of millions of bits per second, Koch et al., 2006) than we are able to process in depth. While we tend not to notice this processing bottleneck, it is easily revealed under laboratory conditions that cause human subjects to routinely fail to notice large, sudden changes in their visual environment (Simons and Levin, 1997; Simons and Rensink, 2005). As a result, we use attention to focus on aspects of the visual input that are most likely to be informative, given our

current needs.

These cognitive influences on vision are worthy objects of study in their own right because humans are highly visual animals and disorders of visual cognition can have profound consequences for behavior. Schizophrenic hallucinations are thought to emerge from expectation gone awry (Corlett et al., 2019), and deficits in attention result in profound disturbances of awareness (Vallar, 1998).

But studying cognitive influences on vision is also a strong leverage point for cognitive neuroscience more generally. Cognitive neuroscience is difficult because the objects of study are latent variables (representations) and processes (computations) that are reflected in behavior and neural activity but can be extremely difficult to infer if not sufficiently constrained. A common strategy in a visual cognition experiment is to present a stimulus drawn from a well-parameterized space (e.g., a set of colors) and then ask the subject to reconstruct the stimulus or render a judgment in that space. Such a design substantially constrains the representational space (the range of possible stimuli) as well as the start- and end-state of a subject's representation on a particular trial. Intervening representations between the presentation of the stimulus and the subject's report can be inferred via behavioral modeling or decoding of neural data. And computations can be inferred from the manner in which representations transform over the course of the trial. Thus, cognitive factors can be studied under relative controlled conditions, and interpretation of neural activity aided by the comparatively well-understood anatomy and physiology of the visual system.

The work in this thesis leverages this strategy to examine how expectation and attention transform visual representations. These studies build on a large body of work examining how these processes influence visual perception. If the principles gleaned from studies of perception can inform cognitive neuroscience more generally, however, it is necessary to test if these principles generalize beyond perception to other types of information processing (and, if not, to understand if these principles can at least provide a useful basis for comparison and understanding). As a step towards that end, we explore how expectation and attention influence visual perception and then examine how these same processes influence visual memory. Our work in each of these domains is summarized in more detail below.

1.2 Transformation of visual percepts and memories by expectation

Visual input is often noisy and ambiguous, posing a challenge for perception: what is the most likely cause of the sensory input? This challenge can be addressed by relying on previous experience and learning to generate a percept reflecting a 'best guess', given the sensory data and one's prior knowledge. Accordingly, humans perceive stimuli faster and more accurately when they have sufficient knowledge to generate informed expectations about what they may see (de Lange et al., 2018; Panichello et al., 2013). Subjects are better at identifying objects when they are shown in their typical environment (e.g., a spatula in a kitchen, Bar, 2004; Biederman, 1972; Biederman et al., 1982; Davenport and Potter, 2004; Palmer, 1975) or are primed with an object drawn from the same context (Gronau et al., 2007, Sachs et al., 2011). More generally, subjects detect noisy stimuli more quickly and at greater levels of degradation when primed with information related to the identity of the object (Eger et al., 2007; Esterman and Yantis, 2010; Melloni et al., 2011; Reynolds, 1985).

While it is clear that expectations influence perception, a precise computational account of how expectations are integrated with incoming sensory information is an active area of investigation. As a standard for comparison, the behavior of humans and other animals is often compared to the hypothetical behavior of an ideal Bayesian observer. Bayesian inference describes the optimal way in which an observer should combine noisy sensory information with prior expectations to infer the state of the world. Imagine that a participant is asked to estimate the horizontal position of a briefly presented stimulus. Imagine also that in addition to the imperfect information obtained from the stimulus, the participant knows from previous trials that central positions are more likely that peripheral ones. These two sources of information can be represented as distributions over the range of possible positions (i.e., the likelihood and the prior, Fig. 1.2). To estimate the horizontal position as a Bayesian observer would, these two distributions should be multiplied, yielding a posterior distribution that reflects information from both sources, weighted by their precision. The result is a more precise estimate that is biased towards the prior. Remarkably, human behavior in perceptual tasks is often consistent with a Bayesian observer (e.g., Girshick et al., 2011; Jazayeri and Shadlen, 2010; Stocker and Simoncelli, 2006).

These results suggest that the nervous system may combine sensory input and expectations in a manner consistent with a Bayesian observer (Aitchison and Lengyel, 2017), or at least implements an approximation of this process (e.g., compare Ma et al., 2006 and Sohn et al., 2019 for accounts which do / do not rely on explicit neural representations of likelihoods and priors). Accordingly,

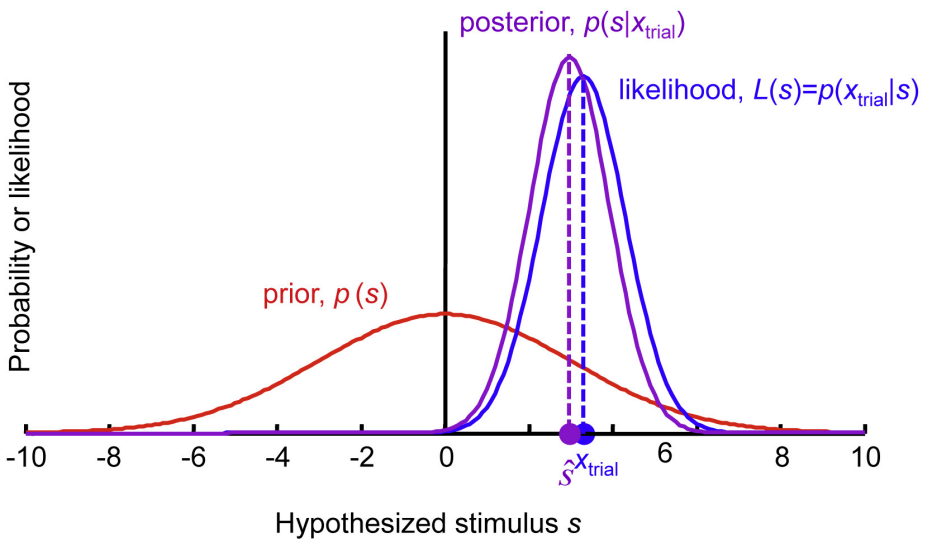


Figure 1.2: Multiplying the likelihood over stimulus values derived from sensory evidence (blue) by the prior (red) yields a lower-variance posterior estimate biased towards the mean of the prior (purple). Image reproduced from Ma, 2019.

neural representations of expected stimuli are more precise (Brandman and Peelen, 2017; Hindy et al., 2016; Kok et al., 2012) and biased towards prior expectations (Kok et al., 2013; van Bergen et al., 2015).

In Chapter 2 of this thesis, we build on his work by explicitly testing if percepts reflect a precision-weighted combination of sensory input and prior expectations, as predicted by Bayesian inference. To accomplish this, we presented participants with faces drawn from a continuous parameterized identity space (analogous to the range of horizontal positions in Fig. 1.2). Additionally, we trained participants to linearly map a continuum of tones onto this stimulus space through associative learning. As a result, tones became predictive of face identity. Thus, after training, we could precisely control the mean of a subject's prior and likelihood function on a trial-by-trial basis by presenting a particular tone and face, respectively. Drawing inspiration from the multisensory integration literature (Ban et al., 2012; Murphy et al., 2013), we generated precise behavioral and neural predictions for the discriminability of specific tone-face pairs assuming optimal Bayesian inference and tested these predictions using psychophysics and fMRI. We show that behavior in this task is consistent with Bayesian inference and provide evidence for neural fusion of sensory inputs and expectations.

In Chapter 3 of this thesis, we explore if this inference process continues beyond perception and into visual memory. Continuing to apply Bayesian inference after encoding is potentially useful because noise continues to accumulate in visual representations as they are held in working memory.

Subjects recall visual stimuli less accurately with increasing memory delays (Pertzov et al., 2017; Rademaker et al., 2018; Schneegans and Bays, 2018; Shin et al., 2017) and the neural representations underlying visual memories have been shown to randomly diffuse over time (Wimmer et al., 2014; Wolff et al., 2020), possible due to Poisson variability in spiking (Bays, 2015; Burak and Fiete, 2012). Therefore, Bayesian inference could help mitigate the deleterious effects of noise in working memory just as it does in perception.

To test this hypothesis, we collected behavioral data from human and non-human primate subjects on a task in which they were asked to remember colors drawn from a circular 'hue' space and reconstruct the colors in that space after a memory delay (Wilken and Ma, 2004). If subjects were applying prior knowledge (accrued over phylogenetic or ontogenetic time) and expected some hues more than others, then their responses on this task should be biased towards expected stimuli. Accordingly, reports on this task clustered around certain hues. Furthermore, if these biases reflect the influence of a prior in accordance with Bayesian inference, then biases should increase with memory noise because the prior will increasingly dominate the Bayesian update. This is exactly what we observed when we increased the delay length and memory load, both of which are predicted to increase the accrual of random error (Bays, 2015; Burak and Fiete, 2012). Finally, to explicitly test if these biases reflected expectations, we manipulated subjects' priors by manipulating the distribution of hues presented during the experiment, and found that these biases shifted towards the (newly) expected hues. We were able to recapitulate these phenomena using a dynamical model of memory which postulates random diffusion of memories coupled with systematic drift towards attractor states reflecting expected hues, linking these behavioral results to a rich theoretical and experimental literature on memory dynamics (reviewed in Brody et al., 2003; Chaudhuri and Fiete, 2016) and suggesting neural architectures that could implement this approximation to Bayesian inference.

1.3 Transformation of visual percepts and memories by attention

Attention is critical for managing the limited capacity of working memory, the mental workspace in which sensory inputs are integrated with remembered relevant sensory and cognitive variables and ongoing goals to guide behavior (Baddeley, 2003). The benefits of attention are easily revealed via visuospatial cueing. When subjects are informed where task-relevant stimuli will appear in the visual environment, they are faster and more accurate at detecting or making simple judgments about those

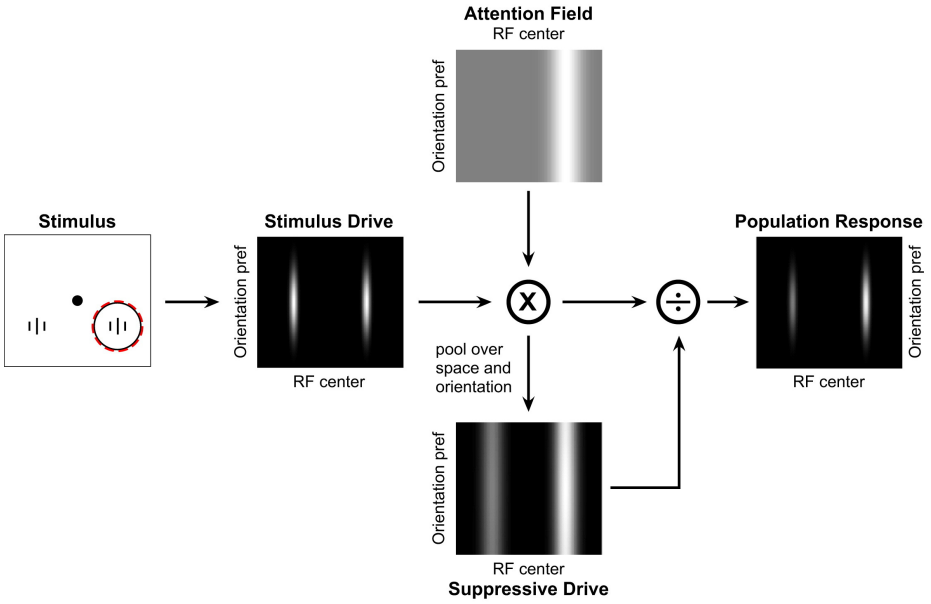


Figure 1.3: The normalization model of attention. An attentional field (top) multiplicatively scales the latent stimulus drive (left) to neurons with with receptive fields within the locus of attention prior to normalization by inhibitory suppressive drive (bottom). The result is an enhanced population responses to attended stimuli (right). Image reproduced from Reynolds and Heeger, 2009.

stimuli, even when their gaze is fixed (Posner, 1980). Superior performance at processing stimuli at these locations comes at the cost of deficits in processing stimuli presented elsewhere, consistent with the role of attention in prioritizing information in a limited capacity system (Desimone and Duncan, 1995).

Visuospatial attention exerts its effects, in part, by modulating evoked neural responses to stimuli in visual cortical and subcortical regions. Attention has been shown to multiplicatively scale neural responses (McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999), sharpen neuronal tuning curves (Spitzer et al., 1988, Martinez-Trujillo and Treue, 2004), and increases contrast gain (Reynolds et al., 2000). This diversity of results is parsimoniously explained by a computational model in which attention multiplicatively scales the latent effective drive to a neuron, which is then normalized by the pooled activity of the population (Reynolds and Heeger, 2009, Fig. 3.3). The precise manifestation of these effects depends on the breadth of attention, but the ubiquitous phenomenon is that neural responses to attended stimuli are enhanced relative to unattended stimuli.

Substantial progress has been made in identifying the networks responsible for controlling these differential responses to attended and unattended stimuli. Neuroimaging work in humans revealed that attending to a spatial location engages a distributed frontal-parietal network, with weaker effects observed in visual cortex, suggesting that attention signals originates in fronto-parietal regions

and influence posterior cortex via feedback (Kastner et al., 1999). This work has been supported by primate electrophysiology studies showing an anterior-to-posterior flow of volitional attentional signals (Buschman and Miller, 2007) and causal work showing that microstimulation of prefrontal cortical regions elicits attention-like effects in visual cortex (Moore and Armstrong, 2003; Moore and Fallah, 2001).

Top-down enhancement of visual evoked responses may thus allow attended stimuli to propagate through the cortical hierarchy and into the lateral prefrontal cortex, a key locus of working memory. Accordingly, neurons in prefrontal cortex are sensitive to task-relevant information presented at attended, but not unattended locations (Everling et al., 2002). Attention may also filter neural signals by selectively routing information via synchrony (e.g. Bosman et al., 2012; Córdova et al., 2016; Saalmann et al., 2012).

While our ability to selectively attend 'externally' to percepts has been extensively studied, our ability to attend 'internally' to memory representations is much less understood. Indeed, it was initially thought that the last point at which attention could influence visual memory representations was shortly after encoding, while perceptual traces of the stimuli were not yet extinguished (Phillips, 1974; Sperling, 1960). Subsequent work, however, revealed that working memory representations that are cued as task-relevant are recalled with greater accuracy (Griffin and Nobre, 2003; Landman et al., 2003; Pertzov et al., 2013; Sligte et al., 2008). Furthermore, such 'retro-cueing' paradigms drive neural activity in the same fronto-parietal network known to mediate visuospatial attention (LaBar et al., 1999; Lenartowicz et al., 2010; Nee and Jonides, 2009; Nobre et al., 2004). However, it is unclear if perceptual and mnemonic attention share a deep mechanistic homology or are distinct processes that simply share common functional consequences.

In Chapter 4 of this thesis, we build on this work by explicitly comparing and contrasting attention in perception and memory. To avoid confusion, we refer to perceptual and mnemonic attention using the distinct terms 'attention' and 'selection'. To study the neural basis of these processes, we recorded and analyzed neural activity from prefrontal, parietal, and visual cortex as monkeys performed a variant of the color working memory task described above. Critically, by spatial cuing stimuli as task-relevant either before or after encoding, we were able to encourage the animals to use attention or selection (respectively) to improve their report of the cued item.

During selection, we found that information about which stimulus was task-relevant emerged first in prefrontal cortex. This suggests that selection, like attention, is directed from prefrontal cortex. Strikingly, in prefrontal cortex similar population codes were used to signal the task-relevant stimulus, suggesting that a common mechanism directs external and internal attention. However, these

control signals were coded differently for attention and selection in other brain regions, suggesting the underlying mechanisms may diverge and uniquely affect stimulus representations. Accordingly, while both attention and selection enhanced information about the cued stimulus in neural firing rates, attention primarily modulated representations in visual cortex during encoding (and this information propagated downstream to parietal and prefrontal cortex), while selection primarily acted within prefrontal and parietal regions. Furthermore, selection did not decrease information about the uncued stimuli, suggesting that, unlike attention, selection does not bias competitive dynamics between representations (Desimone and Duncan, 1995). However, attention and selection displayed an unexpected common feature; both transformed stimulus representations into a new subspace, suggesting a means which the contents of working memory can be selectively 'gated' into a state that enables the preparation of cognitive or motor plans.

Chapter 2

Transformation of percepts by expectation

2.1 Abstract

Humans perceive expected stimuli faster and more accurately. However, the mechanism behind the integration of expectations with sensory information during perception remains unclear. We investigated the hypothesis that such integration depends on "fusion" — the precision-weighted averaging of different cues informative about stimulus identity. We first trained subjects to map a range of tones onto faces from a gender continuum via associative learning; these two features served, respectively, as expectation and sensory cues to gender. We then tested specific predictions about the consequences of fusion by manipulating the congruence of these cues in psychophysical and fMRI experiments. Behavioral judgments and patterns of neural activity in auditory association regions were consistent with fusion of sensory and expectation cues, providing evidence for a precise computational account of how expectations influence perception.

2.2 Introduction

Prior experience and learning guide perception, allowing for fast and accurate processing of sensory input that is often noisy and ambiguous (Hutchinson and Turk-Browne, 2012; Oliva and Torralba, 2007; Panichello et al., 2013). As a result, it has long been suggested that perception may be best understood as a form of probabilistic inferences about the outside world, rather than a veridical

representation of sensory inputs Gregory, 1980; Von Helmholtz, 1866. However, the precise computation by which expectations and sensory information are combined to refine perception remains an active area of investigation (de Lange et al., 2018).

Bayesian inference describes the optimal means by which an observer can combine noisy sensory information with prior expectations to infer the state of the world. Strikingly, human behavior in perceptual tasks has been shown to be consistent with a Bayesian observer (Girshick et al., 2011; Jazayeri and Shadlen, 2010; Stocker and Simoncelli, 2006), engendering proposals that neural systems may combine sensory inputs and expectations in this optimal fashion (for a recent review, see Aitchison and Lengyel, 2017).

Evidence from human neuroimaging has revealed that perceptual representations display characteristics consistent with Bayesian inference, in which neural representations reflect a "fused" precision-weighted average of feature estimates provided by sensory inputs and expectations. The result is a more precise representation that is biased towards the prior expectation. Accordingly, expected stimuli are more easily decoded from patterns of neural activity (Brandman and Peelen, 2017; Hindy et al., 2016; Kok et al., 2012), consistent with an increase in precision. And reconstructed neural representations are biased towards prior expectations (Kok et al., 2013; van Bergen et al., 2015). However, because these studies did not independently manipulate sensory- and expectation-based estimates, the degree to which representations were in fact fused remains unclear.

Here, we build on this work by explicitly testing if percepts reflect a weighted combination of sensory and expectation cues. To investigate this question, we drew inspiration from the multisensory integration and cue combination literatures, which have developed rigorous methods for testing for fused representations in a computationally analogous context (i.e., the fusion of two sensory representations, Ban et al., 2012; Murphy et al., 2013). We combine analysis of psychophysical and neuroimaging data, motivated by formal models, to test the hypothesis that perceptual judgments and neural representations reflect a fusion of sensory and expectation cues.

2.3 Results

2.3.1 Learned tone-face associations bias behavior

To study fusion, we first established a novel set of expectations via associative learning in the domain of face perception. In addition to the gender information conveyed by the visual features of faces (i.e., "sensory" cues) we introduced a novel source of gender information by training subjects to

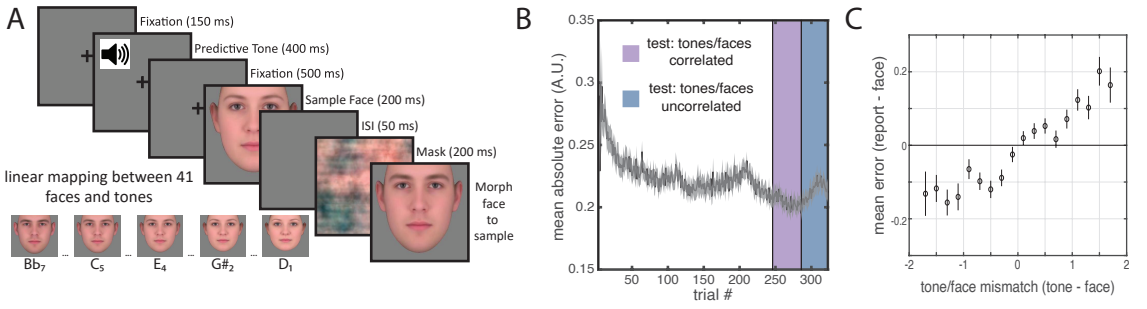


Figure 2.1: Learned tone-face associations bias behavioral reports. (A) Experiment 1 trial structure. On each trial, subjects were presented with a pure tone and an image of a face drawn from a continuous gender space. At the end of the trial, subjects continuously morphed a face through gender space to match the face they had seen. Inset: example mappings for 5 tone face-pairs; in this example, lower notes predict more feminine faces. (B) Mean learning curve across subjects. During an initial training phase (white region), the tones predicted the faces perfectly and subjects received feedback on their performance. The congruent test phase (purple) was identical to this training phase, except that subjects no longer received feedback. During the incongruent test phase (blue), the pairing of tone and face was random. Y-axis reflects error in gender space units: 0.05 units corresponds to 1 step in the 41-step space. Error bars reflect standard error of the mean. (C) Mean signed error (bias) as a function of tone-face mismatch during the incongruent test phase. Positive x-values indicate trials on which the tones predicted a more feminine face than that actually presented. Positive y-values indicate that subjects reported a more feminine face than that actually presented. Both axes are differences in gender space units; error bars are standard error of the mean.

linearly map auditory tones onto a gender continuum (i.e., "expectation" cues; Fig. 2.1a). We were not especially interested in gender per se, but chose this domain because face gender is amenable to multivariate decoding from fMRI (Contreras et al., 2013; Kaul et al., 2011) and because face perception is associated with a well-defined cortical network (Dekowska et al., 2008).

On each trial of the experiment, subjects ($N = 48$) were presented with tone-face pairs (Fig. 2.1a). The faces were drawn from 41 points of a continuous gender space, varying between the average female face and the average male face. During an initial training phase, each of 41 tones, which varied sequentially in pitch, predicted the face with the corresponding index on the continuum deterministically. After the offset of the tone and face, subjects were presented with a face randomly drawn from gender space and had to morph it to match the face they had just seen as closely as possible. Subjects received feedback on their performance and became more accurate over the course of this training phase (Fig. 2.1b, white region). Across subjects, the mean change in error between the first twenty and last twenty trials was 0.054, significantly below zero ($t(47) = 5.019$, $p = 7.88e - 06$).

Following the training phase, there were two test phases during which subjects no longer received feedback on their performance. During the congruent test phase (Fig. 2.1b, purple region) the tones predicted the faces deterministically, as during training. During the incongruent test phase

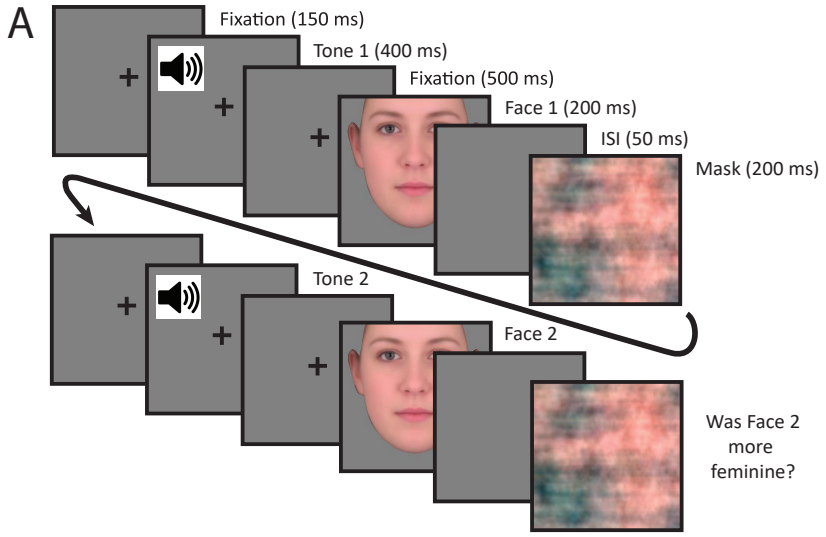
(Fig. 2.1b, blue region), the pairing of tone and face was random. Error was significantly greater during the incongruent test phase than during the congruent test phase ($t(47) = 2.824$, $p = 0.007$). Errors during the incongruent test phase were influenced by the sign and magnitude of the tone-face mismatch. When the tone predicted a more feminine phase then the one that was actually shown, subjects tended to report a more feminine face (and vice-versa, Fig. 2.1c). The mean correlation between tone-face mismatch and mean signed error across subjects was 0.110, significantly greater than zero ($t(47) = 10.204$, $p = 1.66e - 13$). Together, these results suggest that subjects learned the mapping between the tones and the gender space and that this association was sufficient to generate expectations that could bias behavior.

2.3.2 Behavioral evidence for expectation and sensory fusion

In experiment 2, a new cohort of subjects ($N = 60$) was exposed to a linear mapping between the tones and faces. We then tested whether expectations and sensory information were integrated in a manner consistent with fusion using psychophysical techniques originally developed for studying cue combination in depth perception (Ban et al., 2012; Murphy et al., 2013).

After performing a task identical to the training phase from experiment 1, subjects performed a gender discrimination task. On each trial, they were presented with two tone-face pairs and had to report whether the second face was more feminine than the first (Fig. 2.2a). By systematically manipulating the predictive validity of the tones (Fig. 2.1b) we derived two tests for fusion.

The first test relates performance on ΔCongruent trials to performance on ΔFace and ΔTone trials. For a conservative null hypothesis, we assume that subjects still make use of the expectation cues, but that the information along each cue dimension is not fused and remains independent. Under these assumptions, the optimal solution is to recast task as a discrimination problem in a space with two orthogonal cue axes (Fig. 2.3a). The discriminability of the two tone-face pairs on ΔCongruent trials should then be the hypotenuse (root quadratic sum) of the discriminability when only the tones or faces differ (Fig. 2.3b). In the case of fusion, the sensory and expectation dimensions are not independent; observers take a precision weighted average of face and tone information for each pair to produce a single estimate in gender space (Fig. 2.3c). As a result, performance is suppressed in the ΔFace and ΔTone conditions because the difference along one dimension (i.e., face and tone, respectively) is diluted by averaging in the other dimension that does not contain a difference (i.e., tone and face, respectively). Performance in the ΔCongruent condition should thus exceed the root quadratic sum of these suppressed levels (Fig. 2.3d).



B

	ΔFace	ΔTone	ΔCongruent	ΔIncongruent
Pair 1				
face	g	g	g	g
tone	g	g	g	g
Pair 2				
face	g + Δg	g	g + Δg	g + Δg
tone	g	g + Δg	g + Δg	g - Δg

Figure 2.2: Discrimination task for testing fusion. (A) Discrimination task trial structure. On each trial, subjects were presented with two tone-face pairs and reported whether the second face was more feminine than the first. (B) Discrimination task trial types. g refers to a point in masculine-to-feminine gender space linked to a particular tone and face stimulus. Δg is calculated separately for each condition and varies over trials according to a staircasing algorithm that identifies a 75% accuracy threshold (final Δg = JND). For all trials, the first tone accurately predicted the first face. On ΔFace trials, the second face differed from the first. On ΔTone trials, the second tone differed from the first. On ΔCongruent trials, both the second tone and face differed from the first, and the second tone predicted the same gender as the second face. On ΔIncongruent trials, both the second tone and face differed from the first, but the second tone predicted a different gender than the face.

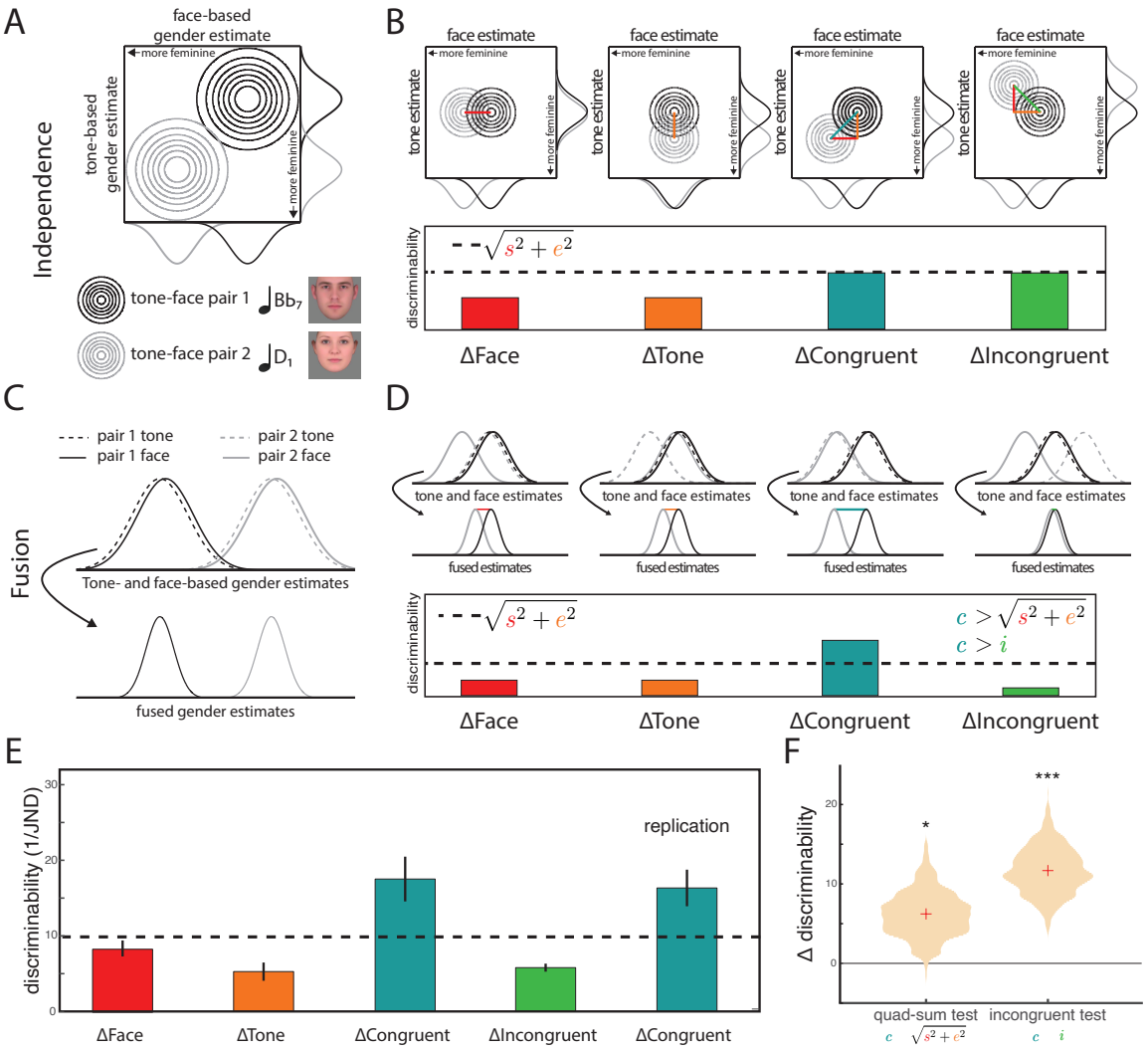


Figure 2.3: Theoretical predictions and behavioral results. (A) If the gender estimates elicited by the tones and faces are independent, the task can be recast as a linear discrimination problem in a 2D space. One axis represents gender estimates derived from the tones, and the other from the faces. Differences between the two pairs along either or both dimensions facilitate discrimination. The example depicts a ΔCongruent trial. (B) Predictions for each of the 4 trial types described in Figure 1c under independence. Because performance is proportional to the distance between means, performance in the ΔCongruent condition can be predicted by performance in the ΔFace and ΔTone according to the Pythagorean theorem (dotted line). Because cue axes are orthogonal, performance on ΔCongruent and ΔIncongruent trials is equal. (C) Under fusion, observers take a precision-weighted average of gender estimates from the faces and tones, resulting in a 1D discrimination problem. (D) As a result, performance in the ΔFace and ΔTone conditions will be suppressed because an uninformative cue has been averaged in. Because both cues are informative, performance in the ΔCongruent condition is unchanged relative to independence and thus exceeds the root quadratic sum (dotted line). ΔCongruent performance also exceeds the ΔIncongruent condition because the conflicting cues partially cancel each other out. (E) Sensitivity is measured as 1/JND, where JND reflects the Δg in each condition that produced 75% discrimination accuracy. Error bars reflect standard error of the mean across subjects. (F) Mean of the two fusion metrics across subjects (computed using the first of the two analyzed congruent staircases). Violin plots reflect bootstrapped distribution of the mean. Asterisks reflect bootstrap tests vs zero: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The second test for fusion compares performance on ΔCongruent vs. ΔIncongruent trials. An independence mechanism predicts that performance in the ΔCongruent and ΔIncongruent conditions should be equivalent because the distance between tone-face pairs in this bivariate space is the same (Fig. 2.3b). In contrast, under fusion, the averaging of conflicting cues in the ΔIncongruent condition will reduce the differences between pairs and hamper discrimination relative to the ΔCongruent condition (Fig. 2.3d).

To quantify the discriminability of the faces, we measured sensitivity in each subject to increments in gender space. Using a staircase procedure, we calculated the just noticeable difference (JND) in each condition as the distance between the two stimuli that resulted in 75% accuracy. We included twice as many ΔCongruent trials (with the original tone-face mappings intact) as other trial types to mitigate interference from the ΔIncongruent, ΔFace, and ΔTone conditions that violated these mappings. However, to equate power in calculating sensitivity, we dummy coded the ΔCongruent trials into two sets and ran independent staircases on them, with the second serving as a replication. Consistent with a fusion mechanism: (1) sensitivity in both ΔCongruent conditions exceeded the root quadratic sum of ΔFace and ΔTone ($t(59) = 2.218$, $p = 0.030$; $t(59) = 2.053$, $p = 0.045$), and (2) sensitivity in both ΔCongruent conditions exceeded ΔIncongruent ($t(59) = 3.995$, $p < 0.001$; $t(59) = 4.359$, $p < 0.001$, Fig. 2.3e-f).

2.3.3 Neural evidence for expectation and sensory fusion

Finally, in experiment 3, we used fMRI to explore how fusion was reflected in neural representations. A new cohort of subjects was first trained on tone-face mappings outside the scanner. Then, in the scanner, they performed an oddball task (Fig. 2.4a) while viewing a single tone-face pair on each trial, and we manipulated the validity of the tone-face relationship across trials.

Separate machine learning classifiers were trained to decode gender from trials in which (Fig. 2.4b): both the tone and face conveyed the same gender (ΔCongruent), the tone conveyed one gender and the face conveyed the other (ΔIncongruent), the tone conveyed gender and the face was neutral (ΔTone), and the tone was neutral but the face conveyed gender (ΔFace). That is, rather than having subjects discriminate tone-face pairs within trial (as in experiment 2), we used classifiers to measure discriminability across trials within each condition. Discriminability was quantified with d-prime, based on the proportion of test trials correctly vs. incorrectly labeled — e.g., how often the classifier guessed "male" for male trials (hits) vs. for female trials (false alarms).

We modeled the design and analysis of this imaging study after experiment 2 because the in-

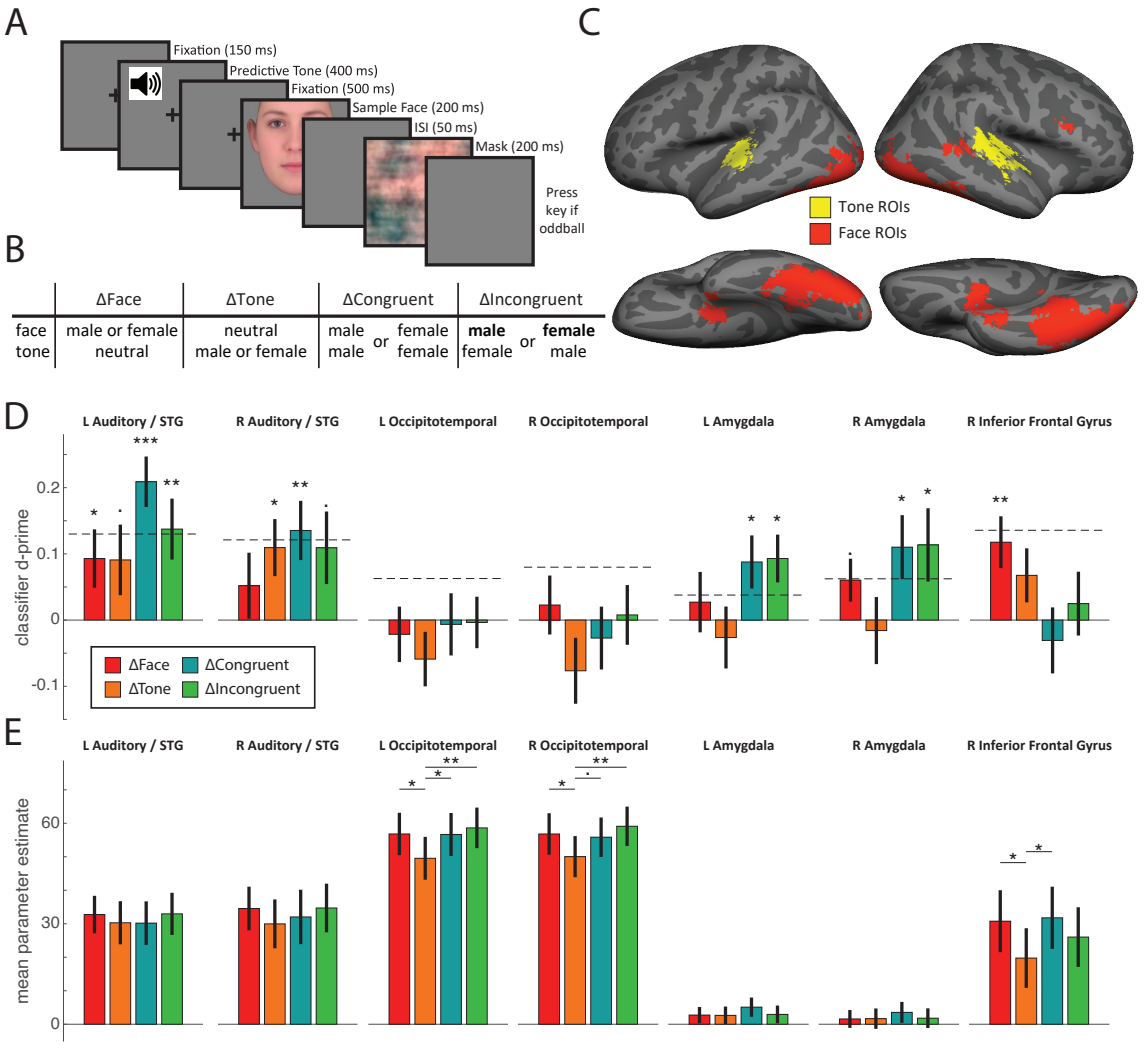


Figure 2.4: Neuroimaging design and results. (A) Each fMRI trial contained one tone-face pair. To ensure attention, subjects pressed a key on infrequent oddball trials where either the tone or face was replaced by two rapid tones or faces, respectively, and otherwise withheld their response. Oddball trials were discarded from analysis. (B) There were four cue conditions: only the face indicating gender (ΔFace), only the tone indicating gender (ΔTone), the face and tone indicating the same gender (ΔCongruent), and the face and tone indicating different genders (ΔIncongruent). Within each condition, the gender could either be male or female, and ΔIncongruent trials were labeled based on the gender of the face. We assessed the gender information in each condition by attempting to discriminate neural patterns for male and female trials using a classifier. (C) Regions of interest (ROIs), generated using automated meta-analyses of published neuroimaging data (Yarkoni et al., 2011). (D) Performance of the four classifiers (in units of d-prime) for each ROI. Dotted line indicates root quadratic sum of ΔFace and ΔTone d-prime, as in Figure 3. Error bars are standard error of the mean. L/R indicate left/right hemisphere, STG = superior temporal gyrus. Asterisks reflect bootstrap tests vs zero. (E) Mean univariate parameter estimates for each condition and ROI, averaged across trials, voxels, and subjects. Error bars are standard error of the mean across subjects. Lines and asterisks reflect dependent t-tests: · $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

dependence model remains a strong null hypothesis. Indeed, any region that contains a mixture of face- and tone-selective voxels will display a pattern consistent with independence. The introduction of cue conflicts on ΔFace and ΔTone trials is a calculated design decision that allows us to test that superior performance in the ΔCongruent condition is, in fact, due to fusion, as described above.

We performed an exploratory analyses to identify brain areas in which neural representations were consistent with fusion. To that end, we defined seven regions of interest (ROIs, Fig. 2.4c) that together cover a broad swath of face- and tone-sensitive areas. ROIs were defined using independent, automated metanalyses (Yarkoni et al., 2011). Single-trial parameter estimates were extracted from each voxel within each ROI and served as the features for classification analyses.

Across the seven ROIs, the pattern of performance across the four classifiers was most consistent with fusion in the left auditory cortex and superior temporal gyrus, with a weaker qualitative trend in right auditory regions (Fig. 2.4d). Accordingly, both fusion metrics were trending towards significance in this ROI (Fig. 2.5a-b, quadsum test: $p = 0.0868$, incongruent test: $p = 0.079$, bootstrap).

Classification performance was poor in the two ventral face ROIs (all classifiers $p > 0.05$ vs zero, Fig. 2.4d), perhaps due to weak topographic organizations for gender information. Classification in left and right amygdala was consistent with independence. In both regions, performance in the ΔCongruent and ΔIncongruent conditions was statistically indistinguishable (Fig. 2.5b), and while performance in the ΔCongruent was numerically greater than the ΔFace and ΔTone conditions, it did not exceed quadratic summation (Fig. 2.5a). Right inferior temporal gyrus displayed an unexpected pattern in which performance in the ΔCongruent and ΔIncongruent conditions were statistically equivalent and tended to be worse than in the ΔFace and ΔTone conditions. The comparison of ΔCongruent vs ΔFace was significant ($p = 0.018$, randomization test, all other $p > 0.10$). Such a pattern could be generated by a region in which separate populations of voxels encode face and tone information and engage in mutually inhibitory interactions.

The pattern of classification performance in left auditory cortex / STG was unrelated to the overall BOLD activity in each condition (Fig. 2.4e). Indeed, repeated measures ANOVA revealed that the mean parameter estimate in this region was not modulated by condition ($F(3, 93) = 0.42$, $p = 0.740$). Mean parameter estimate was significantly modulated by condition in left ($F(3, 93) = 3.04$, $p = 0.033$) and right ($F(3, 93) = 3.13$, $p = 0.030$) occipitotemporal cortex, as well as the inferior frontal gyrus ($F(3, 93) = 2.89$, $p = 0.040$). Post-hoc t-tests revealed that this was due to relatively lower parameter estimates in the ΔTone condition (Fig. 2.4e).

This fMRI design allowed us to plan, a priori, an additional "transfer" test for fusion (Ban et al.,

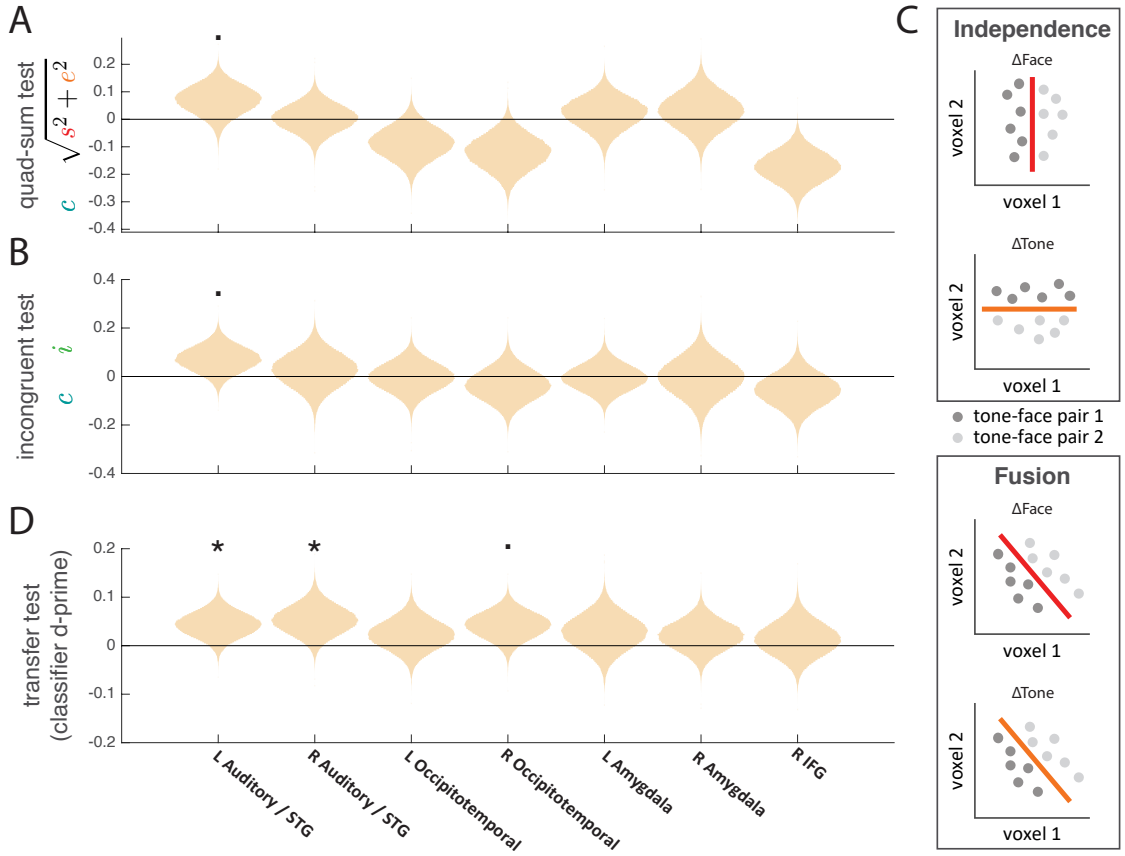


Figure 2.5: Classification-based fusion tests. (A) The mean quad-sum test statistic (difference in ΔCongruent d-prime and the root sum-squared ΔFace and ΔTone d-prime) for each ROI. (B) The mean incongruent test statistic (difference in ΔCongruent and ΔIncongruent d-prime) for each ROI. (C) Motivation for the transfer test. Under independence, face and tone information are coded along orthogonal axis in state space. A classifier trained to discriminate male and female faces will fail to discriminate the corresponding tones (and vice-versa). Under fusion, face- and tone-information are coded along a common axis, allowing a classifier to generalize from the ΔFace to the ΔTone condition (and vice-versa). Dots reflect hypothetical trials in a 2-voxel state space; red and orange lines reflect trained classification boundaries. (D) The mean transfer test statistic for each ROI. Violin plots reflect the bootstrapped distribution of the mean. Asterisks reflect bootstrap tests vs zero: · $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

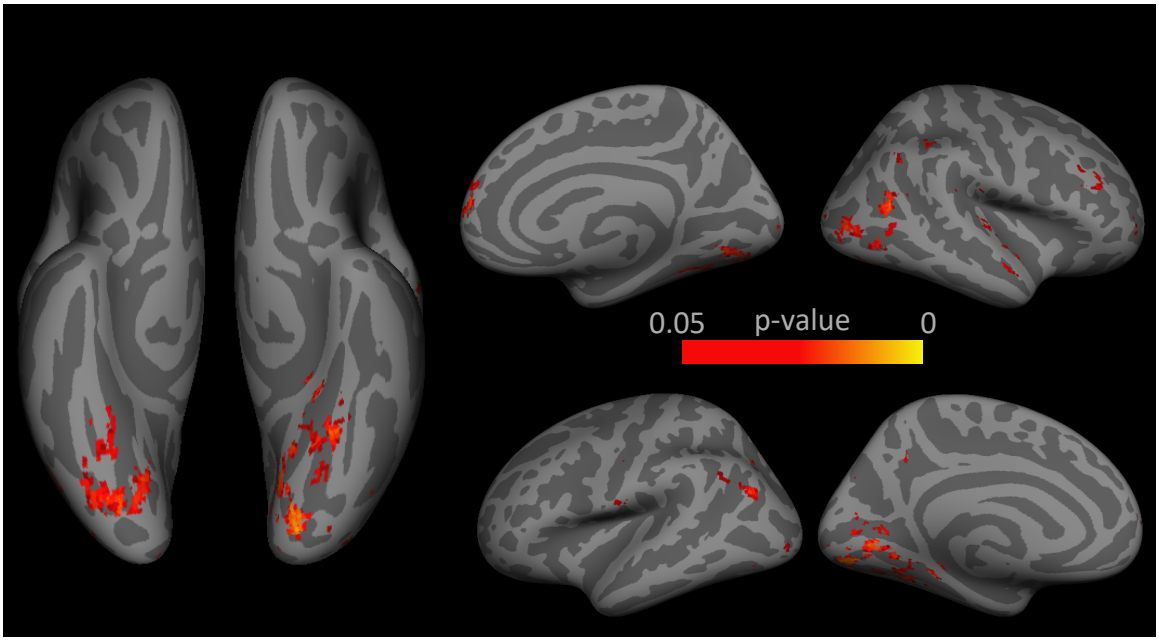


Figure 2.6: Results of transfer test searchlight. Thresholded p-map of the transfer test searchlight (corrected for multiple comparisons). Positions of assigned values correspond to searchlight centers.

2012; Murphy et al., 2013). Under fusion, gender estimates from tones and/or faces are encoded in the same representational space. Therefore, a gender classifier trained on ΔTone trials should be able to decode ΔFace trials, and a gender classifier trained on ΔFace trials should be able to decode ΔTone trials (Fig. 2.5c). In contrast, when tone and face information are independent, a classifier trained on ΔFace trials should not successfully decode ΔTone trials, and vice versa. The transfer test statistic was thus the average d-prime of a classifier trained and tested in this manner, averaged across folds. Consistent with the first two fusion tests, we observed significant transfer in left auditory cortex / superior temporal gyrus (Fig. 2.5d, $p = 0.042$). We also observed significant transfer in right auditory cortex / STG ($p = 0.033$) and a trending effect in right occipitotemporal cortex ($p = 0.060$). No other regions showed significant transfer.

Together, these analyses suggest auditory cortex / STG as a candidate region in which fusion may occur, particularly in the left hemisphere. Although it is important to emphasize that the first two fusion tests were trending in this region (i.e., $p < 0.10$), the probability of observing a trending or significant result for all three fusion tests in any region by chance was low ($p = 0.046$, randomization test).

To ensure that we did not miss any areas with strong patterns of fusion that fell outside our regions of interest, we re-ran the three fusion tests across cortex using a searchlight approach. These searchlights did not reveal any regions that passed the first two fusion tests after correcting for

multiple comparisons. However, the transfer test searchlight revealed two large clusters in bilateral inferior temporal cortex, as well as smaller clusters in right auditory cortex (Heschl's gyrus) and frontal, occipital, and parietal regions (Fig. 2.6, Section 2.3.3). These results suggest that, after training, sensory stimuli and learned cues that evoke expectations about those stimuli drive neural representations in a common manner throughout cortex.

Anatomical region	Hemi	Cluster size	min(p)	Coordinates (x y z)
Lingual Gyrus	L	1349	0.011	52 19 32
Occipital Fusiform Gyrus	R	1053	0.012	28 25 34
Paracingulate Gyrus	R	134	0.025	40 90 43
Lateral Occipital Cortex	L	87	0.021	69 28 48
		33	0.029	62 31 45
		25	0.041	61 18 33
		12	0.038	59 30 53
Heschl's Gyrus	R	70	0.032	18 54 43
Precentral Gyrus	L	34	0.040	76 61 42
Middle Frontal Gyrus	R	32	0.034	24 78 48
		29	0.029	18 72 52
Lateral Occipital Complex	R	28	0.037	18 34 58
Superior Temporal Gyrus	R	27	0.040	19 60 28
Occipital Pole	R	22	0.040	35 18 39
Central Opercular Cortex	L	17	0.030	66 57 48
Posterior Cingulate Gyrus	L	14	0.029	48 45 42
		11	0.041	50 38 52
Frontal Pole	R	12	0.041	30 88 39

Table 2.1: Transfer test searchlight results. Each line describes a cluster of significant voxels (corrected for multiple-comparisons) and the coordinates and location of the voxel with the minimum p-value in that cluster

2.4 Discussion

This study provides evidence that observers incorporate expectations into perceptual processing by fusing them with sensory inputs. Conflicting sensory and expectation cues led to a specific pattern of behavioral deficits in perceptual decision-making, consistent with fusion models of cue integration in which feature estimates from cues are weighted by their precision. Pattern classifiers trained to perform an analogous set of discriminations based on neural activity in left auditory regions displayed a similar pattern of performance. These results provide evidence that fusion is instantiated at the neural level and suggests a computational mechanism by which expectations enhance the discriminability of perceptual representations (Brandman and Peelen, 2017; Hindy et al., 2016; Kok et al., 2012).

Note that while we did not observe a decrease in mean bold activity in auditory cortex, as

observed in some other studies in regions which show enhanced discriminability of congruently cued stimuli (e.g., Kok et al., 2012), these results are not incompatible with proposals that expectations sharpen neural representations (de Lange et al., 2018; Kok et al., 2012). Indeed, by analogy, highly successful models of visual attention marry mechanisms which can sharpen neural representations with inhibitory dynamics that maintain constant levels of overall neural activity (Reynolds and Heeger, 2009).

Previous work in multisensory integration and cue combination has demonstrated that humans can fuse highly stable cues that are genetically programmed or acquired over a lifetime of experience (Alais and Burr, 2004; Ban et al., 2012; Dekker et al., 2015; Ernst and Banks, 2002; Murphy et al., 2013; Nardini et al., 2010). In particular, superior temporal gyrus and left auditory cortex have been shown to be sensitive to the conjunction of highly familiar visual and auditory cues (e.g. video and audio of a person speaking, Callan et al., 2003; Hein et al., 2007; Kreifelts et al., 2007; Miller and D'Esposito, 2005). Here we show that the human brain flexibly leverage similar computational principles to integrate newly predictive information. This might explain how humans deploy recently learned environmental regularities in the service of faster and more accurate perceptual judgments (e.g., Esterman and Yantis, 2010; Turk-Browne et al., 2010). Whether similar learning mechanisms govern both the rapid emergence of fusion in adults and the slower development of cue integration in children remains an intriguing and open question.

The present work has several limitations that should encourage further investigation. First, we explored fusion for only one type of feature: the gender of face stimuli. Future work could examine whether the present findings generalize to other features and feature-selective cortical areas. Second, while we did not observe any evidence for fusion in ventral visual regions, this absence of an effect should be interpreted with caution because classification performance in these regions was generally poor. This may be driven in part by less clear topography for identity-level information in these regions, in contrast with the tonotopic organization of auditory cortex. Alternate cover tasks that require more explicit judgements of face identity, as in previous work (Contreras et al., 2013; Kaul et al., 2011), may provide sufficient SNR to reveal clear patterns of discrimination performance.

Bayesian inference provides a computational account of how expectations and sensory information interact in perception. The mechanism by which this integration is accomplished is an active area of investigation and is likely to depend on the type of expectation under consideration. For example, expectations may be embedded in the structural organization of cortex, or actively applied in the form of input from other brain regions (de Lange et al., 2018). Recent work suggests that expectations may be generated by the hippocampus when they depend on recently learned arbi-

trary associations (Hindy et al., 2016; Kok and Turk-Browne, 2018), raising the possibility that the signatures of fusion we observe in auditory cortex may depend on hippocampal input.

2.5 Methods

2.5.1 Subjects

Forty-eight subjects participated in Experiment 1 (28 female, mean age 19.6). Sixty subjects (37 female, mean age 19.5) participated in Experiment 2. Thirty-two subjects (20 female, mean age 21.8) participated in Experiment 3. All subjects had normal or corrected-to-normal vision and provided written informed consent to a protocol approved by the Princeton IRB.

2.5.2 Stimuli

Visual stimuli consisted of 41 gender-morphed face stimuli (Zhao et al., 2011). Stimuli were generated by interpolating features between a composite male face and a composite female face. The gender of the faces was coded using an arbitrary numerical reference scheme ranging from -1 to 1 in 0.05 increments, with -1 denoting the composite male face and 1 denoting the composite female face. Faces were presented centrally at fixation and spanned 4° of visual angle. Analysis of behavior from experiment 1 revealed that a large central proportion of this stimulus space was perceptually uniform. Stimuli drawn from this range were subsequently used for experiments 2 and 3 (see 'Experiment 1 Procedure', below).

Auditory stimuli consisted of 41 pure tones corresponding to musical notes ranging from D1 to B♭7 (36.7 to 3,951 Hz) in whole-step intervals. This tone space is perceptually uniform according to the MIDI pitch standard. The 41 tones were also assigned a numerical reference ranging from -1 to +1 in 0.05 increments. For all experiments, the tone-face mapping was counterbalanced such that higher frequency tones were mapped to more masculine faces for half of the subjects and to more feminine faces for the other half of subjects. The amplitude of the tone stimuli was adjusted to correct for increasing subjective loudness with increasing pitch.

2.5.3 Experiment 1

In experiment 1, we exposed subjects to a linear mapping between the face and tone spaces and tested if this association could bias behavior.

Subjects performed 325 trials of a delayed estimation task (Fig. 2.1a). On each trial, after being presented with a tone-face pair, subjects had to morph a second face stimulus to match the gender value of the face they had just seen as closely as possible. Subjects morphed the face by dragging a mouse cursor to the left or right edge of the screen, which either smoothly incremented or decremented the gender value of the face at the center of the screen. If a subject morphed the face to the edge of the gender space while the cursor was still at the screen edge, then morphing began to reverse direction in gender space. After identifying a desired face for their response, subjects halted morphing by returning their cursor to the center of the screen and submitted their response by pressing the space bar.

The first 246 trials of this task constituted a training phase in which the tones were perfectly predictive of the faces and subjects received feedback on their performance. Subjects received feedback in the form of points. To encourage precision, points increased logarithmically as error approached zero, up to a maximum of value of 2,000. Negative points were awarded for errors greater than 0.30 units in gender space (6 steps in the 41-step space). Each tone-face pair was presented six times. Trial order was generated randomly for each subject.

Subjects then completed two test phases (41 trials each) during which they no longer received feedback on their performance. During the first test phase, the tones remained perfectly predictive of the faces. During the second test phase, the mapping between tones and faces was randomly shuffled for each subject such that tone conveyed no information about the subsequent face. Within each test phase, each tone and face stimulus was presented once. Trial order was randomly generated for each subject.

Analysis of biases in subject's reports during the training and first test phase revealed that a large proportion of the face stimulus space was approximately perceptually uniform. Across the interval from -0.7 to 0.7, mean absolute bias was 0.038 and the maximum absolute bias was 0.096, or less than one and two steps in the 41-step space, respectively. Stimuli were therefore restricted to this range in Experiments 2 and 3 to satisfy the assumptions of the independence and fusion models.

2.5.4 Experiment 2

In experiment 2, subjects were again exposed to a linear mapping between the tones and faces, and subsequently performed a discrimination task designed to test whether tone and face information were integrated in a manner consistent with fusion.

Subjects first completed 123 trials of a delayed estimation task identical to the training phase of

experiment 1 (Fig. 2.1a). This corresponded to three exposures to each of 41 tone-face pairs. Trial order was generated randomly for each subject.

Subjects next completed a gender discrimination task. On each trial, they were shown two tone-face pairs and asked to report whether the gender value of the second face was more feminine than the first (Fig. 2.2a). For the first pair, the tone continued to predict the gender of the face with 100% validity. The gender space value of this first tone-face pair was randomly assigned to either 0.25, 0.20, 0.15, -0.15, 0.20, or 0.25 on each trial ('g' in Fig. 2.2b). For the second pair, however, the gender value of the second tone and/or face was systematically manipulated in a manner that sometimes corrupted the predictive validity of the tone (Fig. 2.2b): (1) On ΔFace trials, the second tone was identical to the first, but the second face differed in gender from the first by some increment in gender space. (2) On ΔTone trials, the second tone differed in gender from the first by some increment in gender space, but the second face was identical to the first. (3) On ΔCongruent trials, the second tone and face differed from the first tone and face by the same increment in gender space (the second tone on these trials was valid). (4) On ΔIncongruent trials, the second tone and face differed from the first tone and face by equal but opposite increments in gender space.

We measured the sensitivity of subjects to increments in gender space for each of these four trial types using separate staircases. Subjects were not told about the existence of the different trial types or staircases. Subjects began the discrimination task with 41 trials of ΔCongruent trials. Gender increments on each trial were selected using a Bayesian adaptive algorithm (Watson and Pelli, 1983) to converge on the increment at which subjects were correct 75% of the time. The purpose of this initial staircase was to avoid presenting invalid tones early on, which, coupled with the change in task phase, may have cued subjects that the relationship between the tones and faces had changed. Results from this staircase were not analyzed. After the initial 41 trials, five additional and separate 41-trial staircases began concurrently. Depending on the trial type, gender increments for the second tone and face were drawn from increments determined by a ΔFace staircase, a ΔTone staircase, a ΔIncongruent staircase, or one of two ΔCongruent staircases. Two ΔCongruent staircases were included to increase the overall validity of the predictive relationship, although these staircases were analyzed separately to equate statistical power across conditions. At the end of the staircasing procedure, the five estimated just noticeable difference values (in gender space units) were converted to sensitivity scores by taking their inverse (Ban et al., 2012).

2.5.5 Experiment 3

In experiment 3, subjects were again exposed to a linear mapping between the tones and faces, and subsequently participated in an fMRI experiment designed to identify regions supporting fusion.

The training task was identical to that used in experiments 1 and 2. Subjects underwent training over the course of two days, completing 369 trials on the day prior to their scan and an additional 123 trials immediately before the scan.

In the scanner, subjects were exposed to individual tone-face pairs while performing an oddball cover task that demanded attention to the tone and face stimuli. On each trial, subjects were presented with a tone and then a face with latencies identical to the training task, except that a second face never appeared for warping (Fig. 2.4a). Oddball trials occurred 18% of the time, containing either two tones or two faces in rapid succession in place of the typical one tone and one face. Subjects were asked to report the presence of oddballs with a button press and these trials were discarded from further analysis.

Subjects completed eight fMRI runs of 98 trials each (18 oddball trials, 80 non-oddball). Non-oddball trials consisted of eight different trial types (10 trials per condition), corresponding to the cross of gender (male or female) by tone-face relationship (ΔFace, ΔTone, ΔCongruent, ΔIncongruent; Fig. 2.4b). As in the behavioral experiment, we sought to measure the separability of gender representations as a function of condition. Rather than fix discriminability (i.e., behavioral accuracy at 75%) and measure the distance in stimulus space required, here we fixed the distance of the tones and faces in stimulus space and measured discriminability using multivariate pattern classifiers. Stimuli labeled "male" had a gender value of -0.6 (with +/- 0.1 units of jitter), stimuli labeled "female" had a gender value of 0.6 (with +/- 0.1 units of jitter), and neutral stimuli had a value of 0 (with +/ -0.1 units of jitter) in our gender coding scheme.

2.5.6 Image acquisition and analysis

Structural and functional MRI data were collected on a 3T Siemens Skyra scanner with a 16-channel head coil. Structural data was acquired using a T1-weighted magnetization prepared rapid acquisition gradient-echo (MPRAGE) sequence (1 mm isotropic). Functional data consisted of T2*-weighted multiband echo-planar imaging sequences with 48 oblique axial slices aligned to the AC-PC line acquired in an interleaved order (1,500 ms repetition time [TR], 40 ms echo time, 2 mm isotropic voxels, 96 x 96 matrix, 192 mm field of view, 64˚ flip angle). Data acquisition in each functional run began with 12 s of rest in order to approach steady-state magnetization. A B0 field map was

collected at the end of the experiment.

The first four volumes of each functional run were discarded for T1 equilibration. Functional data were preprocessed and analyzed using FSL (www.fmrib.ox.ac.uk/fsl), including correction for head motion and slice-acquisition time, spatial smoothing (5-mm FWHM Gaussian kernel), and high-pass temporal filtering (128-s period). Data were manually inspected for motion artifacts, spiking, and low SNR.

Regions of interest (ROIs) were defined based on automated meta-analysis in Neurosynth (Yarkoni et al., 2011) using "face" and "tone" as the search terms. ROIs were created by downloading statistical images from Neurosynth and binarizing the images such that significant voxels had a value of 1. Clusters with greater than 100 voxels were saved as masks (Fig. 2.4c), registered to each subject's functional space, and then re-binarized.

Classifier analyses were calculated based on the output of a single trial GLM (Aly and Turk-Browne, 2016; Hindy et al., 2016), which contained 98 task-related regressors: one for every trial in the run, modeled as 1.5 s boxcars from fixation onset to mask offset. All regressors were convolved with a double-gamma hemodynamic response function. The six directions of head motion were also included as nuisance regressors. Autocorrelations in the time series were corrected with FILM prewhitening. Each run was modeled separately in first-level analyses. First-level parameter estimates were registered to the participant's T2 image.

Classifier analyses were performed using custom scripts in MATLAB. These multivariate analyses were computed for each run and then averaged across runs (Aly and Turk-Browne, 2016). For each subject, ROI, and condition (ΔFace, ΔTone, ΔCongruent, ΔIncongruent), we trained a regularized logistic regression classifier (penalty = 1) to distinguish patterns of parameter estimates from individual "male" and "female" trials. Classifier performance was assessed using leave-one-out cross-validation (train on 19 trials, test on one). The average classifier accuracy across folds and runs was calculated separately for male and female test trials, and was converted to d-prime using the formula z(hit) – z(false alarm), where correct female test trials were coded as hits and incorrect male trials (i.e., labeled as female) were coded as false alarms. Two neural fusion metrics based on these classifiers were computed for each subject and ROI:

$$M1 = d'_{\Delta Congruent} - \sqrt{(d'_{\Delta Face})^2 - (d'_{\Delta Tone})^2} \tag{2.1}$$

$$M2 = d'_{\Delta Congruent} - d'_{\Delta Incongruent} \tag{2.2}$$

A third neural fusion metric was defined based on the ability of the classifier to generalize across face and tone information. Specifically, within each run, classifiers were trained to discriminate male and female trials from the ΔFace condition and tested on the ΔTone condition, and vice-versa. Generalization performance was averaged across these two folds and across runs.

The significance of each fusion metric for each ROI was assessed by computing the bootstrapped distribution of the mean (resampling subjects with replacement). To combine across fusion tests and control for multiple comparisons across ROIs, we also computed the probability that any of the 7 regions we investigated would display trending or significant effects for all three fusion tests by chance. To do this, we repeated the entire classification and bootstrapping procedure 1000 times, randomly permuting condition labels for each subject (e.g., all ΔFace trials could be relabeled ΔIncongruent), and recorded the number of instances in which at least one ROI displayed p-values < 0.10 for all three fusion tests. This tested the null hypothesis that there was no meaningful pattern of classification performance across the four conditions.

We additionally used searchlight analyses to compute the three neural fusion metrics across cortex. The procedure was identical to that described above for the ROIs, except that parameter estimates were registered to 2-mm MNI space and analyses were repeated for all 27-voxel cubes (3 x 3 x 3) centered on voxels in cortex according to the Harvard-Oxford structural atlas (Desikan et al., 2006). Group analyses comparing each test to zero across subjects were performed using random-effects nonparametric tests (as implemented by the 'randomise' function in FSL), corrected for multiple comparisons with threshold-free cluster enhancement (Smith and Nichols, 2009).

2.5.7 Model-based predictions for independence and fusion

Predictions for the discriminability of the tone-face pairs under independence and fusion were generated as follows (following Ban et al., 2012; Murphy et al., 2013). Each model takes as input the gender value of the two tones being discriminated (t_1 and t_2) and the gender value of the two tones being discriminated (f_1 and f_2). Additionally, each model has two parameters: the noise in the gender estimate from the face (σ^2_{face}), and the noise in the gender estimate from the tone (σ^2_{tone}). The qualitative predictions from each model described in the text do not depend on these parameters, except when σ^2_{face} and/or σ^2_{tone} are extremely large relative to the corresponding experimental manipulations.

Under independence, the gender representation for each tone-face pair are Gaussian distributions in a bivariate gender space. One axis corresponds to gender estimates derived from faces. The second

axis corresponds to gender estimates derived from tones. The means μ and covariance matrix C of these distributions are parameterized as follows:

$$\mu_1 = \begin{bmatrix} f_1 \\ t_1 \end{bmatrix} \tag{2.3}$$

$$\mu_2 = \begin{bmatrix} f_2 \\ t_2 \end{bmatrix} \tag{2.4}$$

$$C = \begin{bmatrix} \sigma^2_{face} & 0 \\ 0 & \sigma^2_{tone} \end{bmatrix} \tag{2.5}$$

The axis of discrimination was taken as the line of optimal discrimination, i.e., the line passing through the mean of both distributions. Each bivariate distribution is projected onto this axis, resulting in two 1D Gaussian distributions with means μ'_1 and μ'_2 and variance σ^2. D-prime is then calculated as

$$d' = \frac{\mu'_1 - \mu'_2}{\sqrt{\sigma^2}}. \tag{2.6}$$

Under fusion, gender representations are gaussians in a univariate gender space that reflects a weighted average of face and tone information. The means μ of the two distributions are defined as:

$$\mu_1 = k_{face} f_1 + k_{tone} t_1 \tag{2.7}$$

$$\mu_2 = k_{face} f_1 + k_{tone} t_1 \tag{2.8}$$

Where k_{face} and k_{tone} sum to 1. Assuming optimal Bayesian inference and gaussian noise (Ernst and Bülthoff, 2004), k_{face} and k_{tone} are:

$$k_{face} = \frac{\frac{1}{\sigma^2_{face}}}{\frac{1}{\sigma^2_{face}} + \frac{1}{\sigma^2_{tone}}} \tag{2.9}$$

$$k_{tone} = \frac{\frac{1}{\sigma^2_{tone}}}{\frac{1}{\sigma^2_{face}} + \frac{1}{\sigma^2_{tone}}} \tag{2.10}$$

The variance of each gender representation is equal to the sum of the variances of the constituent face and tone estimates, multiplied by the square of their weights:

$$\sigma^2 = k_{face}^2\sigma_{face}^2 + k_{tone}^2\sigma_{tone}^2 \tag{2.11}$$

As before, d-prime is calculated as

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2}}. \tag{2.12}$$

2.6 Acknowledgments

This work was supported by US National Institutes of Health grant R01 EY021755 to N.B.T.-B. and a National Defense Science and Engineering Graduate Fellowship to M.F.P. The authors thank Mariam Aly, Nick Hindy, Judy Fan, and Daniel Takahashi for helpful discussions.

2.7 Collaborators

The work described in this chapter was conducted in collaboration with Nick Turk-Browne. At the time of this writing, these results have been published on bioRxiv (Panichello and Turk-Browne, 2020)

2.8 Author contributions

M.F.P. and N.B.T.-B. designed experiments. M.F.P. collected and analyzed data. M.F.P. and N.B.T.-B. discussed the results and wrote the paper.

Chapter 3

Transformation of memories by expectation

3.1 Abstract

Working memory is critical to cognition, decoupling behavior from the immediate world. Yet, it is imperfect; internal noise introduces errors into memory representations. Such errors have been shown to accumulate over time and increase with the number of items simultaneously held in working memory. Here, we show that discrete attractor dynamics mitigate the impact of noise on working memory. These dynamics pull memories towards a few stable representations in mnemonic space, inducing a bias in memory representations but reducing the effect of random diffusion. Model-based and model-free analyses of human and monkey behavior show that discrete attractor dynamics account for the distribution, bias, and precision of working memory reports. Furthermore, attractor dynamics are adaptive. They increase in strength as noise increases with memory load and experiments in humans show these dynamics adapt to the statistics of the environment, such that memories drift towards contextually-predicted values. Together, our results suggest attractor dynamics mitigate errors in working memory by counteracting noise and integrating contextual information into memories.

3.2 Introduction

In this chapter we extend our focus from perception to working memory, our ability to maintain information without direct sensory input. It allows us to decouple behavior from the immediate world, serving as the substrate for planning and problem solving (Baddeley, 2003). Despite its fundamental role in cognition, information in working memory is not stored with perfect fidelity. Errors accrue over time (Pertzov et al., 2017; Rademaker et al., 2018; Shin et al., 2017; Zhang and Luck, 2009) and with the number of items simultaneously held in working memory (Adam et al., 2017; Bays et al., 2009; Fougnie et al., 2012; Luck and Vogel, 1997; van den Berg et al., 2012; Zhang and Luck, 2008).

Theoretical work suggests that the impact of noise can be mitigated if memories are stored using a finite set of stable states known as discrete attractors (Brody et al., 2003; Chaudhuri and Fiete, 2016; Kilpatrick et al., 2013; O'Reilly et al., 1999; Renart et al., 2003). In such systems, memory representations drift towards the attractor states. Once there, memories are stable and therefore resistant to diffusive noise. However, this comes at the the cost of discretizing continuous information, reducing precision and inducing bias into memory.

Here, we test whether the brain uses discrete attractor dynamics to mitigate the impact of noise on working memory. By fitting a flexible dynamical systems model to data from individual subjects, we estimate the forces governing the temporal evolution of working memory representations in both humans and monkeys. We show that discrete attractor dynamics better explain behavior than competing models of memory dynamics. Indeed, discrete attractor dynamics account for the distribution, bias, and precision of working memory reports and the accumulation of error in memory over time. Furthermore, these dynamics adapt to changes in context and memory load in a way that minimize errors in working memory.

3.3 Results

3.3.1 Systematic error in memory increases with load and time

To understand the dynamics governing working memory representations, we examined the behavior of humans (N=90) and monkeys (N=2) performing a delayed estimation task (Wilken and Ma, 2004, Fig. 3.1a). Subjects were instructed to remember the color of 1 to 3 simultaneously-presented stimuli located at different positions on the display (humans saw 1 or 3 items; monkeys saw 1 or 2). After a variable memory delay, subjects reported the remembered color at a cued target location using a

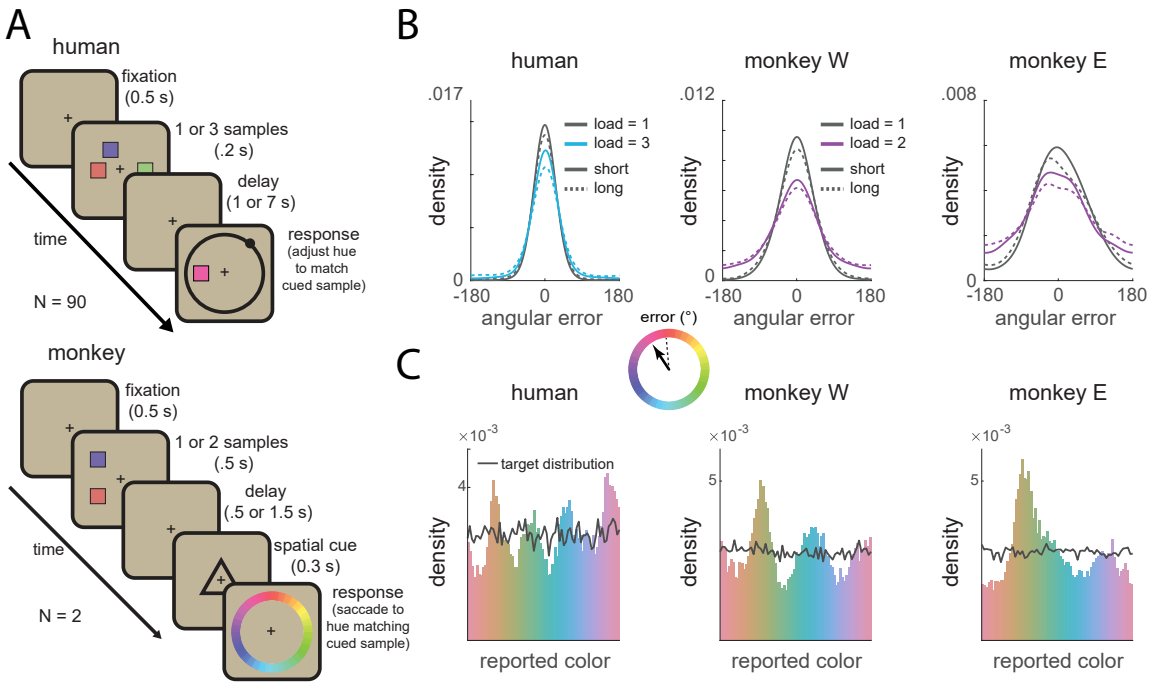


Figure 3.1: Memories cluster in a continuous working memory task. (A) Top: humans (N=90) performed a color delayed-estimation task in which they reported the color of a spatially-cued sample after a variable delay. Humans made their report by adjusting the hue of the response probe by rotating a response wheel (black circle) using a mouse. We rotated the mapping between wheel angle and color on each trial to avoid spatial encoding of color memories. Bottom: monkeys (N=2) performed a similar task. A symbolic cue indicated which sample to report (top or bottom). Monkeys reported a specific color value using an eye movement to a color wheel that was rotated on each trial. (B) Distribution of angular error for humans (top) and monkeys (bottom). Error increased with load and delay time. Gray lines = low load, blue lines = high load, solid lines = short delay, dashed lines = long delay. Inset: Error is calculated as the angular deviation between the color of the cued sample and the reported color in color space. (C) Non-uniform distribution of reported colors for humans (top) and monkeys (bottom). Gray line shows the distribution of target colors.

continuous scale. Stimulus colors were drawn uniformly from an isoluminant circular color space. We quantified error as the angular deviation between the target color and the subject's report. As expected (Adam et al., 2017; Bays et al., 2009; Fougnie et al., 2012; Luck and Vogel, 1997; Pertzov et al., 2017; Rademaker et al., 2018; Shin et al., 2017; van den Berg et al., 2012; Zhang and Luck, 2008, 2009), the average absolute error increased as a function of delay and working memory load in both humans and monkeys (Fig. 3.1b; humans (H): load, $F(1, 89) = 147.23$, $p < 1 \times 10^{-15}$; delay, $F(1, 89) = 85.44$, $p = 1.17 \times 10^{-14}$; load x delay, $F(1, 89) = 13.92$, $p = 3.36 \times 10^{-4}$, analysis of variance; monkey W (W): load, $p < 0.001$; delay, $p = 0.006$; load x delay, $p = 0.495$, bootstrap; monkey E (E): load, $p < 0.001$; delay, $p = 0.009$; load x delay, $p = 0.303$, bootstrap).

Despite the uniform distribution of target colors, the responses of both human and monkey subjects were significantly non-uniform (Bae et al., 2015; Bae et al., 2014; Hardman et al., 2017;

Pratte et al., 2017, Fig. 3.1c, humans and monkeys $p < 0.001$ against uniformity, Hodges-Ajne test; $p < 0.001$ against target distribution, permuted Kuiper's test). This was reflected in a significant decrease in the entropy of the response distribution relative to the target distribution (H: 2.54 vs. 2.61 bits, $t(89) = 13.90$, $p < 1 \times 10^{-15}$, t-test; W: 2.61 vs. 2.65 bits, $p < 0.001$, bootstrap; E: 2.58 vs. 2.65 bits, $p < 0.001$, bootstrap). Responses clustered around specific colors, seen as peaks in the response histogram (Fig. 3.1c). Clustering increased with delay time ($F(1, 89) = 9.56, p = 0.003$, analysis of variance) and with memory load in humans ($F(1, 89) = 5.45, p = 0.022$; Fig. S3.1-2), suggesting that clustering is the result of a load-dependent dynamic process that unfolds over the course of encoding and the memory delay.

3.3.2 Attractor dynamics influence memory representations

Motivated by these results, we tested the hypothesis that discrete attractor dynamics underlie the evolution of working memory representations. Attractor states can be conceptualized as local minima in an energy landscape over mnemonic (color) space, such that memories drift towards nearby attractors over time (Fig. 3.2a). These dynamics could provide a mechanistic explanation for the observed clustering of memory reports.

To test for the existence of discrete attractors, we developed a model to characterize the dynamics governing working memory representations. The model describes memory error as a combination of diffusion from noise in the neural representation (Burak and Fiete, 2012; Compte et al., 2000; Wimmer et al., 2014) and drift towards attractor states. Diffusion was quantified as a random walk from the current location in mnemonic space with no bias ($\mu = 0$) and a variance (σ_L^2) that depended on the number of colors presented (L = memory load). Discrete attractor dynamics were modeled by fitting a function $G(\theta)$ that describes how a remembered color θ will drift as a function of its current value (Fig. 3.2b). Positive drift values reflect a clockwise drift (to the right in Fig. 3.2b) while negative values reflect a counterclockwise drift (to the left). Thus, attractors are points in mnemonic space that 1) are fixed, such that they have no drift, and 2) pull nearby memories towards themselves, indicated by a negative slope in the drift function (Fig. 3.2b, dashed lines). Subjects displayed the same number and location of clusters in their distribution of memory reports regardless of load condition (Fig. S3.2), so we assumed that the pattern of drift did not vary with load (i.e. the shape of the function $G(\theta)$ was the same across loads). However, as with diffusion, the strength of the drift was allowed to vary across memory load (i.e. $G(\theta)$ is scaled by β_L).

Together, drift and diffusion define the temporal evolution of memories during the delay (Fig. 3.2c);

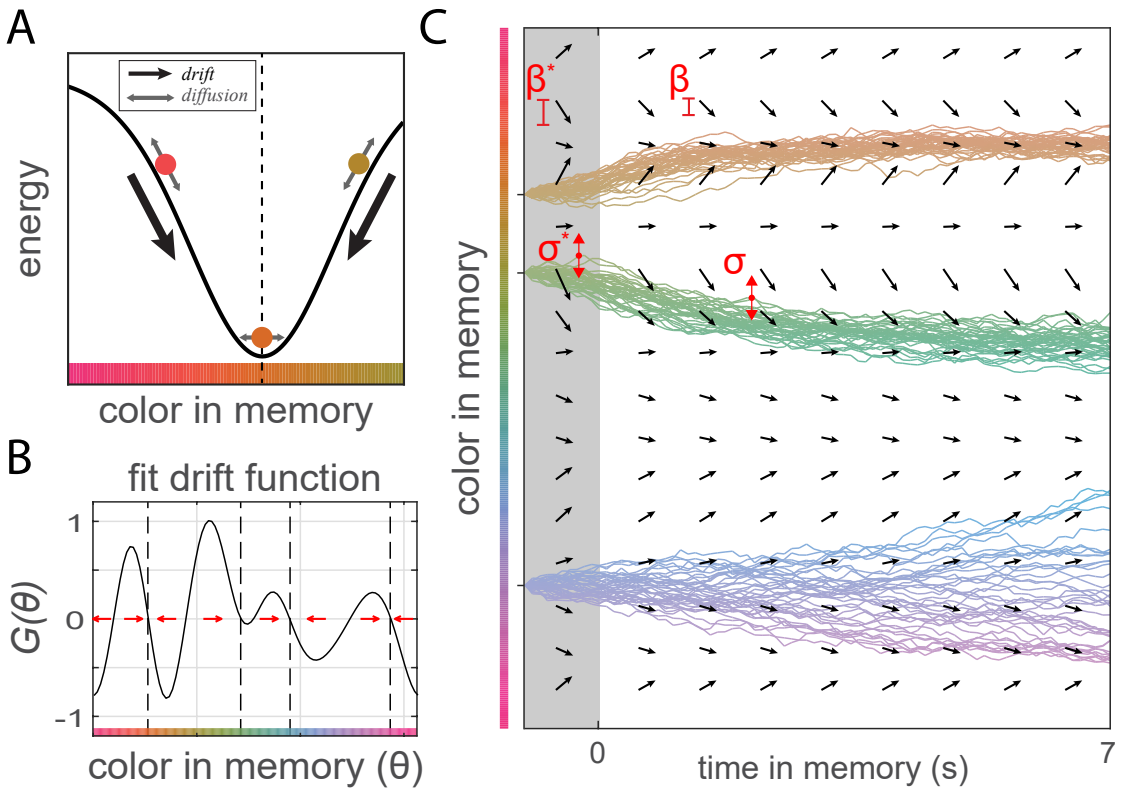


Figure 3.2: Structure of the dynamical model. (A) Illustration of the influence of attractors on color memory. Attractors (dashed line) cause memories to drift over time (black arrow), introducing bias in reports. Noise causes memories to randomly diffuse (grey arrows). (B) The drift function $G(\theta)$ describes how a memory will change based on its current state. Red arrows show the direction of drift; attractors have converging drift (dashed lines). We estimated $G(\theta)$ for each subject using a linear combination of von mises derivatives. (C) The simulated evolution of three color memories during a hypothetical trial. Memory evolves over time according to the drift function (vector field) and random noise. Each line indicates the temporal evolution of a remembered color under a different realization of the noise process. Terms described in main text.

dynamics evolve according to the differential equation $d\theta = \beta_L G(\theta)dt + \sigma_L \mathcal{N}(0, dt)$. Previous work has shown that reports of perceived colors are clustered, although clustering is greater for colors held in working memory (Bae et al., 2015). To capture clustering and other sources of error (Bays et al., 2011; Buschman et al., 2011) that emerge during encoding, inputs were first passed through an encoding stage governed by a similar drift and diffusion process with the same drift function $G(\theta)$. However, the strength of drift and diffusion during encoding was set independently by two additional parameters (β_L^* and σ_L^*; see Methods for details). This allowed us to test for discrete attractor dynamics during both encoding and the memory delay (Fig. 3.3). Finally, three additional terms in the model captured errors due to forgetting of memories (Zhang and Luck, 2008), responses to non-targets (Bays et al., 2009), and noise introduced at decoding (see Methods for de-

tails). Model parameters were estimated by maximizing the joint likelihood of the observed memory reports across individual trials for each subject. Critically, the model did not assume attractor dynamics (Fig. 3.3a); when β_L and β_L^* are zero, memories are only influenced by diffusion, forgetting, and responses to non-targets, as in previous models.

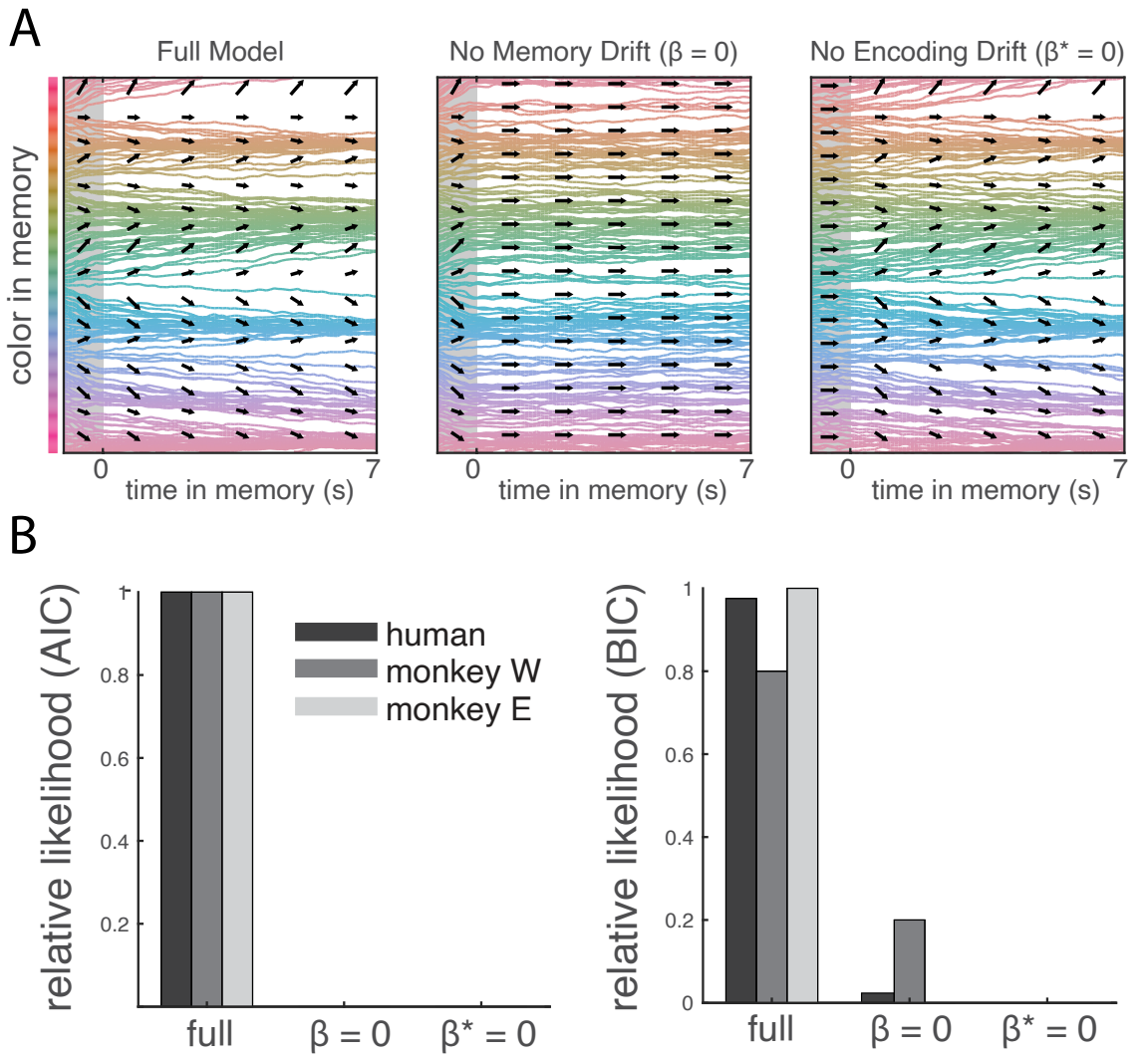


Figure 3.3: Behavior is best explained by attractor dynamics during encoding and memory. (A) Simulated memory trajectories from the best-fitting model for one subject. Left: The full dynamical model includes drift towards attractor states during both encoding and memory. Reduced models include drift only during encoding (middle) or memory (right). (B) AIC and BIC model weights (normalized relative likelihood) for the full model compared with models with zero drift during encoding ($\beta^* = 0$) and memory ($\beta = 0$). Values indicate the probability that the given model is the best model in the set Wagenmakers and Farrell, 2004.

Discrete attractor dynamics provide a better account of behavior than models in which memories only diffuse randomly (Fig. 3.3). To demonstrate this, we compared the full model with drift and diffusion to reduced models without drift towards attractor states during encoding or memory

($\beta^* = 0$ or $\beta = 0$, Fig. 3.3a). Three model comparison statistics (AIC, BIC, and cross-validated likelihood) all indicated that the full model performed best (Fig. 3.3b and Tables S3.1 and S3.2; H: relative likelihood of full model = 1.00 (AIC) and 0.98 (BIC); W: 1.00 and 0.80; E: 1.00 and 1.00). Thus, both the encoding and delay periods are characterized by drift of memories towards attractor states.

As seen in previous work (Bae et al., 2015; Bae et al., 2014), memory reports clustered at certain points in color space, and the bias and precision of reports vary systematically around points of peak clustering. Here, we show that the discrete attractor dynamics explain these variations. First, discrete attractor dynamics predict a clustered distribution of memory reports because memories tend to accumulate at attractor states. Accordingly, colors near attractor states identified by each subject's best-fit model were reported more frequently than average (Fig. 3.4a, H: $t(89) = 43.49$, $p = 9.54 \times 10^{-62}$; W: $p < 0.001$, bootstrap; E: $p < 0.001$, bootstrap). The distribution of memory reports predicted by each subject's best-fit model provides an excellent fit of the empirically-observed distribution of memory reports (Model: Fig. 3.4b, H: $r(70) = .909$, $p = 2.57 \times 10^{-28}$; W: $r(70) = .741$, $p = 9.93 \times 10^{-14}$; E: $r(70) = .934$, $p = 4.21 \times 10^{-33}$, Pearson's r).

Second, discrete attractors explain bias in working memory reports. Memories of a particular target color will consistently drift towards the closest attractor state, inducing systematic bias in subjects' reports. This is evident in subjects' behavior: memories for target colors counter-clockwise to an attractor location tended to drift clockwise, while targets clockwise to an attractor tended to drift counter-clockwise (Fig. 3.4C; H: mean slope -0.40 less than zero, $t(89) = -12.60$, $p = 1.73 \times 10^{-21}$, t-test; W: -0.59, $p < 0.001$, bootstrap; E: -0.73, $p < 0.001$, bootstrap). Model-free analyses showed similar effects. The peaks in the response histogram provide independent estimates of attractor locations. Aligning the bias around peaks in the response histogram reveals a similar pattern with a negative slope (Fig. S3.3; Bae et al., 2015; Bae et al., 2014) . Furthermore, the model provides a good qualitative fit to the pattern of bias across color space (Fig. 3.4D). The model's predicted pattern of biases for each target color was highly correlated with the empirically observed pattern of biases in both human and monkeys (H: $r(88) = .939$, $p = 1.41 \times 10^{-42}$; W: $r(58) = .864$, $p = 6.95 \times 10^{-19}$; E: $r(58) = .850$, $p = 8.13 \times 10^{-18}$, Pearson's r).

Third, discrete attractors explain the precision of working memory reports. Memories near attractors are more stable: as diffusive noise drives a memory representation away from an attractor, drift will pull it back towards the attractor, resulting in a narrow response distribution. For both humans and monkey subjects, the standard deviation (SD) of memory reports was lower for targets near attractors identified by each subject's best-fit model (Fig. 3.4E; H: ΔSD=-1.96, $t(89) = -4.90$,

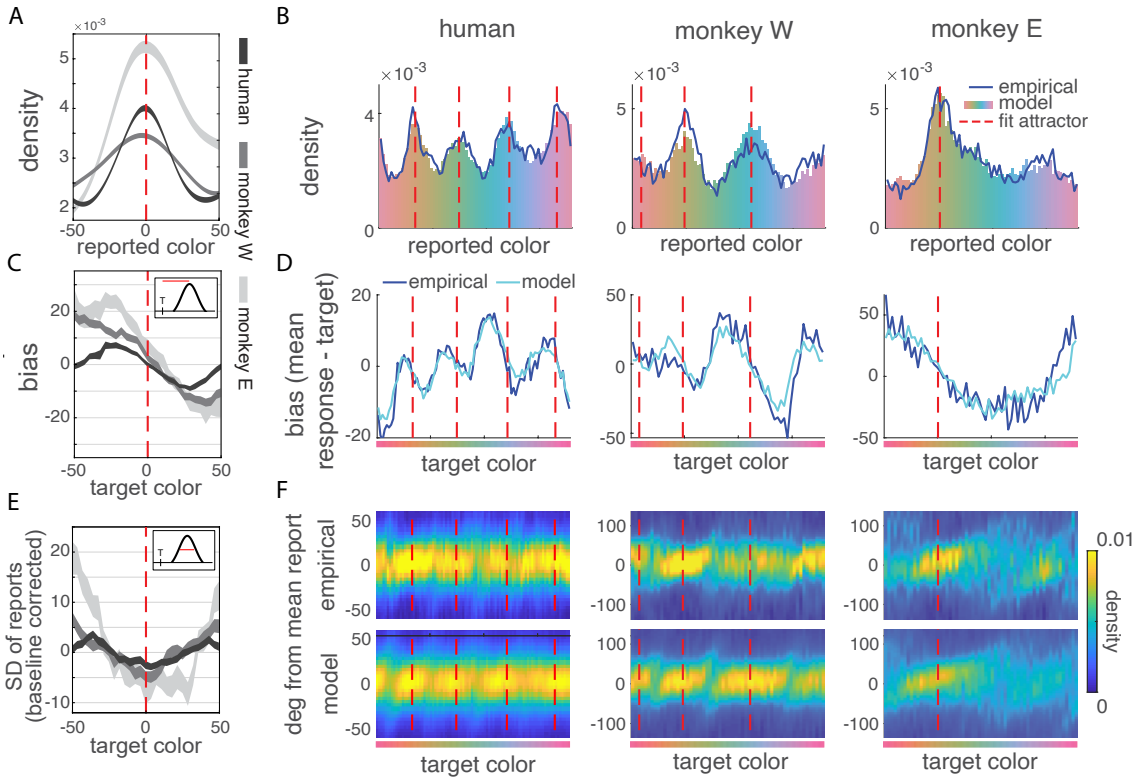


Figure 3.4: Attractor dynamics explain the distribution, bias, and precision of working memory reports. (A) Observed distribution of memory reports around fit attractors (red dashed line). X-axis: degrees in color space. (B) Distribution of simulated memory reports from the full model fit to human and monkey subjects. Blue line shows empirical distribution of reports. (C) Observed bias +/- SEM around fit attractors. Inset: bias is calculated as the angular distance between the target and mean report. Positive values indicate clockwise (CW) drift; negative values indicate counter-clockwise (CCW) drift. (D) Mean bias in reports as a function of target color (dark blue line), with predictions from best-fit models (light blue line). (E) Observed SD of reports +/- SEM around fit attractors. (F) Distribution of memory reports around their mean as a function of target color (top row), with predictions from best-fit models (bottom row). More precise memory reports are indicated by tighter distributions around their mean.

$p = 4.20 \times 10^{-6}$, t-test; W: -2.96, $p < 0.001$, bootstrap; E: -5.59, $p < 0.001$, bootstrap). Model-free analyses again showed similar effects: SD was significantly reduced at the peaks in the response histogram (Fig. S3.3. As with bias, discrete attractor dynamics predict the pattern of precision across color space (Fig. 3.4F). The model's predicted pattern of precision as a function of target color was correlated with the empirically observed values in both human and monkeys (Fig. 3.4F, H: $r(88) = .370$, $p = 3.27 \times 10^{-4}$; W: $r(58) = .377$, $p = 0.003$; E: $r(58) = .630$, $p = 6.88 \times 10^{-8}$, Pearson's r).

We can exclude several other possible explanations for the non-uniform distribution of memory reports. One alternative explanation is that clustering is driven by subjects guessing with a biased

distribution on a subset of trials. However, if true, then bias would not display an 'attractive' positive-to-negative transition at cluster peaks and precision would not depend on the identity of the item in memory (Fig. S3.4). A second alternative is that clustering could be driven by a non-linear mapping between the stimulus space chosen by the experimenter and the subject's true perceptual space. However, such a model predicts the opposite pattern of bias across color space (Fig. S3.5).

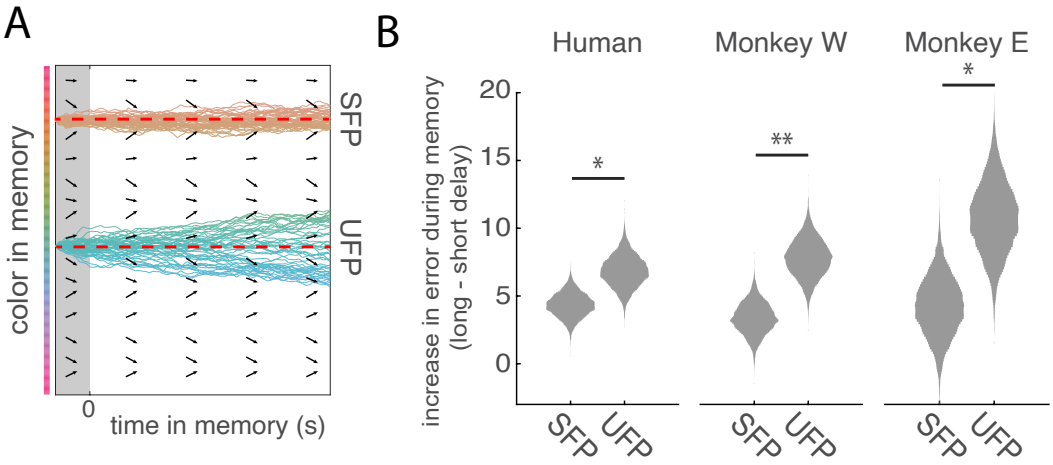


Figure 3.5: Memories near attractor states are more stable. (A) If discrete attractors underlie working memory, then memories of different target colors will accumulate error at different rates over time. Memories of target colors near stable fixed points (SFPs) accumulate a relatively small amount of error over time because perturbations away from SFPs due to random noise are corrected by drift back towards the SFP. Unstable fixed points (UFPs) lie in between attractors; perturbations away from UFPs due to random noise are exacerbated by drift away from the UFP. Red dashed lines the indicate location of two of the fixed points. (B) Mean increase in error at SFPs and UFPs identified by the fit model. Distributions reflect bootstrapped values. * $p < 0.05$, ** $p < 0.01$, bootstrap.

The discrete attractor model also predicts how errors in working memory evolve over time. First, the discrete attractor model accurately recapitulates the increase in error over the delay. To measure the change in mean error over the delay, we measured error for memory delays ranging from 1 to 7 seconds (Experiment 1b; Fig. S3.6a; 120 new human subjects). The discrete attractor model provided a good fit to the increase in error with memory delay (Fig. S3.6).

Second, the discrete attractor model makes the specific prediction that memories of different target colors are expected to accumulate error at different rates. Attractors are 'stable fixed points' because they counteract perturbations of memory due to random noise. Perturbations are corrected by drift back towards the stable fixed point. Because this process occurs continuously over time, memories of target colors near stable fixed points are not only more precise overall (i.e., as in Fig. 3.4E), but also accumulate error at a relatively slow rate over time (Fig. 3.5a). In contrast, target

colors near 'unstable fixed points' accumulate error relatively quickly over time because random perturbations away from these points are exacerbated by drift away from the unstable fixed point (Fig. 3.5a). To test this prediction, we first identified stable and unstable fixed points for each subject by identifying target colors with attractive bias (zero with a negative slope) or repulsive bias (zero with a positive slope). We then calculated how much error increased on long delay trials relative to short delay trials for target colors near stable and unstable fixed points. For both humans ($p = 0.036$, bootstrap) and monkeys (W: $p < 0.0001$, E: $p = 0.024$, bootstrap), error increased more over time for target colors near putative unstable fixed points (Fig. 3.5b).

3.3.3 Attractor dynamics strengthen with load

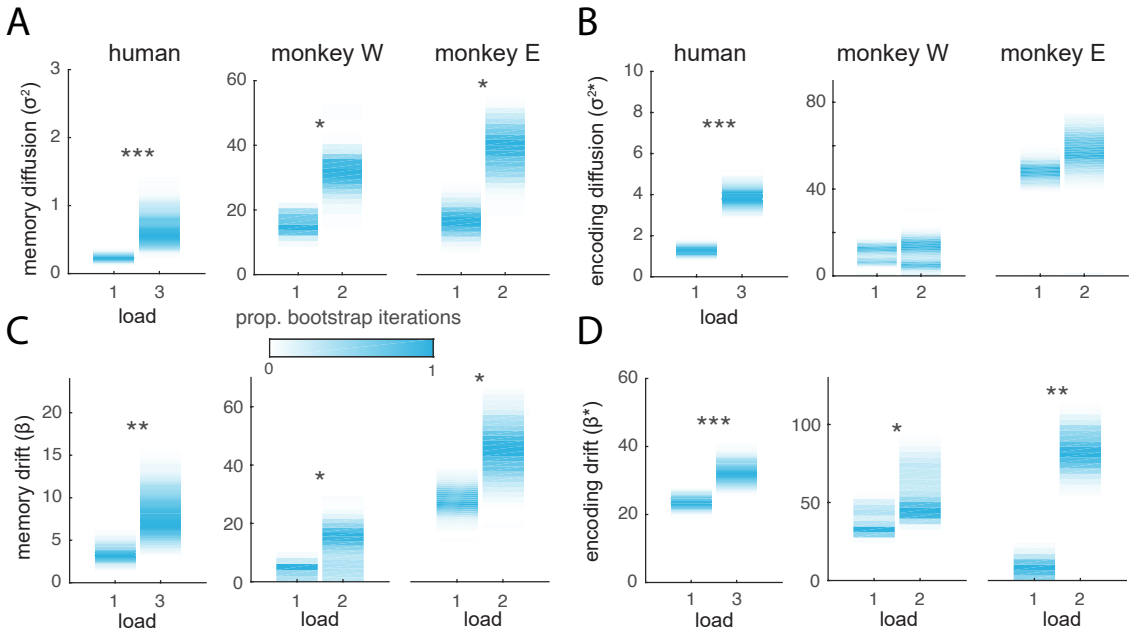


Figure 3.6: Drift and diffusion increase with memory load. Experiment 1a maximum likelihood parameter fits for the diffusion (A-B) and drift (C-D) scaling parameters during memory and encoding. Color intensity reflects normalized proportion of bootstrap iterations. As detailed in the Methods, all parameters are rates (change per second); dynamics during encoding evolve over a fixed period of time (simulated as 1 second), while memory dynamics evolve over the memory delay, which varied from trial to trial. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, bootstrap.

The error-correcting properties of attractors may be especially critical when memory load is high. High memory load decreases the magnitude of neural responses (Buschman et al., 2011), which is thought to render memories more susceptible to noise and, therefore, increase diffusion (Bays, 2015; Burak and Fiete, 2012). Indeed, as estimated by the model fits to experiment 1a, diffusion during the memory delay increased with memory load (Fig. 3.6a, σ_L^2, H: $p = 0.001$; W:

$p = 0.021$, E: $p = 0.010$, bootstrap) although changes during encoding were mixed (Fig. 3.6b, σ_L^{2*}, H: $p < 0.001$; W: $p = 0.459$, E: $p = 0.100$, bootstrap). Consistent with the theory that attractor dynamics compensate for diffusion, we saw a commensurate increase in drift during the memory delay (Fig. 3.6c, β_L, H: $p = 0.002$; W: $p = 0.026$, E: $p = 0.026$, bootstrap) and during encoding (Fig. 3.6d,β_L^*, H: $p = 0.001$; W: $p = 0.024$, E: $p = 0.009$, bootstrap). Similarly, two model-free measures of drift, clustering of responses and mean absolute bias, increased with load (Fig. S3.1). Note that although the rate of drift and diffusion during memory is less than that during encoding, their effects accumulate over the course of the memory delay.

3.3.4 Attractor dynamics are shaped by experience

While discrete attractors compensate for diffusion, they also induce systematic error into working memory. Thus, there is a trade-off between the finite error caused by drifting toward an attractor and the ever-increasing error associated with diffusion. To test whether discrete attractors improved overall performance, we simulated memory dynamics for the full discrete attractor model ('drift + diffusion') and from the same model with encoding and memory drift set to zero ('diffusion', $\beta = \beta^* = 0$). Thus, we can ask how memory accuracy would change if diffusion were held constant and we manipulated only the presence or absence of discrete attractor states. As shown in Fig. 3.7a, the two models accumulate error at different rates over time. Initially, the mean absolute error is greater in the drift + diffusion model due to memory corruption by drift towards attractor states during encoding and the early delay period ($p < 0.05$ for $t < 11s$, bootstrap). However, discrete attractors also counteract diffusive noise and so, as the delay increases, the drift + diffusion model performs significantly better than the diffusion model ($p < 0.05$ for $t >= 33s$, bootstrap), with the crossover in performance occurring at $t \sim 17s$. Thus, attractor dynamics have a greater impact the longer information is held in working memory.

Discrete attractor dynamics are most beneficial when they adapt to the current context. For example, the statistics of many visual features in the real world are not uniform across perceptual space (including color Yendrikhovskij, 2001). In this case, errors can be reduced if attractor states reflect the statistics of the environment, such that attractors occur at the location of common stimuli. To demonstrate this, we tested the performance of the full discrete attractor model in different environments. Environments varied in the proportion of target colors drawn from within 10 degrees of an attractor. For example, when 50% of targets were drawn from nearby an attractor, the 'drift + diffusion' model significantly reduced working memory error for all t (Fig. 3.7a, red

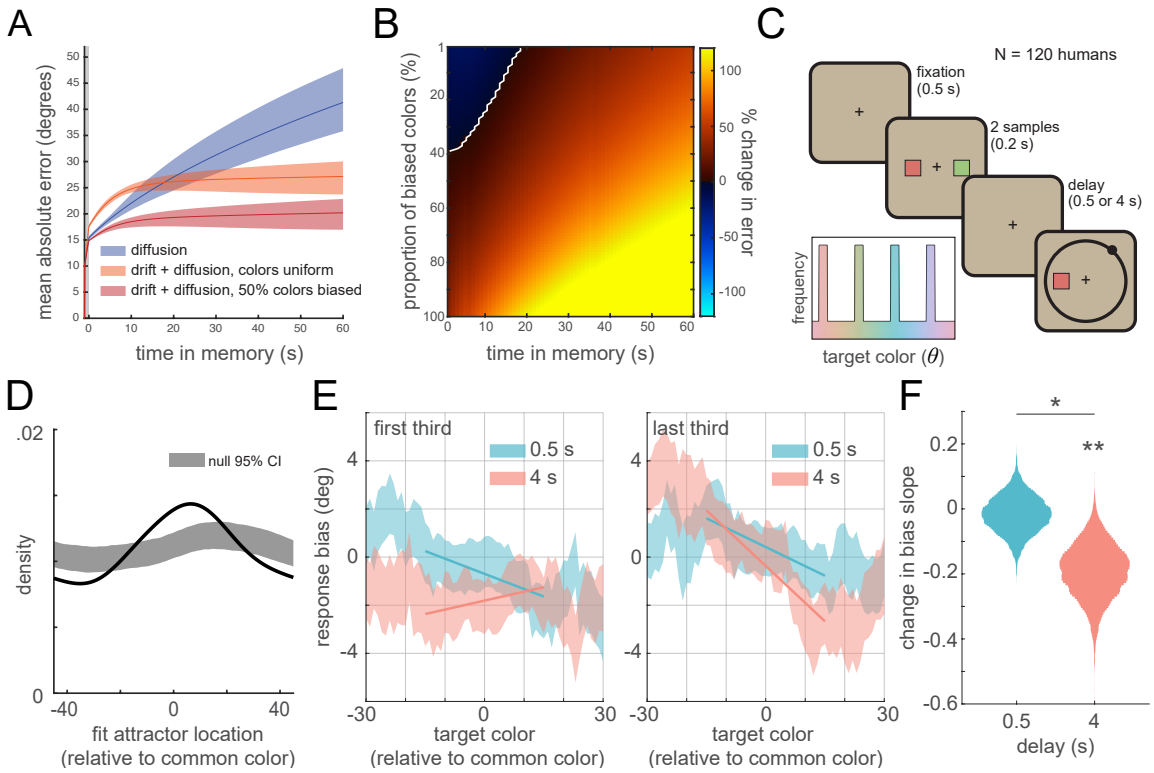


Figure 3.7: Attractor states reflect environmental statistics. (A) Simulated mean absolute error in the representation of the target color over time. The best-fitting discrete attractor model in Experiment 1a ('drift + diffusion'; orange and red) is compared to the same model without encoding or memory drift ('diffusion'; blue). Error is simulated for two target distributions: one in which targets were sampled uniformly (colors uniform) and one in which 50% of targets were drawn more frequently from colors near attractors. Note that the performance of the diffusion model does not depend on the distribution of target colors. Results are based on the mean parameter fits for the high load condition; similar results were observed for low load. Shaded regions reflect variability across subjects (95% CI). (B) Percent increase in simulated error for diffusion model compared to full 'drift + diffusion' model as a function of time in memory (x-axis) and the degree of bias in the target distribution (y-axis). Warm colors indicate attractors reduce memory error; white line indicates no change. (C) Experiment 2. Humans (N=120) performed a color delayed estimation task with two stimuli. Inset: Example color distribution for one subject. Four groups of colors were presented more frequently. Common colors were equally spaced in color space and differed for each subject. (D) Probability of attractor position (estimated from dynamical model fits) relative to common colors (black). Shaded region indicates 95% confidence intervals of the null distribution, based on randomly permuting the location of common colors across subjects. (E) Bias for targets around common colors during the first and last third of trials, with regression lines. Error bars reflect standard error of the mean. (F) Difference in the slope of bias at common colors between the first third and last third of trials for short and long-delay trials. Distributions reflect bootstrapped values. * $p < 0.05$, ** $p < 0.01$, bootstrap.

trace). Parametrically varying the proportion of biased colors revealed that discrete attractor states tuned to the statistics of the environment reduced memory error, even with modest biases in the color distribution (Fig. 3.7b). These results suggest that, in order to minimize working memory errors, attractor dynamics should adapt to the statistics of the current environment.

To test whether memory dynamics adapt to context, we collected data from 120 additional human subjects in a continuous working memory task with a biased stimulus distribution (Experiment 2, Fig. 3.7c). During this task, the statistics of the environment were such that half of all stimuli were drawn from one of four common colors (randomly chosen for each subject) while the other half were drawn from a uniform distribution.

Both model-free and model-based analyses suggest that participants developed attractor states at the common color locations. First, attractor states, as identified by fitting the dynamical model, were significantly more likely to occur at the location of common colors than expected by chance (Fig. 3.7d, $p < 0.001$, randomization test, model fits were limited to trials in which the target color was drawn from a uniform distribution). Second, consistent with the accumulation of memories at attractor states, subjects were significantly more likely than chance to report common colors, even on the half of trials when the target was drawn from a uniform distribution (Fig. S3.7a, $p < 0.001$, randomization test). Third, over the course of the experiment, the pattern of bias around common colors became more consistent with attractor states. As shown in Fig. 3.4c, attractors pull in nearby memories, resulting in a positive-to-negative transition in bias. The more negative the slope, the stronger the attractor. Attraction towards common colors increased with experience: the slope of bias around common colors was significantly more negative during the last third of trials than during the first third (Fig. 3.7e, $p = 0.0138$, bootstrap).

To determine if changes in bias were driven by differences in encoding or memory dynamics, we analyzed short memory delay and long memory delay trials separately. If learned biases toward common colors manifest during encoding, then the bias slope should become more negative for both short and long trials. In contrast, if biases manifest during memory, then the change in bias should be especially strong for long delay trials because the biases in memory dynamics have more time to accumulate. Non-parametric regression revealed a main effect of delay length on bias slope ($p = 0.026$) modulated by a delay x epoch (first or last third of trials) interaction ($p = 0.039$). The bias slope around common colors on short delay trials did not differ between the first third and last third of trials (Fig. 3.7f, $p = .384$, bootstrap) but became significantly more negative for long-delay trials ($p = .006$, bootstrap). Directly comparing the two delay conditions, bias slope was more negative for long-delay trials than short delay trials in the last third of trials ($p = .0411$, bootstrap).

These results suggest that learning modified dynamics during memory rather than encoding.

To ensure that these results were not due to subjects strategically reporting common colors based on explicit knowledge of the stimulus distribution, we analyzed debriefing data collected from the participants in Experiment 2 and 1b. Subjects were not better than chance at identifying whether they were exposed to a biased or uniform stimulus distribution (see Methods for details). Furthermore, participants in Experiment 2 displayed the same pattern of results regardless of whether or not they correctly reported that the stimulus distribution was biased during debriefing (Fig. S3.8).

Finally, if attractors emerge at common color locations, then this should alter the distribution of reported colors over the course of the experiment. Indeed, we found the clustering of memory reports across subjects decreased from the first third to the last third of trials (2.62 to 2.63 bits, $p < 0.001$, randomization test; Fig. S3.7b. This is consistent with a strengthening of attractors at the contextually-predicted locations, which were uncorrelated across subjects. However, it is important to note that, although weaker, clustering is still partially evident at baseline locations in the last third of trials (Fig. S3.7b), and the slope of bias around these baseline locations did not change in strength between the first and last third of the experiment (p = 0.5701, bootstrap). This suggests that the learning rate governing changes in the dynamics is low, ideal for extracting statistical regularities (McClelland et al., 1995).

3.4 Discussion

Our results highlight the dynamic nature of working memory representations. Using both model-based and model-free analyses, we show that two forces drive the evolution of visual representations during encoding and maintenance: 1) random diffusion and 2) drift towards discrete attractor states. Together, these forces provide a parsimonious explanation of the distribution, bias, and precision of memory reports and the accumulation of error in memory over time. These results build on previous models that do not explain why errors in working memory differ as a function of the content (e.g., Compte et al., 2000; Fougnie et al., 2012; van den Berg et al., 2012) or how memory representations dynamically evolve (e.g., Bae et al., 2015).

Previous psychophysical, theoretical, and neurophysiological work has shown noise in neural activity can cause memories to diffuse away from their original representation, leading to errors in working memory (Burak and Fiete, 2012; Compte et al., 2000; Schneegans and Bays, 2018; Wimmer et al., 2014). Our results suggest attractor dynamics within mnemonic space can counteract this noise by pulling memories towards a few stable representations. Consistent with previous theoretical work

(Brody et al., 2003; Chaudhuri and Fiete, 2016; Kilpatrick et al., 2013; O'Reilly et al., 1999; Renart et al., 2003), we provide experimental evidence that the stability of representations at attractors limits the effect of random diffusion. Furthermore, the fact that discrete attractors are evolutionarily conserved across monkeys and humans emphasizes the benefits of error-correction. Indeed, this may be a general phenomenon in the brain: attractor dynamics are thought to minimize the impact of noise in long-term, associative memory (Hopfield, 1982, 1984 and in decision making Inagaki et al., 2017; Piet et al., 2017).

From an information-theoretic perspective, working memory can be conceptualized as a band-limited information channel (Koyluoglu et al., 2017). In this framework, discrete attractors compress working memory representations by discretizing the continuous mnemonic (color) space. Discretization reduces the information needed to encode a memory, allowing it to be more accurately stored in a noisy, band-limited system (Koyluoglu et al., 2017; Nassar et al., 2018). This is particularly important when storing multiple items in working memory. Increasing the number of items in working memory leads to interference between items, reducing memory accuracy (Almeida et al., 2015; Buschman et al., 2011; Pertzov et al., 2017). Consistent with this, we observed an increase in diffusive noise as more items are held in working memory. However, drift also increased in strength, compensating for the increase in noise. In other words, strengthening discrete attractor dynamics increases compression of memories; this reduces the fidelity of memories as they are further discretized, but also makes them more robust to noise and interference. Note that this increase in attractor strength with load cannot be explained by interference among items because item identity is random and so any such interactions would lead to random, not systematic, biases in memory. Several neural mechanisms might account for the increase in attractor strength with load, including increased drive into the network (Cohen et al., 1990; Wang et al., 2018) or changes in f-I gain via neuromodulation (Musslick et al., 2018; Servan-Schreiber et al., 1990).

Finally, our results suggest attractor dynamics adapt to context: attractors emerged at the position of commonly occurring stimuli. The relatively slow rate of change in dynamics (over hundreds of trials) is consistent with theoretical work that suggests such learning could be driven by synaptic plasticity (Kilpatrick, 2018). Indeed, such a mechanism with a slow learning rate is ideal for extracting the statistical regularities of the environment. Intriguingly, we found encoding dynamics adapted to changes in the environment more slowly than memory dynamics. This raises the possibility that encoding and memory dynamics may rely on different neural mechanisms.

By moving to reflect the statistics of the environment, attractors will pull memories towards likely stimuli. In this way, attractor dynamics act to integrate prior beliefs with noisy stimulus

information. This process is analogous to Bayesian inference applied over time. At each timestep in memory, drift applies the prior (embedded in the attractors) to each item in memory, which reflects the posterior of the previous timestep plus random noise. Thus, as time in working memory increases (and stimulus information diffuses), memory representations drift towards prior expectations. Such a process could constitute the mechanism by which sensory history influences working memory (Akrami et al., 2018; Papadimitriou et al., 2015; Papadimitriou et al., 2017). Beyond working memory, attractor dynamics could be a neurally-plausible mechanism for integrating prior beliefs with sensory information in other cognitive behaviors, such as decision making and perception.

3.5 Methods

3.5.1 Subjects

Thirty-three human subjects participated in Experiment 1a at Princeton University. Seventy-three additional subjects participated in an online version of Experiment 1a via Amazon Mechanical Turk (https://www.mturk.com). One-hundred twenty-five subjects participated in Experiment 1b via Amazon Mechanical Turk. One-hundred fifty-five subjects participated in Experiment 2 via Amazon Mechanical Turk. We screened subjects for a minimum of engagement in the task by estimating their probability of random guessing in the task using 3-component mixture model (Bays et al., 2009). Subjects with an estimated guess rate greater than 20% across all trials were excluded from further analysis, yielding thirty laboratory subjects and sixty online subjects for Experiment 1a, one-hundred twenty online subjects for Experiment 1b, and one-hundred twenty online subjects for Experiment 2. This threshold of 20% was set independently based on analysis of a separate pilot cohort of online subjects (N = 57). Subjects recruited online via Mechanical Turk have previously been used to study working memory and have performance comparable to lab subjects (Brady and Alvarez, 2011, 2015). We observe similar qualitative behavior between online and lab subjects (Fig. S3.9) and report their behavior together in the main text. All subjects attested that they had normal or corrected-to-normal vision. We confirmed that subjects had normal color vision using the Ishihara Color Blindness Test. Subjects provided informed consent in accordance with the Princeton University Institutional Review Board.

Two adult male rhesus macaques (8.9 and 12.1 kg) performed the Experiment 1a in accordance with the policies and procedures of the Princeton University Institutional Animal Care and Use Committee.

3.5.2 Experiment 1a - humans

For the laboratory version of Experiment 1a we presented stimuli on a CRT monitor positioned at a viewing distance of 60 cm. We calibrated the monitor using an X-Rite i1Display Pro colorimeter to ensure accurate color rendering. During the experiment, participants were asked to remember the color and spatial location of either 1 or 3 square sample stimuli. The color of each sample was drawn from 360 evenly spaced points along an isoluminant circle in CIE L*a*b* color space. This circle was centered at (L = 60, a = 22, b = 14) and the radius was 52 units. Colors were drawn pseudorandomly, with the caveat that colors presented on the same trial had to be at least 22° apart in color space. The samples measured 2° of visual angle (DVA) on each side. Each sample could appear at one of eight possible spatial locations. All possible locations had an eccentricity of 4.5 DVA and were positioned at equally spaced angles relative to central fixation (0, 45, 90, 135, and 180° clockwise and counterclockwise relative to the vertical meridian). The dimensions of the stimuli for the online experiment were defined by pixels rather than degrees of of visual angle. The samples had an edge length of 30 pixels and were presented at an eccentricity of 170 pixels.

Participants initiated each trial by clicking the mouse and by fixating a cross at the center of the screen (Fig. 3.1a). After 500 ms of fixation, one or three samples (the load) appeared on the screen. The samples were displayed for 200 ms and then were removed from the screen. Participants then experienced a memory delay of 1 second or 7 seconds, after which a response screen appeared. The response screen consisted of the outline of a square at one of the previous sample locations (the probe sample) and a response interface consisting of a circle on a ring. Participants used the mouse to drag the circle around the ring, which changed the color of the probe sample. The angular position of the circle on the ring corresponded to a particular angle in color space. The mapping between circle position and color space was randomly rotated on each trial to exclude the use of spatial memory. We instructed participants to adjust the color of the probe sample to match the color of the sample that had previously appeared at that location as closely as possible. We told participants that accuracy was more important than speed but that they should respond within a few seconds. There was no time limit on the response. All human participants completed 200 trials.

We monitored the eye position of the lab participants using an Eyelink 1000 Plus eyetracking system (SR Research). Participants had to maintain their gaze within a 2° circle around the central cross during initial fixation and sample presentation, or else the trial was aborted and excluded from analysis.

3.5.3 Experiment 1a - monkeys

We presented stimuli on a Dell U2413 LCD monitor optimized for color rendering. The monitor was positioned at a viewing distance of 58 cm. We calibrated the monitor using an X-Rite i1Display Pro colorimeter to ensure accurate color rendering. Sample colors were drawn from 64 evenly spaced points along an isoluminant circle in CIE L*a*b* color space. This circle was centered at (L = 60, a = 6, b = 14) and the radius was 57 units. Slightly different color wheels were used for the humans and the monkeys to accommodate the gamut of the different monitors used in each experiment. Nevertheless, colors corresponding to the same angle in each color wheel are extremely similar in appearance. The edges of the samples measured 2° of visual angle. Each sample could appear at one of two possible spatial locations: at 5 DVA eccentricity from fixation and 45° clockwise and counterclockwise from the horizontal meridian.

We adapted Experiment 1a so that it could be performed by non-human primates. The animals initiated each trial by fixating a cross at the center of the screen. After 500 ms of fixation, one or two samples appeared on the screen. The samples were displayed for 500 ms, followed by a memory delay of 500 ms or 1500 ms. Next, a symbolic cue was presented at fixation for 300 ms. This cue indicated which sample (top or bottom) the animal should report in order to get juice reward. The response screen consisted of a ring 2° thick with an outer radius of 5°. The animals made their response by breaking fixation and saccading to the section of the color wheel corresponding to their report. This ring was randomly rotated on each trial to prevent motor planning or spatial encoding of memories. The animals received a graded juice reward that depended on the accuracy of their response. The number of drops of juice awarded for a response was determined according a circular normal (von mises) distribution centered at 0° error with a standard deviation of 22°. This distribution was scaled to have a peak amplitude of 12, and non-integer values were rounded up. When response error was greater than 60°, no juice was awarded and the animal experienced a short time-out of 1 to 2 seconds. Responses had to be made within 8 seconds; in practice, this restriction was unnecessary as response times were on the order of 200-300 ms. We analyzed all completed trials (trials on which the animal successfully maintained fixation and saccaded to the color wheel, regardless accuracy). Monkey W completed 15,787 trials over 26 sessions and Monkey E completed 16,601 trials over 17 sessions.

We monitored the eye position of the animals using an Eyelink 1000 Plus eyetracking system (SR Research). The animals had to maintain their gaze within a 2° circle around the central cross during the entire trial until the response, or else the trial was aborted and the animal received a

brief timeout. Trials during which the animal broke fixation were excluded from analysis.

3.5.4 Experiment 1b

The stimuli and procedures were similar to those for the online version of Experiment 1a, except that participants were presented with two samples on every trial and the delay varied continuously between 1 seconds and 7 seconds. Model predictions (Fig. S3.6) were generated from the best fitting model. As in Experiment 1a, the full model provided the best fit to the data (mean increase in cross-validated log-likelihood over worst-fitting model, full: 7.45, $\beta = 0$: 7.41, $\beta^* = 0$: 0.20).

3.5.5 Experiment 2

The stimuli and procedures for Experiment 2 (Fig. 3.7c) were similar to those for the online version of Experiment 1a. We shortened the memory delays to 500 ms and 4000 ms to reduce the length of the experiment. Participants saw 2 samples on each trial. Critically, the color of the samples were no longer always drawn uniformly from the circular color space. Rather, for each sample, there was a 50% chance that the color of that sample would be drawn from a biased distribution (Fig. 3.7c). This biased distribution consisted of four equally-spaced clusters of common colors. Each cluster was 20° in width. Each participant was exposed to a unique set of common colors as the cluster means were shifted by a single random phase for each subject.

3.5.6 Subject debriefing

Participants in Experiments 1b and 2 were presented with following debriefing question: "During this experiment, some participants are shown target colors at random. Others are shown some colors more often than others. Which group do you think you are in?". The response options were "I was shown all colors about equally often" or "I was shown some colors more often than others". When presented with this two-alternative forced choice at the end of the experiment, 49.2% of participants in Experiment 2 correctly reported that the distribution of targets was biased, while 48.3% incorrectly reported a uniform distribution of targets (3 participants abstained). We estimated the false alarm rate for this question by analyzing responses of participants in Experiment 1b to the same question: 49.2% incorrectly reported a biased distribution, 50.0% reported a uniform distribution, 1 abstained. The proportion of subjects reporting a biased distribution was not significantly different between Experiments 1b and 2 ($\chi^2(1) = 0.015$, $p = 0.902$, chi-squared).

3.5.7 Effects of load and time on mean error

Throughout the text, all t-tests are two-tailed and all randomization tests are one-tailed, unless otherwise indicated.

We analyzed mean absolute error for human subjects using a 2x2 repeated measures ANOVA with factors load, delay time, and their interaction. We analyzed each monkey's data by fitting the equivalent regression model to their mean error in each condition. We obtained bootstrapped confidence intervals for each regression coefficient by re-sampling trials with replacement from each monkey's dataset and refitting the regression model on each iteration (1000 iterations). We also used this method to analyze the effect of load and time on clustering and mean bias (Fig. S3.1), and the effect of task epoch and time on bias slope (Fig. 3.7e-f).

3.5.8 Clustering metric

We observed that the distribution of reported hues $\hat{\theta}$ are clustered relative to the distribution of target hues Θ. To quantify this phenomenon, we developed a simple clustering metric. This metric relies on the fact that entropy is maximized for uniform probability distributions. In contrast, probability distributions with prominent peaks will have lower entropy. Because the target hues are drawn from a circular uniform distribution, the entropy of the targets $H(\Theta)$ will be relatively high. If a subject's responses are clustered, their entropy $H(\hat{\theta})$ will be relatively low. Taking the difference of these two values yields a clustering metric C. Negative values of C suggest greater clustering:

$$C = H(\hat{\theta}) - H(\Theta) \tag{3.1}$$

where:

$$H(x) = -\sum_{x=1}^{360} f(x) \log_2 f(x) d\hat{x} \tag{3.2}$$

To account for the fact that this estimate of entropy is biased, we subsampled the data such that there was an equal number of trials in each condition. We estimated the pdf of the responses $f(\hat{\theta})$ and the targets $f(\Theta)$ using kernel density estimation (Matlab CircStat toolbox, kernel width = 10°). Note that our goal was to quantify the clustering of reports for items in memory; random guesses (Bays et al., 2009; Zhang and Luck, 2008) confound this analysis by contributing a uniform component to the response distribution that varies systematically as a function of load and time. To address this, we estimated the proportion of responses due to guessing using mixture models (Bays et

al., 2009; Zhang and Luck, 2008) and removed a uniform component from the response distribution $f(\hat{\theta})$ and the target distribution $f(\theta)$ equal in area to the guess rate and then renormalized each.

3.5.9 Bias and standard deviation of memory reports

To dissociate systematic and unsystematic sources of error in memory, we calculated the bias and standard deviation of memory reports across color space. We used 4° bins for humans and 6° bins for monkeys to accommodate their coarser sampling of color space (64 target colors). Bias refers to the distance between the the target color and the mean reported color. We calculated the slope of bias around negative-slope zero-crossings in each subject's fit drift function (Experiment 1a), around significant peaks in each subject's response histograms (Experiment 1a), and around commonly presented presented colors (Experiment 2) by fitting a line to the bias +/- 15° around the point of interest. Mean standard deviation around these points was calculated around these points using the same window (+/- 15°). For monkey subjects, we boostrapped confidence intervals for slope and standard deviation by resampling trials with replacement.

To compute the bias and SD for the non-uniform guessing strategy (Fig. S3.4), we performed 1000 iterations of a randomization test where memory reports were shuffled with respect to the target colors and report the mean bias and SD for each target color across iterations.

To identify significant peaks in subjects' response histograms (Experiment 1a), we first estimated the PDF of subjects' responses using kernel density estimation. We identified possible peaks as samples larger than their two neighboring samples and recorded their amplitude. We then repeated this analysis on the distribution of targets, resampling with replacement to create a null distribution of peak amplitudes. Peaks in the original response distribution with an amplitude greater than the 95th percentile relative to the null were deemed significant. We identified negative-slope zero-crossings in the fit drift function of each subject by identifying peaks in the numerical integral of the drift function. Peaks with a prominence in the 20th percentile or lower across subjects were excluded from analysis.

Finally, to generate model predictions for bias and standard deviation, we fit the discrete attractor model to each subject's data and generated synthetic datasets (1,000 trials for each human subject and 20,000 trials for each monkey) by simulating responses from each subject's best fit model. We then analyzed the bias and standard deviation of these simulated reports as above. Model performance was assessed by correlating model predictions with empirical results across target colors.

3.5.10 Dynamical model

We developed a quantitative model to describe how items in memory change over time. We assume that two distinct influences may make memory dynamic. First, systematic biases may cause memories to drift towards stable attractor states over time. Second, memories may be perturbed by unsystematic random noise. We model memory using a stochastic ordinary differential equation that captures both of these influences:

$$d\theta = \beta_L G(\theta)dt + \sigma_L dW \tag{3.3}$$

This equation describes the time evolution of a color memory θ (a circular variable corresponding to an angle in our circular color space) under the influence of some deterministic dynamics defined by G (the drift) as well as an additive white noise process W with variance σ^2. β_L sets the gain of the drift. Thus, $\beta_L G(\theta)dt$ describes influence of drift and $\sigma_L dW$ the influence of random noise on memory. To test the hypothesis that memory load influences these dynamics we fit a separate β and σ for each load n.

Based on the clustering we observe in the data, it seems likely that $G(\theta)$ is a nonlinear function. We needed a relatively parsimonious way of describing $G(\theta)$ that still gave us enough flexibility to describe this nonlinearity. So, for each subject, we defined $G(\theta)$ using a basis set consisting of twelve first derivatives of the von mises distribution separated by 1 standard deviation on the interval $(0, 2\pi)$:

$$G(\theta) = \sum_{j=1}^{12} w_j \frac{d}{d\theta}\phi\left(\frac{2\pi}{12}j, \frac{2\pi}{12}\right) \tag{3.4}$$

where ϕ is a von mises distribution parameterized by a mean and standard deviation. We then divided $G(\theta)$ by its maximum absolute value. This normalization procedure aids the interpretation of β: it is the maximum instantaneous drift rate. Our choice of 12 basis functions was to minimize AIC in comparison to function estimates with higher or lower number of basis functions.

To fit the model described in equation 3.3 to subject data, we needed to describe the time evolution of θ probabilistically. So, we rewrote equation 3.3 as a Fokker-Planck equation, a partial differential equation that tracks probability density function of θ over time:

$$\frac{\partial}{\partial t}p(\theta, t) = -\frac{\partial}{\partial \theta}\beta_L G(\theta)p(\theta, t) + \frac{\sigma_L^2}{2}\frac{\partial^2}{\partial \theta^2}p(\theta, t) \tag{3.5}$$

In order to track probability mass, we discretized our 1-dimensional state space (the value of θ)

into 100 evenly spaced bins from 1° to 360°. Once discretized, the change in $p(\theta, t)$ over a given timestep dt can be described by a Markov transition matrix M_L:

$$\frac{\partial}{\partial t} p(\theta, t) = M_L p(\theta, t) \tag{3.6}$$

This discretized approximation can be solved analytically in time, yielding:

$$p(\theta, t) = e^{M_L t} p(\theta, 0) \tag{3.7}$$

where $p(\theta, 0)$ is the initial state of memory after encoding.

We wanted to dissociate load-driven changes in the dynamics of memory and encoding. To capture differences in encoding, we allowed the state of a memory at the start of the delay, $p(\theta, 0)$, to vary as a function of load. To simulate the encoding process, we first initialized a narrow probability density $P_0(\Theta)$ that reflects the color of the target stimulus. P_0 is a von mises distribution with mean equal to the target color Θ and a standard deviation of 0.1 radians:

$$P_0(\Theta) = \phi\left(\Theta, 0.1\right) \tag{3.8}$$

We then allowed P_0 to propagate for a 1 second encoding period according to the following differential equation:

$$d\theta = \beta_L^* G(\theta) dt + \sigma_L^* dW \tag{3.9}$$

where β_L^* and σ_L^* interact to set the bias and variance of the encoded memory. Therefore, $p(\theta, 0)$ is calculated as:

$$p(\theta, 0) = e^{M_L^*} P_0(\Theta) \tag{3.10}$$

and the final probability distribution describing the memory of the target hue after a memory delay of t seconds on a trial with load n is:

$$p(\theta, t) = e^{M_L t} e^{M_L^*} P_0(\Theta) \tag{3.11}$$

All drift and diffusion parameters (β_L, σ_L, β_L^*, and σ_L^*) are rates; they measure the change in memory over time (either due to drift or diffusion). However, care must be taken when directly comparing the value of these parameters across the encoding and memory periods. This is because

encoding is modeled as occurring over a fixed period (1 second), while the length of the memory delay can change from trial to trial. Therefore, the degree to which memory dynamics influence reports depends on the length memory delay. Drift and diffusion can be compared more directly within the encoding or memory periods.

Equation 3.11 describes the probability distribution for the memory of the target color Θ at time t. However, our goal is to predict the subject's report on a particular trial, $p(\hat{\theta}, t)$, which does not just depend on the color of the target (Bays et al., 2009; Zhang and Luck, 2008). On some trials, subjects may experience complete failures of memory, resulting in random guessing. On other trials, subjects may commit a 'swap' error and report their memory of one of the non-target colors, θ_i^* (note that the memory of non-target colors also evolved according to equation 3.11). Finally, random error may be introduced at decoding. To account for these additional influences, we estimated each subject's probability of committing swap errors and guessing, and, for each trial, computed a mixture of the target memory distribution, the non-target memory distributions, and a uniform component:

$$p(\hat{\theta}, t) = (1 - \lambda - \alpha)p(\theta, t) + \alpha\frac{1}{m}\sum_{i=1}^{m} p(\theta_i^*, t) + \lambda\frac{1}{2\pi} \tag{3.12}$$

where m is the number of non-target colors (0 or 2 for humans, 0 or 1 for monkeys). α and λ represent the probability of swap errors and guesses, respectively. They are linear functions of t parameterized by a slope a and intercept b. We estimated a unique λ and α function for each load (note that α takes on a value of zero when load is 1). To capture decoding error, we circularly convolved the final response distribution with a von mises distribution with a standard deviation $\sigma^\dagger$. As noted below, we found the model with response error fit well to human behavior. However, monkey behavior was best explained without this term.

We found the maximum likelihood estimate (joint likelihood across trials) of the free parameters $\beta_L, \beta_L^*, \sigma_L, \sigma_L^*, a_{\lambda_L}, b_{\lambda_L}, a_\alpha, b_\alpha, w_j$, and $\sigma^\dagger$ (humans only) using gradient descent. To obtained bootstrapped distributions of the parameter distributions for human subjects, we repeatedly resampled the parameters fit to each subject with replacement and took the mean of these values. To obtain bootstrapped distributions for monkey subjects, we repeatedly resampled each monkey's pool of trials with replacement and repeated the fitting process. Model comparison was performed on data pooled across sessions (monkeys) or subjects (humans).

Model fits indicated that random guessing increased with time for human subjects (Fig. S3.10), consistent with previous reports (Pertzov et al., 2017; Rademaker et al., 2018; Shin et al., 2017). Guessing decreased with delay, however, for the two monkeys. We wanted to ensure that trade-

offs between guessing and other parameters, such as the rate of diffusion, were not driving the effects of increased drift and diffusion with load. So, we fit different versions of the model in which we systematically simplified our parameterization of guess rate. Across the two monkeys, model comparison using AIC and BIC indicated that the full model was the best fit to the data. Regardless, for all models, drift and diffusion increased with load, indicating that this is a stable feature (Section 3.10).

Model comparison indicated that the full model including decoding error was clearly better than the model without decoding error in humans. However, the model with decoding error was not clearly better than a model without decoding error across monkeys and so we defaulted to the simpler model (monkey E: $wBIC = 0.01$; monkey W: $wBIC = 1.00$; compared to $wBIC = 1.00$ in humans). Furthermore, in exploratory tests we found decoding error substantially disrupted the ability of the model to predict the clustering and precision of responses in monkey W; with decoding error the correlation between the predicted and observed response distribution in monkey W dropped from .741 to .393 and the correlation between the predicted and observed pattern of precision across colorspace dropped from .377 to .120. Based on this, we concluded that models with decoding error best described the human behavior but that the simpler model without decoding error best described the monkey behavior. Differences in decoding error could reflect different response modalities (moving a mouse for humans, saccade for monkeys) or reflect the fact that monkeys saw the entire color wheel while humans did not.

3.5.11 Simulated error of models over time

We wanted to identify if attractor dynamics might be normative and enhance the fidelity of memory. To do this, we computed the expected mean error for the memory of a target color as a function of delay time for the full dynamic model with attractor dynamics (drift + diffusion) and a model without attractor dynamics (diffusion). The drift and diffusion parameters of the drift + diffusion model were set to the mean fit parameters for the human subjects in Experiment 1a. The parameters of the diffusion model were identical except that β_L and β_L^* were set to zero. To isolate error in the representation of the target color, the probabilities of guessing and swaps were set to zero. To create a representative drift function, we fit our basis set to the numerical derivative of the PDF of the response distribution for human subjects (normalized to have a maximum absolute value of one), which yields attractors at locations in color space where they are most frequently observed (i.e., at commonly reported colors). To create biased target distributions, we parametrically took a weighted

average of a distribution that was entirely uniform over color space and a biased distribution that was uniformly distributed within 10 degrees of attractor states and zero elsewhere.

3.5.12 Nonlinear mapping between stimulus and perceptual space

The color space used to parameterize stimuli in these experiments (CIELAB) is designed to be perceptually uniform, but we sought to demonstrate that inhomogeneties in this space cannot explain our results. To demonstrate this, we analyzed an alternative model (Fig. S3.5a) which assumes a nonlinear mapping between our stimulus parameterization (a circle in CIELAB space) and a hypothetical true perceptual space (a square, although the results generalize to other shapes). The continuous CEILAB and perceptual spaces were discretized into 1024 points. We simulated memory reports by first generating 100,000 angles randomly distributed around our stimulus space (representing the target stimuli) and projecting these points onto the true perceptual space (representing encoding). Memory was simulated as a purely diffusive process of the encoded target colors around the true perceptual space (i.e., there were no discrete attractor dynamics). Simulations were run for 1,000 timesteps (arbitrary units). Diffusive noise at each timestep was modeled as random step between 0 and 4 points in either direction in the discretized perceptual space. Report was simulated by projecting the diffused memory representations back into stimulus space. This model predicts clustering of memory reports (Supplementary Figure 5a) but does not predict attractive bias around cluster peaks (Fig. S3.5b) as observed empirically. We thank and anonymous reviewer for proposing and implementing this alternative model.

3.6 Acknowledgments

We thank A. Piet for suggesting trial-by-trial analysis, B. Morea and H. Weinberg-Wolf for assistance with NHPs, and S. Henrickson, F. Bouchacourt, A. Libby, and P. Kollias for comments. This work was supported by NIMH R56MH115042 and ONR N000141410681 to TJB, an NDSEG fellowship to MFP, and McKnight Foundation, Simons Collaboration on the Global Brain (SCGB AWD1004351) and the NSF CAREER Award (IIS-1150186) to JWP.

3.7 Collaborators

The work described in this chapter was conducted in collaboration with Brian DePasquale, Jonathan Pillow, and Tim Buschman. These results have been published elsewhere (Panichello et al., 2019).

3.8 Author contributions

MFP and TJB conceived of the experiments; MFP, BD, JWP, and TJB designed the dynamical model; MFP and BD implemented the model; MFP collected and analyzed the data; MFP and TJB wrote the original draft; MFP, BD, JWP, and TJB discussed the results and prepared the final draft.

3.9 Code availability

Code for fitting the discrete attractor model to behavioral data from delayed estimation tasks is available at https://github.com/buschman-lab.

3.10 Supplementary figures

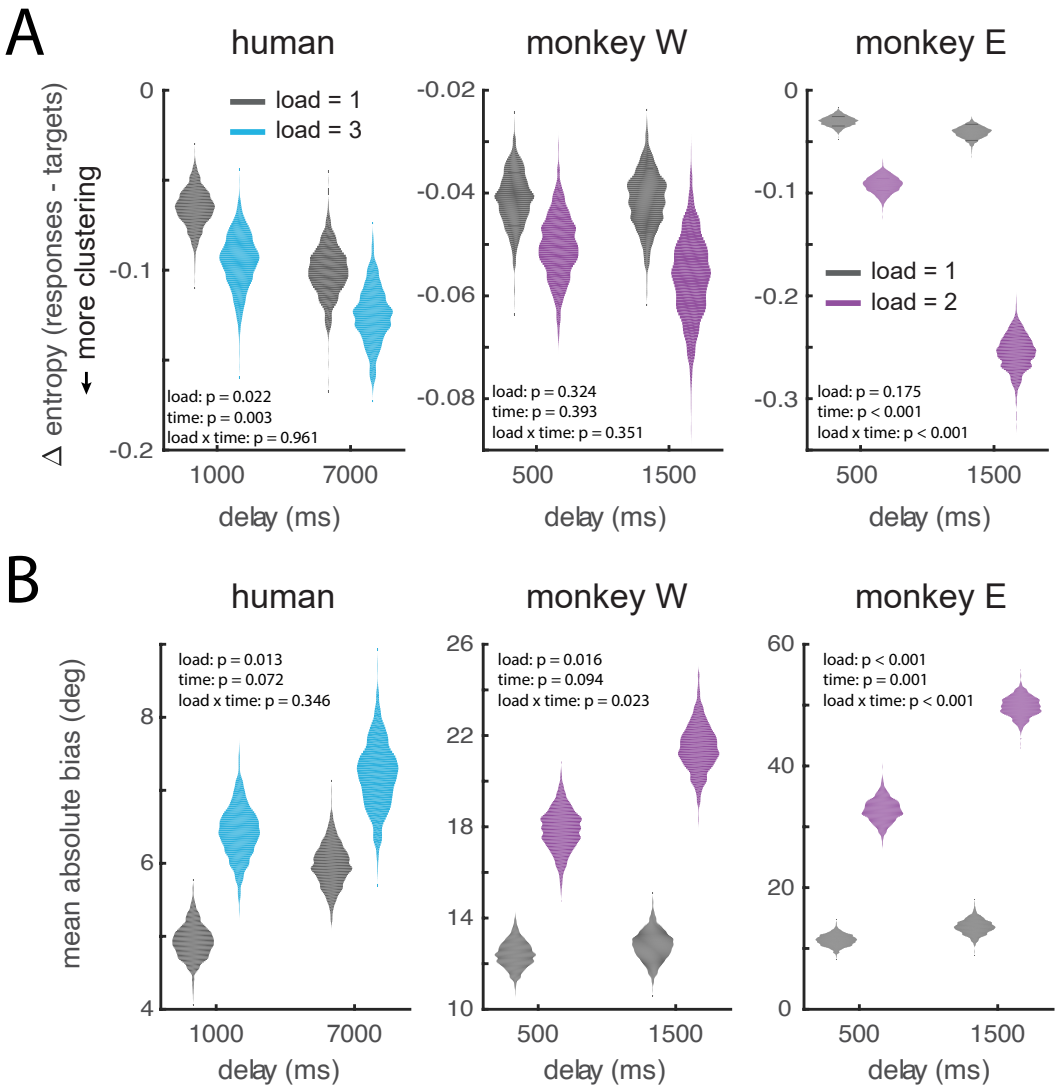


Figure S3.1: Clustering increases with load and delay. (A) Difference in entropy between the response distribution and target distribution for humans and monkeys as a function of load and delay. More negative values indicate more clustered memory reports. (B) Mean absolute bias (averaged across all target colors) for humans and monkeys as a function of load and delay. Violin plots indicate distribution of bootstrapped values. P-values reflect non-parametric regression (bootstrap).

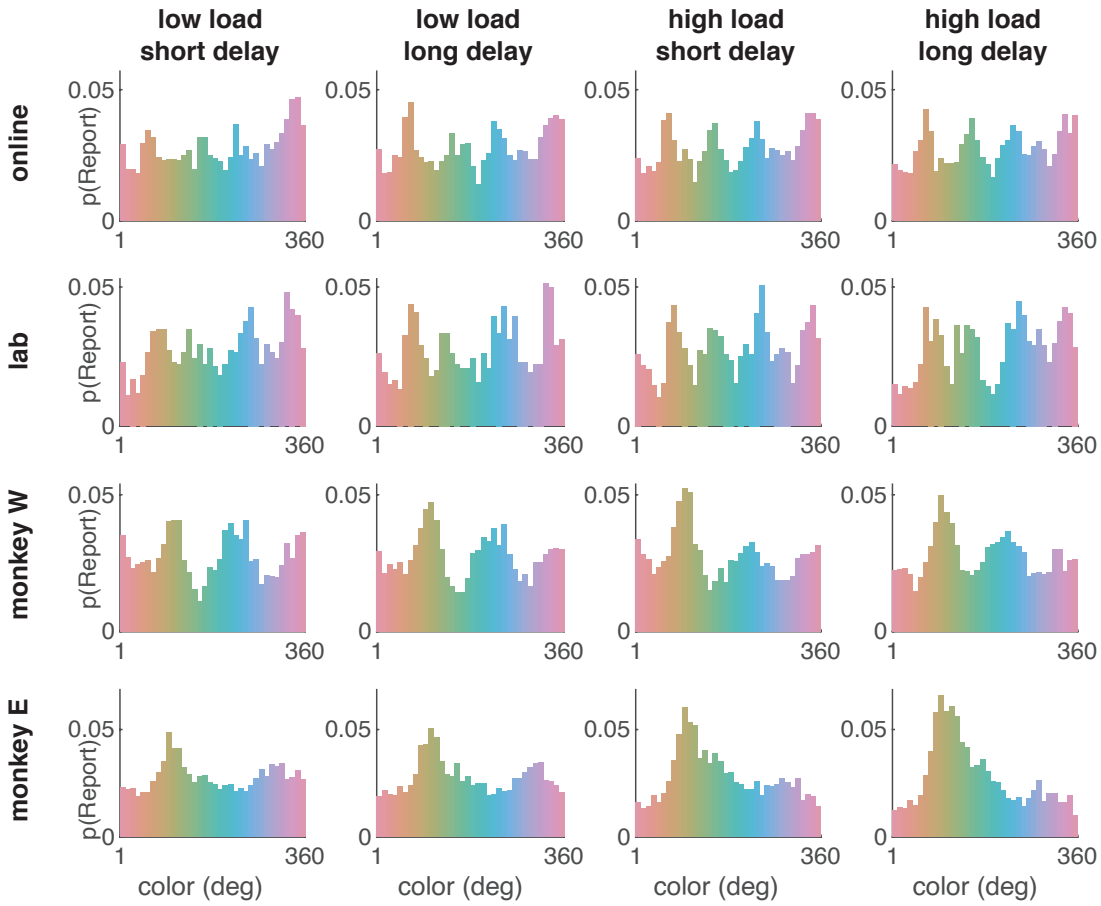


Figure S3.2: Response histograms for humans and monkeys by condition.

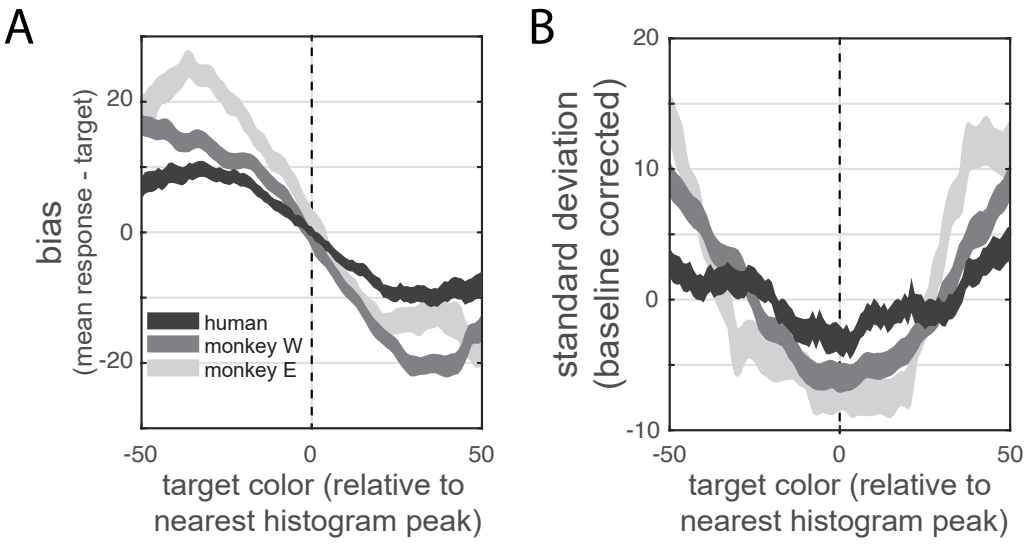


Figure S3.3: Bias and standard deviation of memory reports around putative attractors. Putative attractors are identified as significant peaks in subjects' distribution of reported colors. Error bars reflect standard error of the mean.

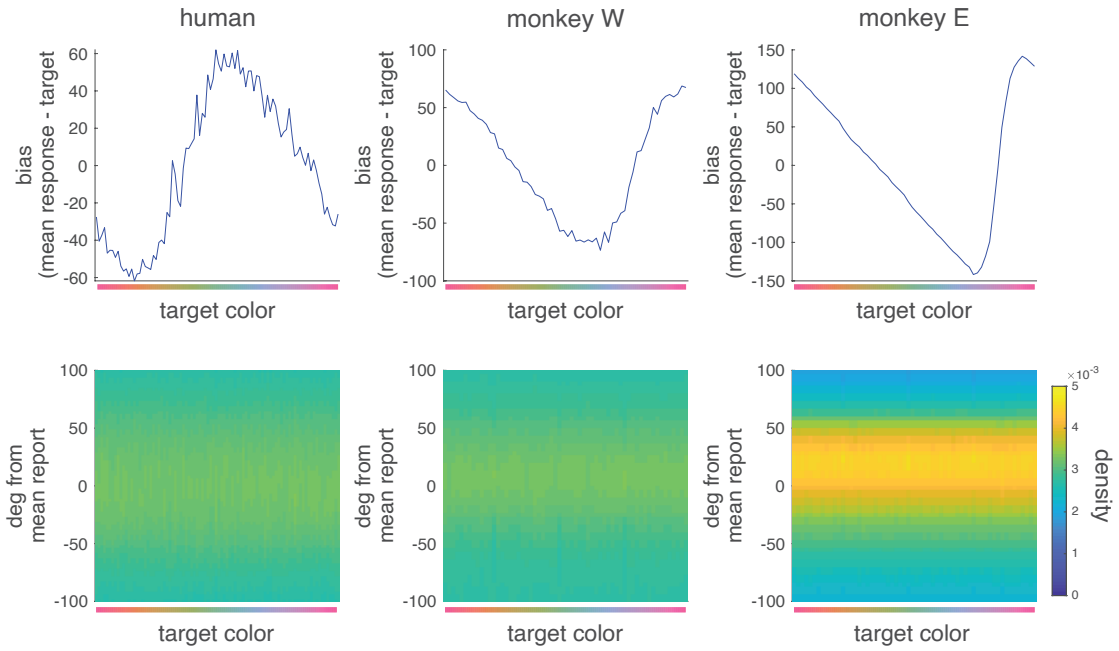


Figure S3.4: Simulated performance of a non-uniform guessing strategy. In the non-uniform guessing strategy, the subject reports one of the frequently-reported colors on a subset of trials, and the color reported is independent from the identity of the target (see Methods). Plots show the expected pattern of bias (top row) and precision around mean report (bottom row) as a function of target color. For the subset of trials on which the subject makes a non-uniform guess, bias depends only on the distance between the target color and the mean reported color across all trials (top row). Additionally, precision does not vary as a function of target color (bottom row).

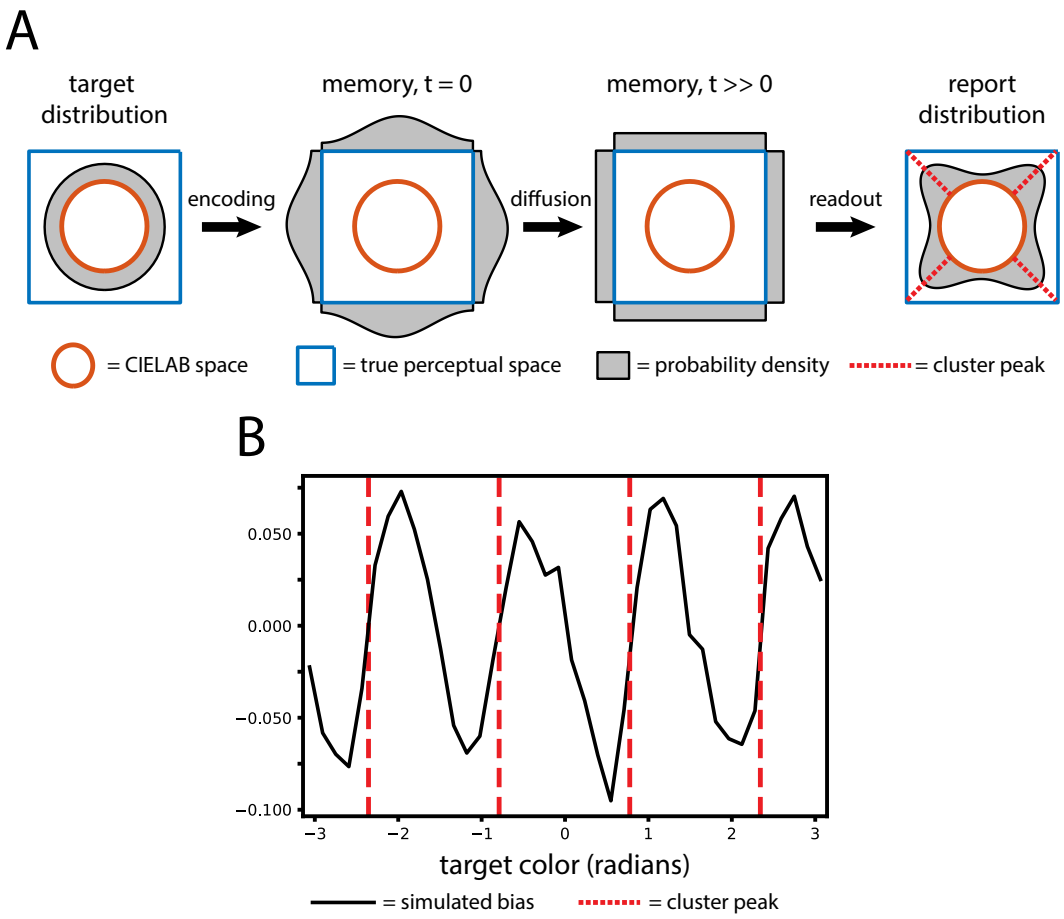


Figure S3.5: Simulated performance based on a nonlinear encoding of CIELAB color space. A nonlinear mapping between stimulus space and a subject's true representational space introduces clustering into memory reports without discrete attractor dynamics but cannot explain memory biases. (A) Model structure. Far left: across trials, target colors are uniformly distributed in CIELAB space (orange circle). Center left: true perceptual space is assumed to be any arbitrary shape (here: a square) other than a circle concentric with CIELAB space. When the uniform target colors are projected into this true space, clusters form at locations where changes in the CIELAB angle θ result in small changes in the true space. Center right: random diffusion in memory erodes the concentration gradient in the true perceptual space over time. For clarity, we show a complete erosion of the concentration gradient at $t >> 0$, but in practice the concentration gradient will only partially degrade for delays of a few seconds when reports are still reasonably accurate. Far right: Projecting the uniform distribution of memories in true space back into CIELAB space results in clustering at locations where changes in the true space result in small changes in θ. For a square perceptual space, this results in clustering at vertex angles, which may be mistaken for attractors. (B) Predicted bias based on 100,000 simulated trials. Counterintuitively, this model predicts repulsive (positive slope) bias around points of peak clustering, inconsistent with empirical results (Fig. S3.3). We thank an anonymous reviewer for proposing and implementing this alternative model.

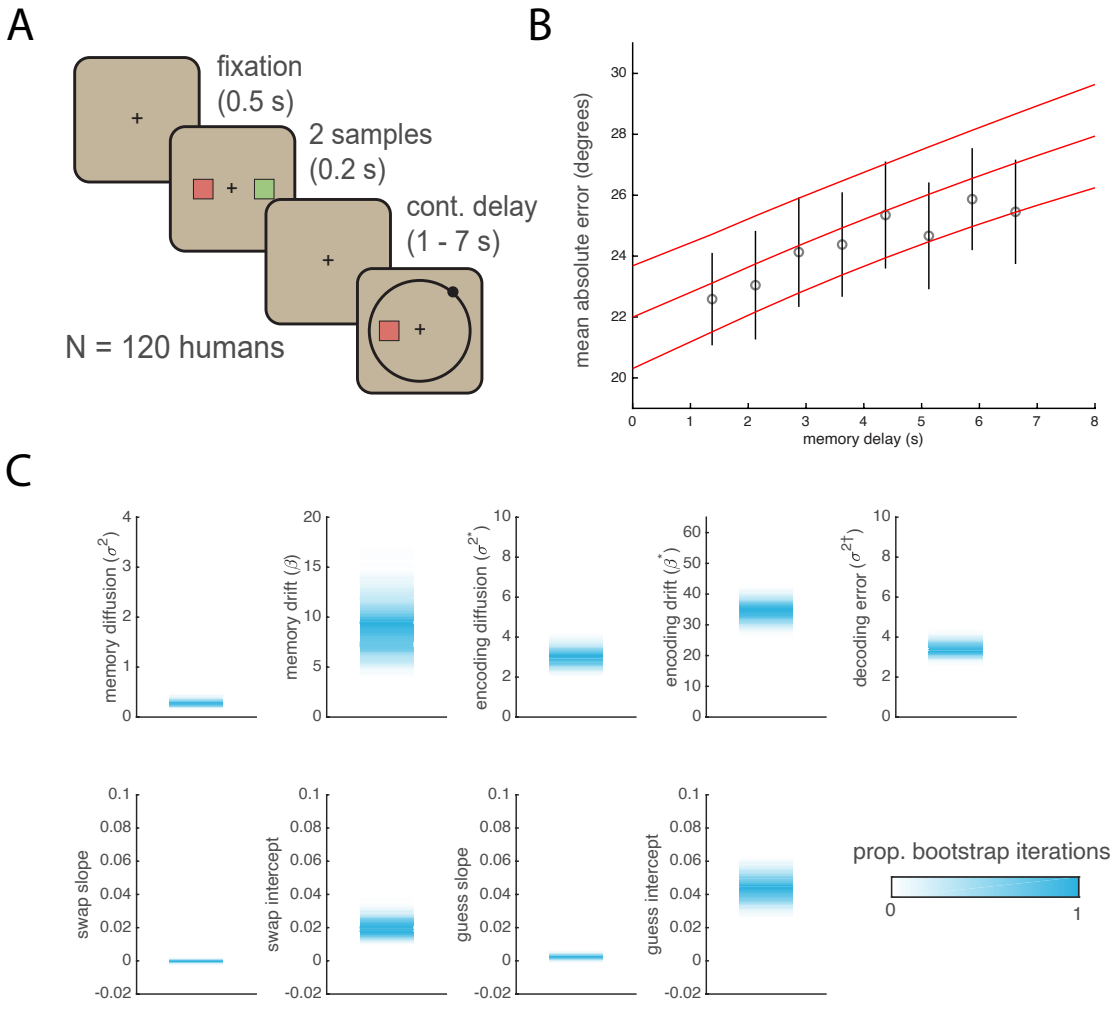


Figure S3.6: Experiment 1b design and results. (A) Experiment 1b design. Experiment 1b was similar to 1a, except that there were always two samples and the delay varied continuously between 1 and 7 seconds (see methods). (D) Mean absolute error +/- 95% CI (bootstrap) as a function of delay length. Red = model fit, black = data. (C) Maximum likelihood dynamic model parameter estimates for Experiment 1b. Color intensity reflects normalized proportion of bootstrap iterations.

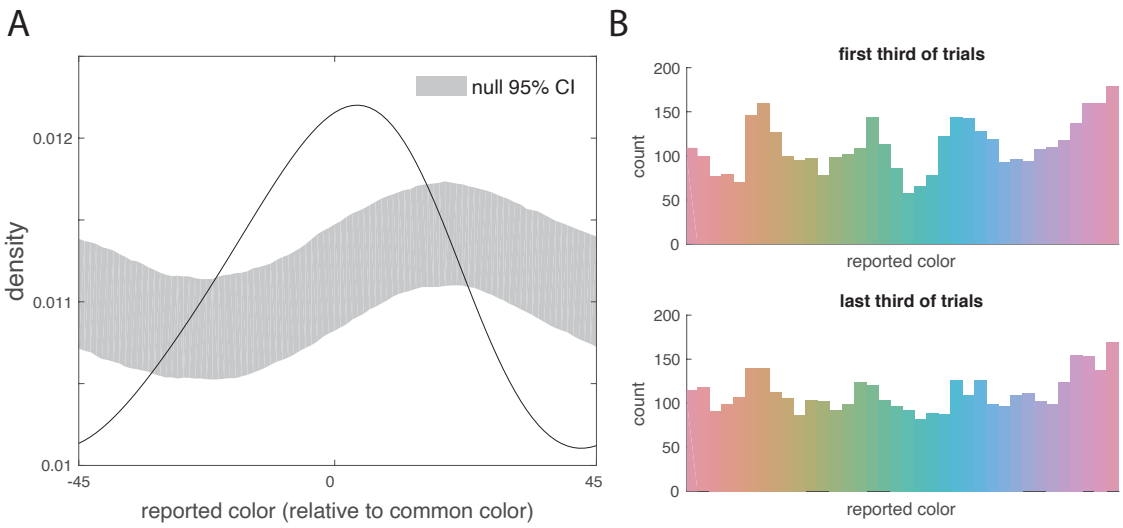


Figure S3.7: Distribution of color reports in Experiment 2. (A) Probability of report relative to common color location in colorspace, computed using the subset of trials in which target colors were distributed uniformly. (B) Distribution of reported colors for the first and last third of trials, computed using the subset of trials in which target colors were distributed uniformly.

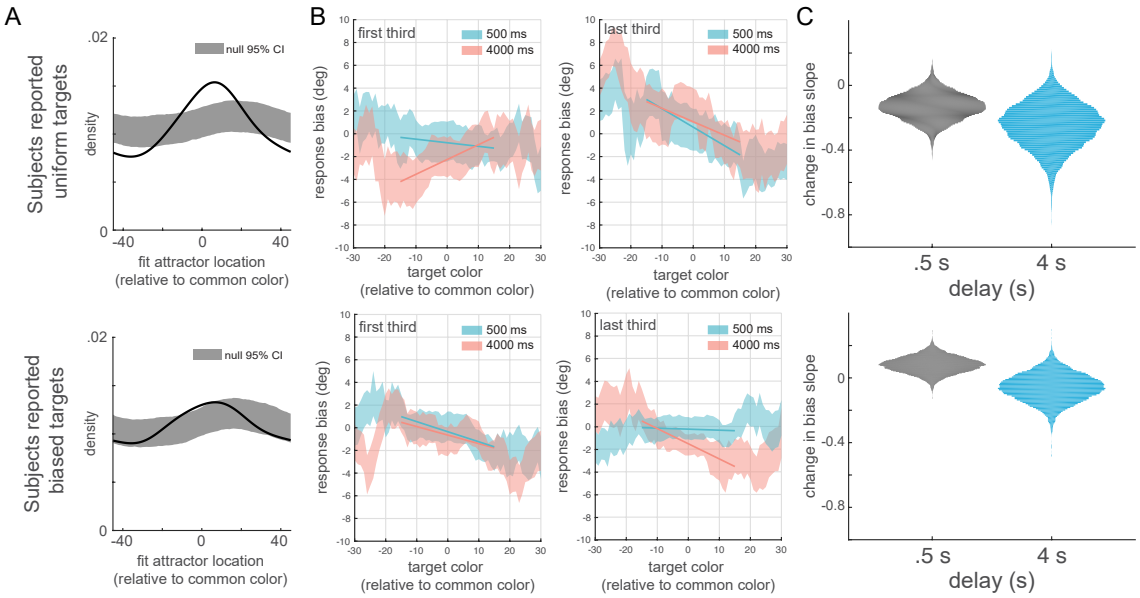


Figure S3.8: Performance of Experiment 2 subjects, grouped by debriefing report. Regardless of whether experiment 2 subjects incorrectly reported that the distribution of target colors was unbiased (top row) or correctly reported that the distribution of target colors was biased (bottom row), both groups were more likely than chance to display attractors at common color locations (A) and both groups showed a numerical trend for slope to decrease more on long-delay trials (B-C).

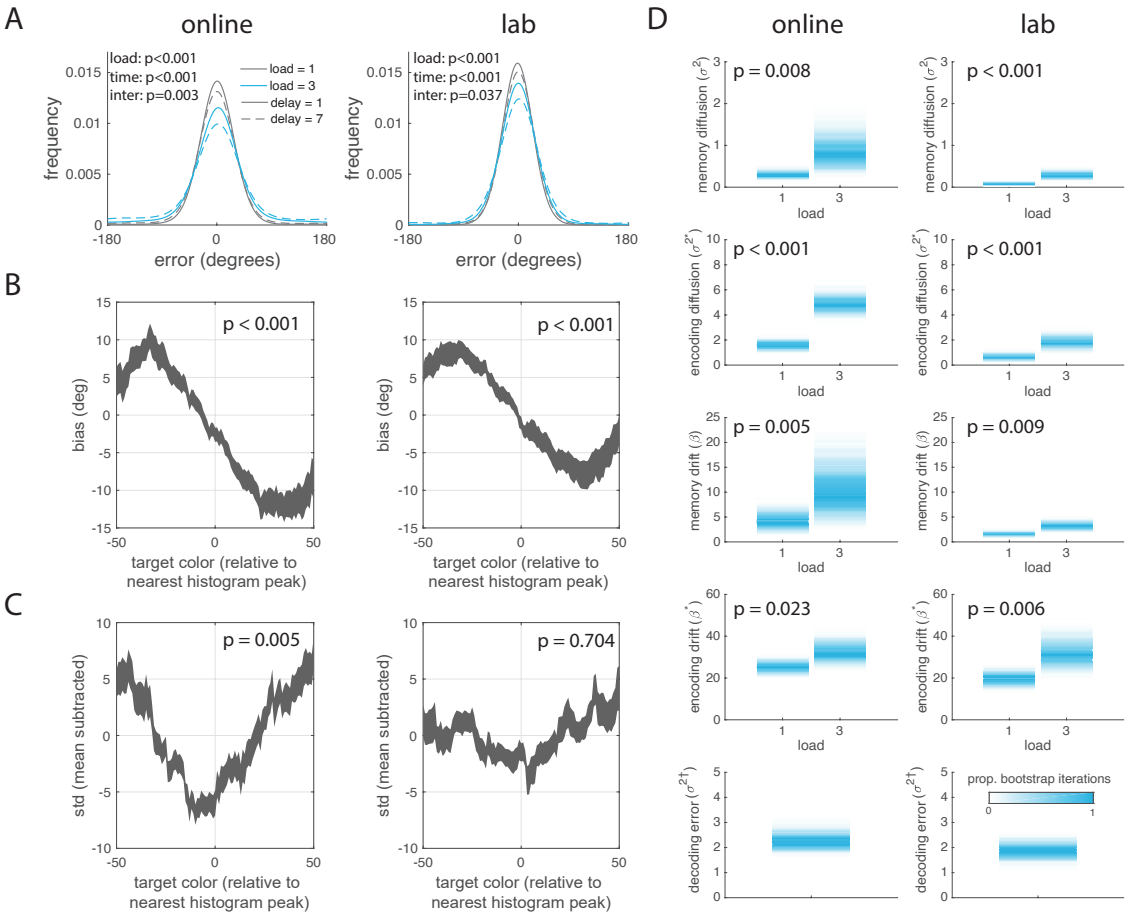


Figure S3.9: Online and lab subjects show qualitatively similar behavior. (A) Distribution of angular error. P-values reflect the results of a repeated-measures ANOVA predicting mean error as a function of load and time, as in text describing Fig. 1b. (B) Bias around putative attractors. P-values reflect a t-test of the slope of bias at histogram peaks vs zero, as in text describing Fig. S3. (C) Precision around putative attractors. P-values reflect a t-test of the relative standard deviation of memory reports at histogram peaks vs zero, as in text describing Fig. ED5. (D) Dynamical model parameter fits. P-values reflect differences in diffusion and drift parameters as a function of load, as in the text describing Fig. 6.

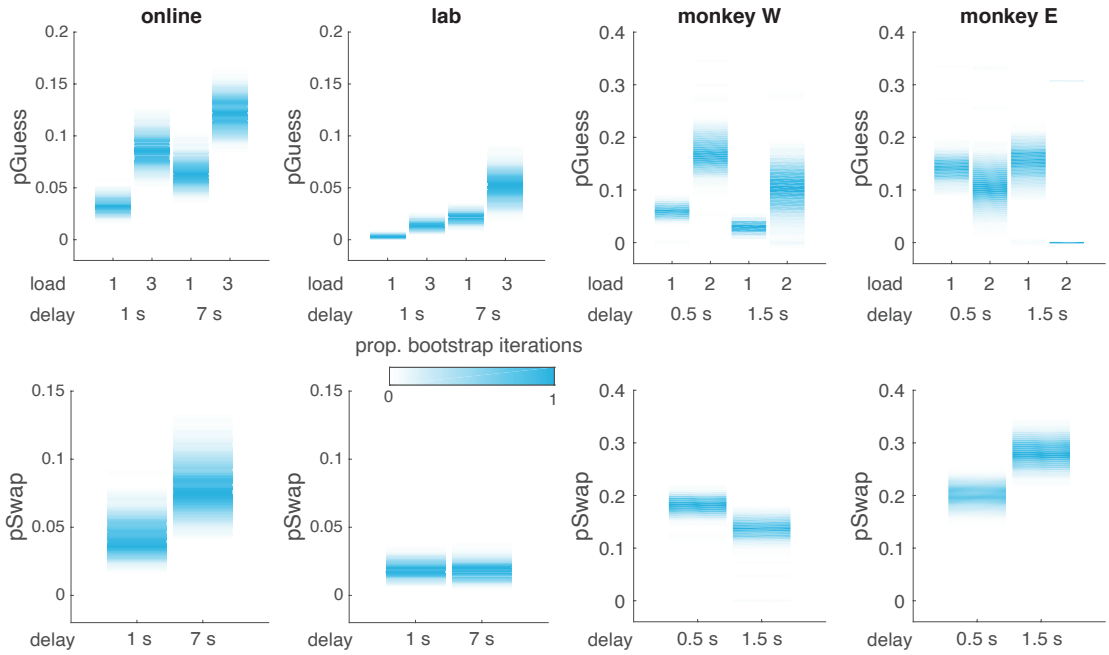


Figure S3.10: Estimated rate of guessing and swap errors. Plots show the maximum likelihood guess and swap probabilities from dynamic model fits for each load and delay. Color intensity reflects normalized proportion of bootstrap iterations.

population	model	no. param	AIC	ΔAIC	wAIC	BIC	ΔBIC	wBIC
human	full	27	119885	0	1.00	120094	0	0.98
	drop β_L	25	119908	23	0.00	120102	7	0.02
	drop β_L^*	25	120500	615	0.00	120694	600	0.00
online	full	27	85108	0	0.99	85308	6.2	0.04
	drop β_L	25	85117	8.6	0.01	85302	0	0.96
	drop β_L^*	25	85502	393	0.00	85686	385	0.00
lab	full	27	34016	0	1.00	34194	0	0.99
	drop β_L	25	34038	22	0.00	34203	8.8	0.01
	drop β_L^*	25	34310	294	0.00	34475	281	0.00
monkey W	full	26	129087	0	1.00	129286	0.0	0.80
	drop β_L	24	129105	18	0.00	129289	2.8	0.20
	drop β_L^*	24	129654	567	0.00	129838	552	0.00
monkey E	full	26	144746	0	1.00	144947	0	1.00
	drop β_L	24	144873	126	0.00	145058	111	0.00
	drop β_L^*	24	144871	125	0.00	145057	110	0.00

Table 3.1: AIC and BIC model comparison. We compared the full model with competing models without attractor dynamics during encoding or maintenance. Model weights (wAIC and wBIC) indicate the probability that the given model is the best model in the set given the data and set of candidate models.

subject	full	drop β_L	drop β_L^*
human	15.3	14.7	0
monkey W	14.5	14.1	0.6
monkey E	5.0	2.0	1.9

Table 3.2: Cross-validated model comparison. Mean difference in 20-fold cross-validated log-likelihood for full model and competing models without attractor dynamics during encoding or maintenance. Values represent the increase in log-likelihood relative to the worst fitting model, averaged across folds.

		p(Guess) eq.				parameter MLE							
subject	model	C	L	D	I	σ_1	σ_2	σ_1^*	σ_2^*	β_1	β_2	β_1^*	β_2^*
monkey W	1	x	x	x	x	15	31	12	13	4	15	33	45
	2	x	x	x		15	29	12	14	4	14	32	44
	3	x		x		15	36	11	17	5	17	32	49
	4	x	x			13	26	14	16	4	13	32	44
	5	x				12	32	14	21	4	16	31	50
	6					13	21	32	69	0	9	161	344
monkey E	1	x	x	x	x	17	39	48	58	28	45	8	82
	2	x	x	x		23	35	44	71	33	39	4	88
	3	x		x		23	35	47	57	32	44	2	85
	4	x	x			17	29	47	76	29	34	6	84
	5	x				17	30	52	60	28	40	4	80
	6					17	29	69	76	25	35	5	85

Table 3.3: Parameter fits for simplified models of guessing and swap behavior. Maximum likelihood estimates for drift and diffusion parameters for models with different parameterizations of guessing probability. An 'x' indicates that a parameter is included in a given model. For the most flexible model (model 1, identical to that reported in the main text), guessing is effectively parameterized by a constant term C, a coefficient determining an effect of load on guessing (L), a coefficient determining an effect of memory delay on guessing (D), and an interaction term (I). Successive models drop combinations of these terms, yielding less flexibility in how guessing changes with load and time. For example, for model 5, p(Guess) is constant across load and time. Regardless of the parameterization, however, drift and diffusion consistently increase with load during both encoding and memory.

Chapter 4

Transformation of memories and percepts by attention

4.1 Abstract

Cognitive control guides behavior by controlling what, where, and how information is represented in the brain. Previous work has shown parietal and prefrontal cortex direct attention, which controls the representation of external sensory stimuli. However, the neural mechanisms controlling the selection of representations held 'in mind', in working memory, are unknown. Here, we show prefrontal cortex controls working memory by selectively enhancing and transforming its contents. Monkeys were trained to switch between two tasks, requiring them to either select an item from a set of items held in working memory or attend to one stimulus from a set of visual stimuli. Simultaneous neural recordings in prefrontal, parietal, and visual cortex found prefrontal cortex played a primary role in selecting an item from working memory, representing selection before parietal and visual cortex. Surprisingly, the same representation that controlled selection also directed attention to an external stimulus, suggesting prefrontal cortex may act as a general controller. Selection acted on memory representations by strengthening the selected item and transforming it in a task-dependent manner. Before selection, when both items were relevant to the task, the identity of each item was represented in an independent subspace of neural activity. After selection, the representation of only the selected item was strengthened and transformed into a new subspace that was used to guide the animal's behavioral report. Together, our results show how prefrontal cortex controls working memory, selectively enhancing and transforming memories to support behavior.

4.2 Introduction

So far, we have provided evidence that expectations transform percepts and memories in a homologous manner. We next examine if another top-down cognitive processes, attention, similar impacts perception and working memory. Items held in working memory (e.g. the list of specials at a restaurant) are thought to be represented in a distributed network of brain regions, including prefrontal cortex, parietal cortex, and sensory cortex (Christophel et al., 2017). A control mechanism can then select a specific item from working memory and use it to guide behavior (Ester et al., 2018; Gazzaley and Nobre, 2012; Myers et al., 2017; Sprague et al., 2016) (e.g. selecting a special to order for dinner). This process is homologous to attention, which selectively enhances task-relevant sensory inputs (Buschman and Kastner, 2015; Desimone and Duncan, 1995). Previous functional imaging work has shown prefrontal and parietal cortex are active when an item is selected from working memory (LaBar et al., 1999; Nee and Jonides, 2009; Nobre et al., 2004). However, because it has never been studied at the level of single neurons, the neural mechanisms of selection remain unknown.

To address this, we simultaneously recorded from the prefrontal, parietal, and visual cortices of two monkeys (*Macaca mulatta*) as they selected one of two items held in working memory. On each trial of the experiment, the animals remembered the color of two squares (Fig. 1A, an 'upper' and 'lower' stimulus). After a memory delay, the animals received a cue that indicated whether they should report the color of the 'upper' or 'lower' square (now held in working memory). This cue was followed by a second memory delay, after which the animals reported the color of the cued square by saccading to the matching color on a color wheel (note, the wheel was randomly rotated on each trial to prevent motor planning). Therefore, to perform the task, the animals had to hold two colors in working memory, select the color of the cued square, and then use it to guide their response to the color wheel.

4.3 Results

4.3.1 Attention and selection reduce behavioral errors

Overall, both monkeys performed the task well; mean angular error between the presented and reported color was 51.8°(Fig. 4.1b-c and Fig. S4.1a). As expected (Bays et al., 2009; Wilken and Ma, 2004; Zhang and Luck, 2008), accuracy depended on the number of items in memory (i.e. the 'memory load') – angular error was greater when two colored squares were presented, compared to

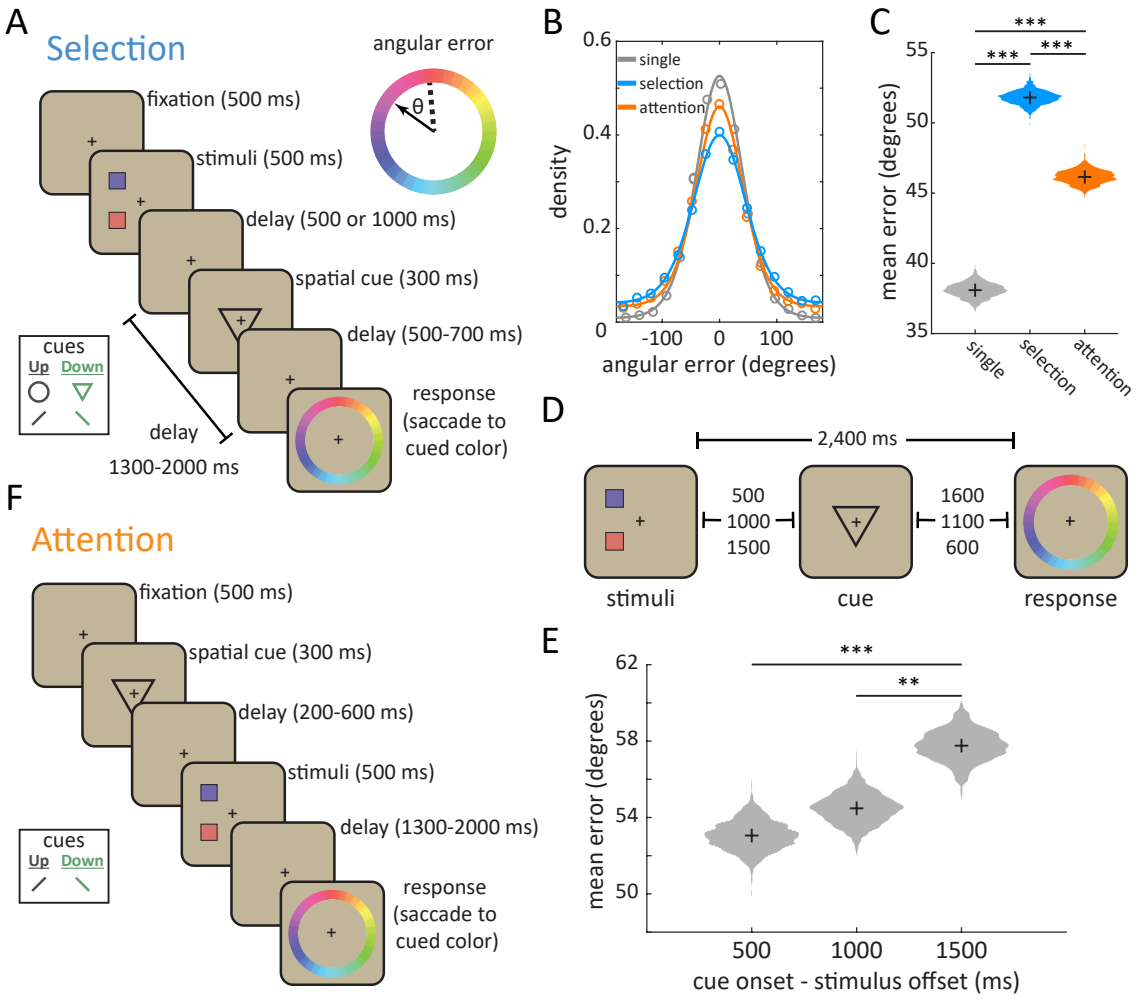


Figure 4.1: Monkeys use selection and attention to control the contents of working memory. (A) Animals were trained to perform two variants of a delayed estimation paradigm. On each trial, two colored squares were presented (one 'upper' and one 'lower' stimulus). Lower inset shows symbolic cues presented at fixation to indicate whether the upper or lower stimulus should be reported at the end of the trial in order to receive reward. In the 'selection' condition, the cue appeared during a memory delay after stimulus offset, requiring the animal to select an item from working memory. Animals received a graded juice reward for making an eye movement to the portion of a color wheel matching the color of the cued stimulus. Color wheel inset shows error was calculated as the angular deviation between the presented color (dashed) and the reported color (solid). (B) Distribution of angular error (circles) with best-fitting mixture models (lines, see methods) for single item trials (gray), selection trials (blue), and attention trials (orange). (C) Bootstrapped distribution of mean angular error in the selection, attention, and single-stimulus conditions (colors as in B). (D) In a separate behavioral experiment, we fixed the total memory delay of the selection condition and systematically varied the length of the delay between stimuli offset and cue onset. (E) Increasing the time before selection increased error. Distribution shows mean angular error as a function of the delay between stimulus and selection cue (bootstrapped). Bars and asterisks reflect paired randomization tests. (F) Animals also performed an 'attention' condition, interleaved with selection trials (in blocks). In this condition, the cue appeared before stimulus onset. Inset: the two symbolic cues used in the attention condition. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

when only one square was presented (Fig. 4.1b-c and Fig. S4.2a; mean absolute error was 38.1°for 1 item, 51.8°for 2 items, $p < 0.001$, randomization test). The increased error with two items in memory is likely due to interference between the items (Bays, 2015; Bouchacourt and Buschman, 2019; Buschman et al., 2011; Sprague et al., 2014). Selecting an item in working memory is thought to reduce such interference(Bays and Taylor, 2018; Pertzov et al., 2013). Consistent with this, reports were more accurate when selection occurred earlier in the trial (Fig. 4.1d-e and Fig. S4.2b, 53.1°, 54.4°, and 57.8°for 0.5, 1, and 1.5 s post-cue, respectively; linear regression, $\beta = 4.67 +/- 1.08$ SEM, $p < 0.001$, bootstrap).

In addition to the selection condition, animals also performed an 'attention' condition. On attention trials, the cue was presented before the colored squares, allowing the animal to attend to the location of the to-be-reported stimulus (Fig. 4.1f). Memory reports were more accurate in the attention condition than in the selection condition (Fig. 4.1b-c and Fig. S4.2b; 46.1°vs. 51.8°,$p < 0.001$, randomization test). In addition, the effect of memory load was reduced; increasing the number of stimuli from 1 to 2 led to a smaller increase in error on attention trials (9.01°vs. 13.7°for attention/selection, $p < 0.001$, bootstrap). This is consistent with attention reducing interference between stimuli(Desimone and Duncan, 1995; Treue and Maunsell, 1996) and modulating what enters working memory(Everling et al., 2002).

4.3.2 Attention and selection share a population code

To understand the neural mechanisms of selection, we simultaneously recorded from four regions known to be involved in working memory (Fig. 4.2a) – lateral prefrontal cortex (LPFC; 682 neurons), frontal eye fields (FEF; 187 neurons), parietal cortex (7a/b; 331 neurons), and intermediate visual area V4 (341 neurons). Neurons in all four regions carried information about which item was selected from working memory (i.e. the upper or lower item; Fig. 4.2b and Fig. S4.3a-b). To quantify information about selection, we trained a logistic regression classifier to decode the location of selection from the firing rates of populations of neurons recorded in each region (Fig. 4.2c; pseudo-populations were constructed across all recording days, see methods for details). As seen in Fig. 4.2d, the classifier could decode the location of selection in all four regions (blue lines). However, significant information about selection emerged first in LPFC and then later in posterior regions (Fig. 4.2e, 175 ms post-cue in LPFC, 245 ms in FEF, 285 ms in parietal, and 335 ms in V4). LPFC was significantly earlier than parietal cortex and V4 ($p = 0.005$ and $p = 0.048$, randomization test; but statistically indistinguishable from FEF, $p = 0.371$). This suggests that signals necessary for controlling selection

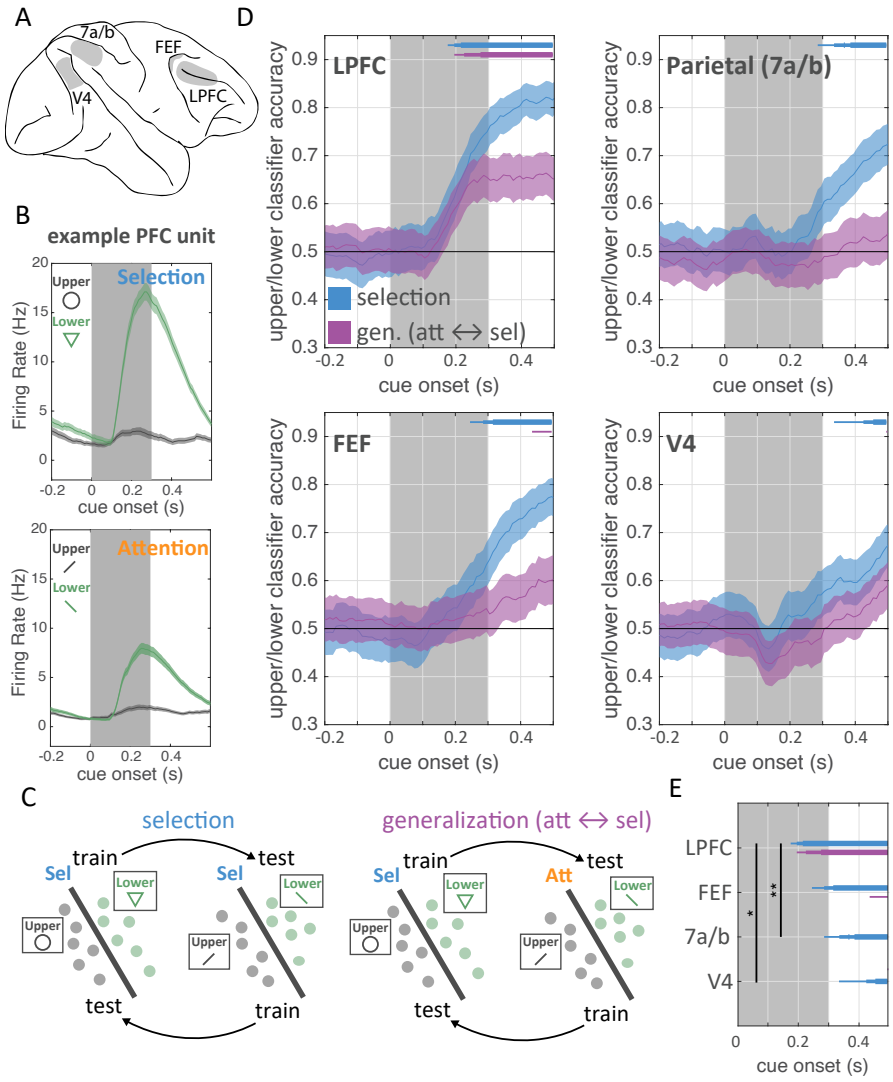


Figure 4.2: Selection is directed from prefrontal cortex and shares a population code with attention. (A) Neural activity was recorded simultaneously from lateral PFC (LPFC), the frontal eye fields (FEF), parietal cortex (area 7a/b), and visual area V4. (B) Firing rate of an example PFC neuron around cue onset when the upper (gray) or lower (green) stimulus was cued in the selection and attention conditions. Shaded regions are standard error of the mean. Inset shows different cues used for selection and attention. (C) Training and testing regime for classifiers designed to quantify information about the location of selection and attention based on population firing rates. Classification accuracy was measured on held-out data both within selection (left) and across selection and attention (right). Classifiers were trained and tested on different cue sets and performance was averaged across these splits. (D) Timecourse of information about the location of selection and attention for each brain region (labeled in upper left). Lines show mean classification accuracy around cue onset for the classifier trained and tested within the selection condition (blue) and across selection and attention (purple). Error bars are standard error of the mean. Bars along top indicate above-chance classification: p ¡ 0.05, 0.01, and 0.001 for thin, medium, and thick lines, respectively. (E) Timepoints of significant classification for each brain region, as in (D). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

emerge first in prefrontal cortex.

There is a functional homology between selection and attention(Gazzaley and Nobre, 2012). They both control neural representations – selection controls 'internal' working memory representations, while attention controls 'external' sensory representations. In both cases, control mitigates interference between representations (Buschman et al., 2011; Desimone and Duncan, 1995; Schneegans and Bays, 2017). Motivated by this functional homology, we investigated whether there was a shared population representation controlling selection and attention. Previous work in humans has shown both attention and selection activate prefrontal and parietal cortex(LaBar et al., 1999; Lenartowicz et al., 2010; Nee and Jonides, 2009; Nobre et al., 2004). However, it is not known if the neural mechanisms controlling selection and attention are the same. To test this, we first examined the responses of single neurons to the 'upper' and 'lower' cues on selection and attention trials. Neurons that encoded the location of selection responded similarly during attention in LPFC (Fig. S4.3c; $r(586) = 0.09$, $p = 0.036$). In contrast, sensitivity for selection and attention were uncorrelated in FEF, V4, and parietal (Fig. S4.3c; FEF: $r(169) = 0.04$, $p = 0.617$; V4: $r(318) = -0.04$, $p = 0.513$; parietal: $r(301) = 0.03$, $p = 0.612$).

Furthermore, classifiers trained to decode the location of selection generalized to decode the location of attention (and vice-versa, Fig. 4.2c; see methods). Consistent with a common mechanism in LPFC, generalization performance was significantly above chance and followed the timecourse of the selection classifier (Fig. 4.2d, purple lines). In contrast, generalization was weaker in FEF and trended towards being delayed relative to LPFC ($p = 0.12$, randomization test) and there was no significant generalization in parietal cortex or V4 (Fig. 4.2d-e; poor generalization was not due to an inability to decode the location of attention, Fig. S4.4). Together, these results suggest a common neural mechanism in LPFC controls attention to sensory inputs and selection of items in working memory.

4.3.3 Selection enhances the representation of task-relevant memories

Next, we explored how selection impacts the neural representations of items in working memory. As noted above, selection improves working memory accuracy(Griffin and Nobre, 2003; Murray et al., 2013; Pertzov et al., 2017; Sprague et al., 2016, Fig. 4.1e). To understand the neural mechanisms, we first measured color information in LPFC, FEF, parietal (7a/b), and V4. Single neurons in all four regions showed strong color selectivity (Fig. 4.3a). Selectivity was quantified by measuring the circular entropy of each neuron's firing rate in response to colors around the color wheel (see

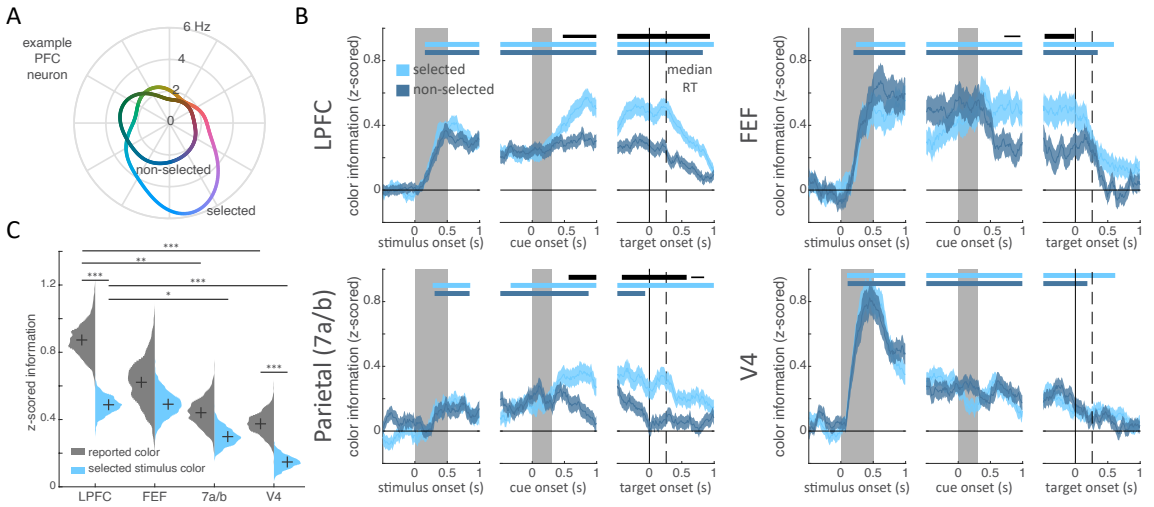


Figure 4.3: Effects of selection on color information in working memory. (A) Color tuning curve for an example LPFC neuron. Firing rate (radial axis) was averaged from 500-700 ms post-cue as a function of the color (angular axis) of the selected (light) and non-selected (dark) stimulus. This neuron carries information about the color of the selected item (i.e. the neural response is non-uniform), but this information is reduced for when the item is not selected. (B) Mean z-scored color information for the selected and non-selected color (in light and dark blue, respectively) in each brain region, averaged across all neurons. Information was quantified by calculating the entropy of the selected/non-selected color tuning curves (high entropy reflects high color information). Error bars are standard error of the mean. Horizontal bars indicate significant information for the selected item (light blue), the non-selected item (dark blue), and significant difference in information about the selected and non-selected items (black). Bar width indicate significance: p ¡ 0.05, 0.01, and 0.001 for thin, medium, and thick, respectively. All tests were cluster-corrected for multiple comparisons (see methods). (C) Information about the presented color of the selected item (light blue) and the reported color of the selected item (gray), averaged across all neurons. Both types of information were calculated on firing rates in a 200 ms window prior to onset of the response color wheel. Distributions show bootstrapped estimates of the mean. Horizontal lines indicate pairwise comparisons. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

methods). A significant proportion of neurons in each region encoded color information about either the upper or lower stimulus during the trial (LPFC: $N = 387/607$ cells; FEF: 114/178; parietal: 181/307; V4: 245/323; all $p < 0.001$, binomial test). Across the population, all four regions carried information about the color of the stimuli during their presentation (Fig. 4.3b, left panels). Color information was then maintained across this distributed network during the first memory delay (Fig. 4.3b; middle panels, before selection). Interestingly, there was less information about color on attention trials compared to selection trials (Fig. S4.5-6). This could reflect a task-specific difference in how memories are stored; recent theoretical work (Masse et al., 2019) suggests a more active representation is needed when manipulating memories (such as in the selection condition).

Selection enhanced memories across prefrontal and parietal cortex. In LPFC, color information about the selected memory was increased relative to the unselected memory, starting at 475 ms

after cue onset (Fig. 4.3b). Similar differences were seen in FEF and parietal (Fig. 4.3b, at 715 and 565 ms, respectively). In V4, selection did not impact memory representations (Fig. 4.3b; although information about the selected item tended to increase with the accuracy of memory reports, Fig. S4.7). The selective enhancement of a memory was related to behavior: when memories reports were inaccurate, the effect of selection was absent (in PFC and 7a) or slightly inverted (in FEF and V4), suggesting that the animal failed to select an item or selected the wrong item (Fig. S4.7-8).

These results are consistent with previous fMRI and EEG work in humans(Ester et al., 2018; Sprague et al., 2016) and suggest selection and attention use similar mechanisms to enhance memory/sensory representations in prefrontal and parietal cortex (see Fig. S4.5 for enhanced representation of attended stimuli in our task). However, in contrast with attention(Cisek and Kalaska, 2005; Desimone and Duncan, 1995; Reynolds et al., 1999), selection did not reduce the response to the unselected memory in LPFC and parietal cortex (Fig. S4.9; but did slightly decrease the response in FEF), suggesting selection may not engage the competitive mechanisms that suppress unattended stimuli(Reynolds et al., 1999).

4.3.4 Attention and selection prepare representations for read-out

Next, we were interested in how the changing task-demands during the trial affected memory representations. Early in the trial, during the first delay, color memories must be maintained in a form that allows the animal to select the cued item (i.e. colors are bound with location information). Later in the trial, after selection, the same information was used in a different way – to guide the visual search of the color wheel (which results in the anima's decision and eye movement). Given this change in how memory information is used during the trial, we tested whether memory representations were transformed by selection. For these analyses, we focused on neural representations in LPFC because activity in this region encoded both stimuli and was tightly linked with behavior (Fig. 4.3c).

Early in the trial, before selection, the color of each item in memory was represented in separate subspaces in the LPFC neural population. Fig. 4.4a shows the representation of color information about the upper and lower item, before selection (projected into a reduced three dimensional space, see methods). Color information showed a clear organization; the responses to four categories of color were separated and coded in color order for both the upper and lower item (i.e. neighboring colors on the color wheel had neighboring representations; note: the response wheel was rotated on each trial, so this does not reflect motor planning). Color representations for each item were constrained

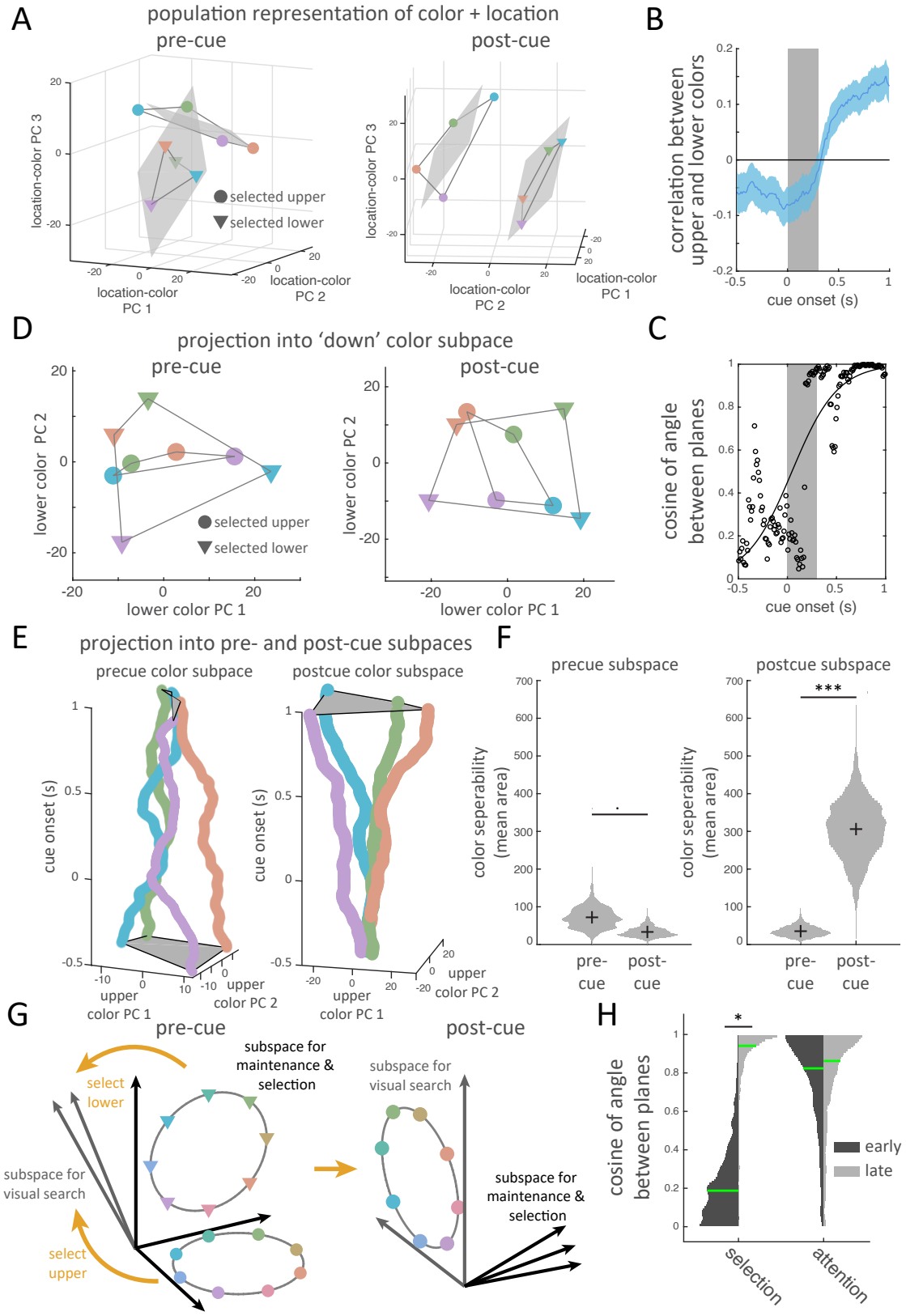


Figure 4.4: (continued on next page)

Figure 4.4: Selection transforms task-relevant information into a common subspace. (A) Population response for selected colors (binned into 4 color bins, indicated by marker color) at different locations (upper vs. lower, indicated by marker shape). Population response is taken as the vector of mean firing rate of all recorded neurons before the cue (pre-cue, left; taken at 400 ms) and after the cue (post-cue, right; taken just prior to target onset, see methods for details). Responses are projected into a reduced dimensionality subspace defined by the first three principle components (PCs) of all 8 color/location pairs. Grey lines connect adjacent colors along the color wheel. Gray shaded region reflects the best fitting planes to each location (see methods). (B) Color representations for upper and lower items become correlated after selection. Line shows the mean correlation between the population representation for each color when it was presented/remembered in the 'upper' or 'lower' position, over time. Correlation was measured after subtracting the mean response at each location (see methods). Error bars reflect standard error of the mean. (C) Color planes (seen in A) become aligned after selection, reflected in an increase in the cosine of the angle between the two color planes around the time of cue onset. Black line shows the best-fitting logistic function. (D) Alignment of color representations before (left) and after (right) selection. Colored markers indicate vector of population firing rate for both upper and lower items (markers as in A). Here, all vectors are projected into the 'lower' subspace, defined by the first two PCs that maximally explain variance in the color of the lower item (defined in the full N-dimensional neural space on held-out data; see methods). Timepoints and markers are as in (A). (E) Timecourse of population responses to the color of the upper item, projected into the upper subspace defined before selection (left) and after selection (right). Upper subspaces were defined as in D, but for the upper item. (F) Before selection, color representations are better separated using the pre-selection subspace. After selection, colors are better separated in the post-selection subspace. Separability was measured as the area of the quadrilateral defined by the population vectors for each color, projected into either the pre-selection or post-selection subspaces (left and right columns in each plot; area averaged across upper and lower items). Subspaces are defined as in D and E. Violin plots show bootstrapped distributions. (G) Schematic of how selection transforms color representations. Initially, the colors of the upper and lower item are encoded in orthogonal subspaces specific to each item's location. The selected item is then transformed into a common subspace, regardless of its initial location. (H) Upper and lower representations become aligned after selection (left column) but immediately after stimulus presentation during attention (right column). Histograms show bootstrapped distribution of the cosine of the angle between the best-fitting planes for the upper and lower stimuli in either an 'early' (150-350 ms post-stimulus offset) or 'late' (200-0 ms before color wheel onset) time period during the delay. Green lines indicate median values. $\cdot$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

to a 'color plane', consistent with a two-dimensional color space (these planes explained $> 97\%$ of variance). As seen in Fig. 4.4a, the upper and lower color planes appeared to be independent from one another, suggesting color information about the upper and lower items were separated into two different subspaces in the LPFC population (before selection).

Consistent with separate subspaces, the color representations of the upper and lower items were anti-correlated before selection (Fig. 4.4b, e.g. the N-neuron population vectors of 'red upper' and 'red lower' were anti-correlated; mean $r = -0.067$ for -300 to 0 ms pre-selection, $p = 0.009$, bootstrap). This weak anti-correlation suggests that tuning curves of neurons were slightly inverted for color across the two spatial locations. Further consistent with separate subspaces, the median angle between the color planes of the upper and lower items was 79.1°(Fig. 4.4c, IQR: 71.4°to 85.1°), suggesting the subspaces were nearly orthogonal. This orthogonality was not because independent

populations of neurons encoded each item: as expected (Rigotti et al., 2013), more LPFC neurons were selective for both items in memory than expected by chance (31% and 35% of neurons selective for the color of upper/lower stimulus were also selective for other item; $p = 1.21e-4$, binomial test). Further reflecting the different upper and lower subspaces, color representations of an item were less separated when they were projected onto the other subspace (e.g. projecting upper colors into the lower color subspace, Fig. 4.4d); each item's color subspace was defined as the 2D space that maximally captured color information in the full N-dimensional neural space. To quantify the separability of colors, we measured the area of the quadrilateral defined by the four color representations - increased separation of colors increases this 'color-area'. Consistent with independent subspaces for upper and lower items, color-area was greater when color representations were projected into their own subspace compared to the other subspace (86.1 vs. 35.2 units^2, $p = 0.041$, bootstrap; all subspaces were defined on held-out data).

After selection, the representation of the selected item was transformed into a different subspace (Fig. 4.4e). Reflecting this transformation, the pre-selection subspace separated colors early in the trial, but this separation collapsed by the end of the second memory delay (Fig. 4.4e, left, and Fig. S4.10). Accordingly, color-area tended to decrease over time from 74.1 to 39.4 units^2 (Fig. 4.4f, left; $p = 0.076$, bootstrap). Instead, after selection, colors were represented in a new 'post-selection' subspace (Fig. 4.4e, right, and Fig. S4.10; color-area increased in the post-selection subspace from 27.8 to 261.9 units^2 across time, Fig. 4.4f, right; $p < 0.001$, bootstrap).

Interestingly, this transformation was such that the post-selection subspaces for the upper and lower items were now aligned (Fig. 4.4A). Consistent with such an alignment, the representation of the color of the selected item shifted from being anti-correlated before selection to positively correlated after selection (Fig. 4.4b, mean $r = 0.139$ for -300 to 0 ms prior to target onset, $p < 0.001$ vs zero and vs pre-cue, bootstrap). Furthermore, the color planes of the upper and lower item shifted from orthogonal to parallel: the angle between the planes was 20.1°after selection (IQR: 11.6°-29.0°). This change was significant over time, as measured by a significant increase in the cosine of the angle between the upper and lower color planes after selection (Fig. 4.4c, $p = 0.006$, bootstrap test of logistic regression). Finally, color representations of an item were now well separated when they were projected onto the other color subspace (Fig. 4.4d; color-area increased from 35.2 to 94.0 units^2 over time, $p = 0.010$, bootstrap). The alignment of color spaces was important for behavior: when memory reports were inaccurate, the increase in cosine of the angle between the two color planes was reduced (Fig. S4.11, $p = 0.0273$, randomization test).

The dynamic re-alignment of neural representations may reflect changing task demands during

the trial. Before selection, coding of the two memories should be independent to allow for selection of a specific color (the independent subspaces of Fig. 4.4a, schematized in Fig. 4.4g, left). After selection, the color of the selected item should move into a shared 'template' subspace that can be used to guide visual search, regardless of whether the upper or lower item was selected (schematized in Fig. 4.4g, right). In this way, location information is abstracted away, as it is no longer relevant.

If the transformation of memories is driven by task demands, then these dynamics should follow a different time course on attention trials. On attention trials, the animals can immediately prepare to search for the cued color after stimulus presentation. Accordingly, the representations of the upper and lower colors were positively correlated immediately after stimulus offset on attention trials (Fig. S4.12). Furthermore, the upper and lower color planes were well-aligned throughout the trial (Fig. 4.4h; early: median angle = 34.5°, IQR = 22.1°to 51.4°; late: median angle = 30.4°, IQR = 18.5°to 46.2°; no change with time, $p = 0.449$, bootstrap; there was a trend towards an interaction between time and attention/selection, $p = 0.067$). These results suggest the transformation of color information is under cognitive control, rather than inherent in the dynamics of the circuit.

Finally, we were interested in whether the same 'template' space was used in the selection and attention tasks. Consistent with a similar template space across tasks, we found a weak, but significant, correlation between color representations at the end of the delay on attention and selection trials (Fig. S4.13, mean $r = 0.06$, $p = 0.015$, bootstrap). This correlation did not exist before selection (mean $r = -0.01$, $p = 0.634$) and increased with time ($p = 0.027$, bootstrap).

4.4 Discussion

Altogether, these results provide novel insight into the mechanisms controlling working memory. Simultaneous recordings from prefrontal, parietal, and visual cortex show that prefrontal cortex directs selection to internal memory representations. Furthermore, selection and attention had overlapping representations in LPFC, with delayed or no generalization in FEF, parietal cortex, and visual cortex. These results suggest lateral prefrontal cortex is a 'domain-general' controller, categorically directing the control of representations regardless of whether they are internal memories or external stimuli. This could be useful for generalizing a task across two cognitive domains, such as sensory processing and working memory (e.g. selecting a dinner special from memory or from a printed list). Conversely, the more differentiated control signals in FEF and parietal cortex may allow for representation-specific control of memories or sensory stimuli.

Selection enhanced working memory representations in prefrontal and parietal cortex, corre-

sponding to an improvement in working memory accuracy. This is similar to attention (Reynolds et al., 1999; Reynolds and Heeger, 2009) and selection of motor actions(Cisek and Kalaska, 2005), suggesting there is a common mechanism for enhancing representations in the brain to overcome interference with competing sensory inputs, motor actions, or memories.

Selection also transformed memory information. Early in the trial, working memory representations were held in location-specific 'upper' and 'lower' spaces, perhaps facilitating the selection of a memory by its associated location. Then, later in the trial, the selected memory shifted into a shared 'template' space, which could be used to guide responses by acting as a template for searching the color wheel (Fig. 4.4a). Interestingly, all three spaces (upper, lower, and template) were approximately orthogonal to one another, which could reduce interference between simultaneously maintained memory representations (i.e. upper and lower) and limit interactions between memory representations and search-related representations. The dynamic transformation of the selected memory from the upper/lower space to the shared template space is reminiscent of the rotation of motor movements from a passive 'null' space to an active 'response' space (Kaufman et al., 2014). Our results build on this work, showing multiple representational spaces can converge onto a single common space (i.e., both lower and upper can transform into the template space). Furthermore, we find these dynamics are under cognitive control and depend on task demands, reflected in the different timecourses of transformation across selection and attention.

More broadly, such dynamic transformations could be a mechanism of cognitive control. Cognitive control is thought to rely on task-specific routing of information (Miller and Cohen, 2001). Previous work has suggested such routing can occur through gain modulation(Miller and Cohen, 2001) or changes in synchrony (Buschman et al., 2012; Fries, 2015; Fries et al., 2001). Our results suggest an alternative mechanism – cognitive control dynamically transforms information in a task-specific manner, allowing information to selectively engage with task-relevant circuits(Stokes et al., 2013). For example, consider a downstream 'visual search' circuit that uses color information from the common template space to guide visual search. Early in the trial, memories are stored in the upper/lower spaces, which are orthogonal to the template space (Fig. 4.4g, left). Thus, colors are not distinguishable to the visual search circuit and so the circuit is not engaged. After selection, memory information is transformed into the shared template space (Fig. 4.4g, right) and the visual search circuit can be engaged. In this way, dynamically transforming representations may allow the brain to control what and when cognitive computations are engaged.

4.5 Methods

4.5.1 Subjects

Two adult male rhesus macaques (Monkey 1 and 2, 12.1 kg and 8.9 kg) performed the experiment. All experimental procedures were approved by the Princeton University Institutional Animal Care and Use Committee and were in accordance with the policies and procedures of the National Institutes of Health.

4.5.2 Behavioral task

Stimuli were presented on a Dell U2413 LCD monitor positioned at a viewing distance of 58 cm. The monitor was calibrated using an X-Rite i1Display Pro colorimeter to ensure accurate color rendering. During the experiment, subjects were asked to remember the color of either 1 or 2 square stimuli presented at two possible locations. The color of each sample was drawn from 64 evenly spaced points along an isoluminant circle in CIELAB color space. This circle was centered at (L = 60, a = 6, b = 14) and the radius was 57 units. The stimuli measured 2°of visual angle (DVA) on each side. Each stimulus could appear at one of two possible spatial locations: 45°clockwise or counterclockwise from the horizontal meridian (in the right hemifield; stimuli are depicted in the left hemifield in Figure 1 for ease of visualization) with an eccentricity of 5 DVA eccentricity from fixation. To perform the selection task, the animal had to remember which color was at each location (i.e., the 'upper' and 'lower' colors).

The animals initiated each trial by fixating a cross at the center of the screen. On selection trials, after 500 ms of fixation, one (20% of trials) or two (80% of trials) stimuli appeared on the screen. The stimuli were displayed for 500 ms, followed by a memory delay of 500 or 1,000 ms. Next, a symbolic cue was presented at fixation for 300 ms. This cue indicated which sample (upper or lower) the animal should report in order to get juice reward. Two sets of cues were used in the experiment to dissociate the meaning of the cue from its physical form. The first set (**cue set 1**) consisted of lines oriented 45°clockwise and counterclockwise from the horizontal meridian (cueing the lower and upper stimulus, respectively). The second set (**cue set 2**) consisted of a triangle or a circle (cueing the lower and upper stimulus, respectively). Cues were presented at fixation and subtended 2 degrees of visual angle. After the cue, there was a second memory delay (500-700 ms), after which a response screen appeared. The response screen consisted of a ring 2°thick with an outer radius of 5°. The animals made their response by breaking fixation and saccading to the section of the

color wheel corresponding to the color of the selected (cued) memory. Importantly, the color ring was randomly rotated on each trial to prevent motor planning or spatial encoding of memories. The animals received a graded juice reward that depended on the accuracy of their response. The number of drops of juice awarded for a response was determined according a circular normal (von mises) distribution centered at 0°error with a standard deviation of 22°. This distribution was scaled to have a peak amplitude of 12, and non-integer values were rounded up. When response error was greater than 60°for Monkey 1 (40°for Monkey 2), no juice was awarded and the animal experienced a short time-out of 1 to 2 s. Responses had to be made within 8 s, although, in practice, this restriction was unnecessary as response times were on the order of 200–300 ms.

Attention trials were similar to selection trials, except that the cue was presented 200-600 ms before the stimuli. After the colored squares, a single continuous delay occurred before the onset of responses screen (1300-2000 ms for Monkey 1 and 1000-2000 ms for Monkey 2). For behavioral analyses and all neural analyses around the response epoch, we only analyzed trials with a minimum of delay of 1300 ms to match the total delay range for attention and selection.

Condition (selection or attention) and cue set were manipulated in a blocked fashion. Animals transitioned among three different block types: (1) attention trials using cue set 1, (2) selection trials using cue set 1, and (3) selection trials using cue set 2. The sequence of blocks was random. Transitions between blocks occurred after the animal had performed 60 correct trials of block type 1 (attention) or 30 correct trials for block types 2 and 3 (selection), balancing the total number of attention and selection trials.

The eye position of the animals was continuously monitored at 1 kHz using an Eyelink 1000 Plus eye-tracking system (SR Research). The animals had to maintain their gaze within a 2°circle around the central cross during the entire trial until the response. If they did not maintain fixation, the trial was aborted and the animal received a brief timeout.

We analyzed all completed trials, defined as any trial on which the animal successfully maintained fixation and made a saccade to the color wheel, regardless of accuracy. Monkey 1 completed 9,865 trials over 10 sessions and Monkey 2 completed 11,131 trials over 13 sessions.

As shown in Fig. S4.1, the behavior of the two animals was qualitatively similar and so we pooled data across animals for all analyses.

4.5.3 Surgical procedures and recordings

Animals were implanted with a titanium headpost to immobilize the head and with two titanium chambers for providing access to the brain. The chambers were positioned using 3D models of the brain and skull obtained from structural MRI scans. Chambers were placed to allow for electrophysiological recording from LPFC, FEF, area 7a/b, and V4.

Epoxy coded tungsten electrodes were used for both recording and microstimulation. Electrodes were lowered using a custom built microdrive assembly that lowered electrodes in pairs from a single screw. Recordings were acute; up to 80 electrodes were lowered through intact dura at the beginning of each recording session and allowed to settle for 2-3 hours before recording. This enabled stable isolation of single units over the session. Broadband activity (sampling frequency = 30 kHz) was recorded from each electrode. We performed 13 recording sessions in Monkey 2 and 10 sessions in Monkey 1.

After recordings were complete, we confirmed electrode locations by performing structural MRIs after lowering two electrodes in each chamber into cortex. Based on the shadow of these two electrodes, the position of the other electrodes in each chamber could be reconstructed. Electrodes were categorized as falling into LPFC, FEF, 7a, and V4 based on anatomical landmarks.

In separate experiments, we identified which electrodes were located in FEF using electrical microstimulation. Based on previous work(Bruce and Goldberg, 1985), we defined FEF sites as those for which electrical stimulation elicited a saccadic eye movement. Electrical stimulation was delivered in 200 ms trains of anodal-leading bi-phasic pulses with a width of 400 s and an interpulse frequency of 330 Hz. Electrical stimulation was delivered to each electrode in the frontal well of each animal and FEF sites were identified as those sites for which electrical stimulation (< 50 μA) consistently evoked a saccade with a stereotyped eye movement vector at least 50% of the time. Untested electrode sites (e.g., from recordings on days with a different offset in the spatial distribution of electrodes) were classified as belonging to FEF if they fell within 1 mm of confirmed stimulation sites and were positioned in the anterior bank of the arcuate sulcus (as confirmed via MRI).

4.5.4 Signal preprocessing

Electrophysiological signals were filtered offline using a 4-pole 300 Hz high-pass Butterworth filter. For Monkey 1, to reduce common noise, the voltage time series x recorded from each electrode was re-referenced to the common median reference(Rolston et al., 2009) by subtracting the median

voltage across all electrodes in the same recording chamber at each time point.

The spike detection threshold for all recordings was set equal to $-4\sigma_n$, where σ_n is an estimate of the standard deviation of the noise distribution:

$$\sigma_n = median(\frac{|x|}{0.6745}) \tag{4.1}$$

Timepoints at which x crossed this threshold with a negative slope were identified as putative spiking events. Repeated threshold crossings within 32 samples (1.0667 ms) were excluded. Waveforms around each putative spike time were extracted and were manually sorted into single units, multi-unit activity, or noise using Plexon Offline Sorter (Plexon, Dallas, Texas).

4.5.5 Statistical procedures

All parametric tests were two-sided. Nonparametric tests were based on resampling trials with replacement or permuting trial labels and were one-sided, unless otherwise indicated.

4.5.6 Mixture modeling of behavioral reports

Behavioral errors on delayed estimation tasks are thought to be due to at least three sources of errors(Bays et al., 2009; Zhang and Luck, 2008): imprecise reports of the cued stimulus, imprecise reports of the uncued stimulus, and random guessing (i.e., from 'forgotten' stimuli). To estimate the contribution of each of these sources of error, we used a three-component mixture model to model behavioral reports (Bays et al., 2009):

$$p(\hat{\theta}) = (1-\gamma-B)\phi_\sigma(\hat{\theta}-\theta) + \gamma\frac{1}{2\pi} + B\frac{1}{m}\phi_\sigma(\hat{\theta}-\theta^*) \tag{4.2}$$

where θ is the color value of the cued stimulus in radians, $\hat{\theta}$ is the reported color value, θ^* is the color value of the uncued stimulus, γ is the proportion of trials on which subjects responded randomly (i.e., probability of guessing, p(Guess)), B is the proportion of trials on which subjects reported the color of the uncued stimulus (i.e., probability of 'swapping', p(Swap)), and ϕ_σ is a von-mises distribution with a mean of zero and a standard deviation σ (inverse precision). Bootstrapped distributions of the maximum likelihood values of the free parameters γ, B, and σ were generated by fitting the mixture model independently to the behavioral data from each session and then resampling the best fitting parameter values with replacement across sessions. In this way, the distribution shows the uncertainty of the mean parameters across sessions.

As noted in the main text, if the animal was able to select an item from memory earlier in the trial, then this reduced the error in the animal's behavioral response (Fig. 4.1d-e). Behavioral modeling showed earlier cues improved the precision of memory reports (Fig. S4.2b, $\beta = 3.95 +/- 1.88$ SEM, $p = 0.012$, bootstrap) but did not significantly change the probability of forgetting (i.e. random responses; Fig. S4.2b, $\beta = 0.03 +/- 0.03$ STE, $p = 0.126$, bootstrap). Furthermore, we found that memory reports were more accurate in the attention condition than in the selection condition (Fig. 4.1b-c). Here, behavioral modeling showed the improvement with attention was due to an increase in the precision of memory reports and a reduction in forgetting (i.e. fewer random reports; Fig. Fig. 4.2b).

4.5.7 Calculation of cue modulation indices

We used a cue modulation index (MI_{cue}) to describe how each neuron's firing rate was modulated by cuing condition ('upper' or 'lower'), defined as:

$$MI_{cue} = \frac{FR_{upper} - FR_{lower}}{FR_{upper} + FR_{lower}} \tag{4.3}$$

where FR_{upper} and FR_{lower} are a neuron's mean firing rate on trials in which the upper or lower stimulus was cued as task relevant, respectively. Modulation indices were either computed using trials pooled across all selection trials (Fig. S4.3a) or calculated separately for each of the three block types (Fig. S4.3b-c, attention with cue set 1, selection with cue set 1, and selection with cue set 2, see above). This analysis included all neurons that were recorded for at least 10 trials per each cued location. The significance of each neuron's modulation index (Fig. S4.3a) was assessed by comparing to a null distribution of values generated by randomly permuting location labels (upper or lower) across trials (1000 iterations). To test if a region had more significant neurons than expected by chance, the percentage of significant neurons was compared to that expected by chance (the alpha level, 5%).

4.5.8 Calculation of cued location

We used linear classifiers to quantify the amount of information about the location of the cued stimulus (upper or lower) in the population of neurons recorded from each brain region (Fig. 4.2c-d). This analysis included all neurons that were recorded during at least 60 trials for each cueing condition (upper or lower) in each block type (attention with cue set 1, selection with cue set 1, and selection with cue set 2, see above). On each of 1000 iterations, 60 trials from each cueing

condition and block type were sampled from each neuron with replacement. The firing rate from those trials, locked to cue onset, was assembled into a pseudo-population by combining neurons across sessions such that pseudo-trials matched both block and cue condition. For each timestep, a logistic regression classifier (as implemented by fitclinear.m in MATLAB) with L2 regularization ($\lambda = \frac{1}{60}$) was trained to predict the cueing condition (upper or lower) using pseudo-population data from one block (e.g., selection with cue set 1) and tested on held out data from another block (e.g., selection with cue set 2). Classification accuracy (proportion of correctly classified trials) was averaged across reciprocal tests (e.g., train on selection with cue set 2, test on selection with cue set 1).

We used a randomization test to test for significant differences in the onset time of above-chance classification across regions (Fig. 4.2e). For each pair of regions, we computed the difference in time of first-significance ($p < 0.05$, using the bootstrap procedure describe above) for each region (the lag). To generate a null distribution of lags, we randomly permuted individual neurons between the two regions and then repeated the above bootstrap procedure to determine the lag in above-chance classification for each permuted dataset. 1000 random permutations were used for each pair of regions. Significance was assessed by computing the proportion of null lags of greater magnitude than the observed lag. Note that this randomization procedure controls for differences in the number of features (neurons) across regions.

To assess the discriminability of the upper and lower attention conditions (Fig. 4.4), we calculated the 10-fold cross validated classification accuracy (averaged across folds). To provide an estimate of variability we repeated this analysis 1,000 times, each time with a different partition of trials into folds.

4.5.9 Quantification of color information

We adapted previous work(Tort et al., 2010) to define a color modulation index (MI_{color}) that describes how each neuron's firing rate was modulated by the colors of the remembered stimuli. After dividing color space into $N = 8$ bins, MI_{color} is defined as:

$$MI_{color} = \frac{\sum_{c=1}^{N} z_c \; log(N z_c \;)}{\log(N)} \tag{4.4}$$

where z_c is a neuron's normalized mean firing rate r_c across trials evoked by colors in the c^{th} bin:

$$z_c = \frac{r_c}{\sum_{c=1}^{N} r_c} \tag{4.5}$$

MI_{color} is a normalized entropy statistic that is 0 if a neuron's mean firing rate is identical across all color bins and 1 if a neuron only fires in response to colors from one bin. To control for differences in average firing rate and number of trials across neurons, we z-scored this metric by subtracting by the mean and dividing by the standard deviation of a null distribution of MI values. To generate this null distribution, the color bin labels were randomly shuffled across trials and the MI statistic was recomputed (1,000 times per neuron).

Z-scored color modulation indices were computed separately for each time point, trial type (attention or selection), and stimulus type (selected/non-selected/attended/non-attended, Fig. 4.3b and Fig. S4.5). This analysis included neurons that were recorded for at least 10 trials in each of these conditions. Selectivity for color was computed without respect to the spatial location of the stimulus (upper or lower). Computing selectivity for colors only presented at a neuron's preferred location did not qualitatively change the results. Z-scored modulation indices were compared to zero or across conditions via t-test (Fig. 4.3b). We corrected for multiple comparisons over time using cluster-correction(Maris and Oostenveld, 2007). Briefly, the significance of contiguous clusters of significant t-tests was computed by comparing their cluster mass (the sum of the t-values) versus that expected by chance (randomization test). Additionally, to summarize changes in selected and non-selected color information after cue onset, we averaged color information for each neuron in two time periods (-300 to 0 ms pre-cue and 200 to 500 ms post-cue) and tested the difference of these values (post-pre) against zero by bootstrapping the mean difference in color information across neurons (Fig. S4.9).

To determine if a neuron displayed significant selectivity for the color at one particular location (upper/lower), we calculated the z-scored information about the cued color at each timepoint over the interval from 0 to 2.5 seconds post-stimulus onset independently for each location. Color selectivity was measured across all conditions, including attention, selection, and single-item trials. As described above, we used a cluster correction to correct for multiple comparisons across time. Neurons with significant color selectivity ($p < 0.05$) at any point during this interval were deemed color selective. Binomial tests compared the proportion of neurons with significant color selectivity for at least one of the two locations to a conservative null proportion of 10% (for two tests with an alpha of 0.05, one test for each location).

To determine if independent populations of LPFC neurons encoded the upper and lower color

during the pre-cue period of selection trials, we counted the number of neurons with significant cluster-corrected selectivity during the 500 ms period before cue onset. Of the 607 neurons entering the analysis, 112 (18.5%) carried information about the upper color and 99 (16.3%) carried information about the lower color. Of these, 35 (5.8%) carried information about both the upper and lower color. A binomial test compared this proportion (5.8%) to that expected by random assignment of top- and bottom-selectivity (i.e., $18.5\% \times 16.3\% \approx 3.0\%$).

To quantify the amount of information each neuron carried about the animal's *reported color*, we followed the same approach as for stimulus color, except that responses were binned by the color reported by the animal rather than by the color of the cued or uncued stimulus (Fig. 4.3c).

To compare the amount of color information in firing rates across the attention and selection conditions (Fig. S4.6), we computed the z-scored color modulation indices as described above for each of the four conditions of interest (selected, non-selected, attended, and non-attended colors). Trial counts were matched across these four conditions to avoid biases in the color information statistic. To assess relative information about cued (selected and attended) and uncued (non-selected and non-attended) color information, we computed the difference in color information between each pair of conditions, for each neuron. The average difference across all neurons was then tested against zero, using the cluster correction described above to correct for multiple comparisons across time(Maris and Oostenveld, 2007).

To compare the amount of color information in firing rates when behavioral performance was relatively accurate or inaccurate (Fig. S4.7-8), we divided selection trials into two groups based on the accuracy of the behavioral report. Trials within each session were split by the median accuracy for that session. Z-scored color modulation indices were computed separately for each split-half of trials (Fig. S4.7). As above, the same number of trials were used for all four conditions (more/less accurate x selected/non-selected). Additionally, to quantify the effect of selection, the difference in color information for selected and unselected colors was computed for each group of trials separately (more or less accurate). This selected-unselected difference was then tested against zero to measure the effect of selection and tested between the two groups of trials to measure the effect of behavioral accuracy. Comparisons were done with a t-test across all neurons and used the cluster correction described above to correct for multiple comparisons across time(Maris and Oostenveld, 2007).

4.5.10 Principal components analysis of color representations

We were interested in understanding the geometry of mnemonic representations of color across the two possible stimulus locations (upper or lower). To explore this, we examined the response of the population of neurons as a function of the color and location of the cued stimulus. The fidelity of these population representations depended on the behavioral performance of the animal. Therefore, for all principle component analyses, we divided trials based on the accuracy of the behavioral report (median split for each session, as above) and separately analyzed trials with lower angular error (higher accuracy, Fig. 4.4) and higher angular error (lower accuracy, Fig. S4.11).

Trials were sorted into $B = 4$ color bins and $L = 2 locations$ (top or bottom), yielding $BxL = M$ (8) total conditions. To visualize these population representations, we projected the population vector of mean firing rates for each of these 8 conditions into a low-dimensional coding subspace (Fig. 4.4a and Fig. S4.12c, similar to previous work (Murray et al., 2017). For each timestep, we defined a population activity matrix $\mathbf{X}$ as an $M \times N$ matrix, where N is the number of neurons:

$$\mathbf{X} = \begin{bmatrix} \mathbf{r}(c_{1,1}) & - \bar{\mathbf{r}} \\ \vdots & \\ \mathbf{r}(c_{B,L}) & - \bar{\mathbf{r}} \end{bmatrix} \tag{4.6}$$

Here, $\mathbf{r}(c_{B,L})$ is the mean population vector (across trials) for the condition corresponding to color bin B and location L and $\bar{\mathbf{r}}$ is the mean population vector across conditions (i.e., the mean of each column is zero).

The principle components of this matrix were identified by decomposing the covariance matrix $\mathbf{C}$ of $\mathbf{X}$ using singular value decomposition (as implemented by pca.m in MATLAB):

$$\mathbf{C} = \mathbf{PDP^T} \tag{4.7}$$

where each column of $\mathbf{P}$ is an eigenvector of $\mathbf{C}$ and $\mathbf{D}$ is a diagonal matrix of corresponding eigenvalues. We constructed a reduced ($K = 3$) dimensional space whose axes correspond to the first K eigenvectors of $\mathbf{C}$ (i.e., columns of $\mathbf{P}$, $\mathbf{P}_K$, assuming eigenvectors are ordered by decreasing explained variance). These first 3 eigenvectors explained an average of 65% of the variance in the mean population response across all examined timepoints. We then projected the population vector for a given condition into this reduced dimensionality space:

$$\mathbf{z}_K \ = \mathbf{P}_K^{\mathbf{T}}(\mathbf{r}(c_{B,L}) \ - \ \bar{\mathbf{r}}) \tag{4.8}$$

where $\mathbf{z}_K$ is the new coordinate along axis K in the reduced dimensionality space.

We observed that, when visualized in the reduced dimensionality space, the population representations for each color bin B within a given location L tended to lie on a plane, referred to as the 'color plane' in the main manuscript (Fig. 4.4a). To identify the best fitting plane, we defined a new population activity matrix $\mathbf{Y}_L$ for each location L with dimensions $B \times K$:

$$\mathbf{Y}_L = \begin{bmatrix} \mathbf{z}(c_{1,L}) \ - \ \bar{\mathbf{z}}_L \\ \vdots \\ \mathbf{z}(c_{B,L}) \ - \ \bar{\mathbf{z}}_L \end{bmatrix} \tag{4.9}$$

where $\mathbf{z}(c_{B,L})$ is the population vector for the condition corresponding to color bin B and location L in the reduced dimensionality space and $\bar{\mathbf{z}}_L$ is the mean population vector across color bins for that location (i.e., the mean of each column is zero). The principle components of this matrix were calculated in the same manner as above and the first two principle components were the vectors that defined the plane-of-best-fit to the points defined by the rows of $\mathbf{Y}_L$.

If the vectors defining the plane-of-best-fit for the upper item are $\mathbf{v_1}$ and $\mathbf{v_2}$ and those for the lower item are $\mathbf{v_3}$ and $\mathbf{v_4}$, then the cosine of the angle between these two color planes can be calculated as:

$$cos(\theta) \ = \ (\mathbf{v_1} \ \times \ \mathbf{v_2}) \bullet (\mathbf{v_3} \ \times \ \mathbf{v_4}) \tag{4.10}$$

For all analyses, population vectors were based on pseudo-populations of neurons combined across sessions. Pseudo-populations were created by matching trials across sessions according to the color and location of the cued stimulus as described above, and following previous work(Rigotti et al., 2013). This analysis only included neurons that were recorded for at least 10 trials for each conjunction of color and location. Confidence intervals for $cos(\theta)$ were calculated using a bootstrapping procedure. On each of 1000 iterations, 10 trials from each of the 8 conditions were sampled from each neuron with replacement. The average firing rates across these sampled trials provided the mean population vector for that condition on that iteration. To assess how $cos(\theta)$ changed around cue onset (Fig. 4.4c and Fig. S4.11), we used a logistic regression model of the form:

$$cos(\theta) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 t))} \tag{4.11}$$

where t is time relative to cue onset. This model was fit to values of $cos(\theta)$ computed at

each timepoint in the interval from 500 ms pre- to 1000 ms post-cue onset on each bootstrap iteration (described above). This yielded a bootstrapped distribution of β_1 estimates which could be compared to zero or across the two groups of trials with more and less accurate behavioral responses (Fig. S4.11).

To define the color subspace in the full neuron-dimensional space, we defined a $(B = 4)xN$ mean population activity matrix for each location L:

$$\mathbf{W}_L = \begin{bmatrix} \mathbf{r}(c_{1,L}) \ - \ \bar{\mathbf{r}}_L \\ \vdots \\ \mathbf{r}(c_{B,L}) \ - \ \bar{\mathbf{r}}_L \end{bmatrix} \tag{4.12}$$

The color subspace was defined as the first two principle components of $\mathbf{W}_L$.

These subspaces were used for two analyses. First, we projected the population vectors of color responses from one item into the color subspace for the other item (Fig. 4.4d). For example, the population vector response to colors of the upper item were projected into the color subspace of the lower item, defined as the first two principal components of $\mathbf{W}_{lower}$, and vice-versa (Fig. 4.4d). Second, by defining the color subspace of each item at different timepoints t_i, we could examine how color representations evolved during the trial (Fig. Fig. 4.4e-f and Fig. S4.10).

Finally, we were interested in quantifying the separability of colors in a given subspace. As seen in Fig. 4.4d-e, the population representation of the four color conditions, projected into the subspace, form the vertices of a quadrilateral with the edges of the quadrilateral connecting adjacent colors on the color wheel (e.g., Fig. 4.4d). To measure separability of the colors, we computed the area of this quadrilateral (polyarea.m function in MATLAB). Bootstrapped distributions of these area estimates were obtained by resampling trials with replacement from each condition before re-computing $\mathbf{W}_L$.

4.5.11 Correlation of color representations

We wanted to understand how similarly color was represented across the upper and lower locations over the course of the trial. To explore this, selection or attention trials were binned based on the color and location of the cued stimulus and then randomly partitioned into two halves. These split halves were used to estimate the degree of noise in the data (Fig. S4.12, described below). Specifically, trials were sorted into $B = 4$ color bins, $L = 2$ locations (top or bottom), and $H = 2$ halves, yielding $B \times L \times H = M$ total conditions. For each of these conditions, at a given timepoint of interest, we computed the average population vector $\mathbf{r}(c_{B,L,H})$.

We then computed the average correlation between each population vector and the population vectors corresponding to the same color bin at the other location (Fig. 4.4b and Fig. S4.12)

$$\rho_{cross} = \frac{1}{B2H} \sum_{i=1}^{H} \sum_{j=1}^{H} \sum_{b=1}^{B} corr(\mathbf{r}(c_{b,1,i}) - \langle \mathbf{r}(c_{B,1,i}) \rangle_{\{B\}}, \mathbf{r}(c_{b,2,j}) - \langle \mathbf{r}(c_{B,2,j}) \rangle_{\{B\}}) \tag{4.13}$$

where $\langle \cdot \rangle_{\{B\}}$ is the average across the set of color bins B. In other words, for each set of B population vectors corresponding to a particular half of the data H and location L, we subtracted the mean across bins to center the vector endpoints around zero. Thus, ρ_{cross} quantifies to what extent color representations are similarly organized around their mean across the two locations.

To obtain an upper bound on potential values of ρ_{cross} given the degree of noise in the data, we also computed the average correlation of each population vector with itself across the two halves:

$$\rho_{self} = \frac{1}{BL} \sum_{b=1}^{B} \sum_{l=1}^{L} corr(\mathbf{r}(c_{b,l,1}) - \langle \mathbf{r}(c_{B,l,1}) \rangle_{\{B\}}, \mathbf{r}(c_{b,l,2}) - \langle \mathbf{r}(c_{B,l,2}) \rangle_{\{B\}}) \tag{4.14}$$

Finally, to understand how similarly color was represented across the two cueing conditions, trials were sorted into $B = 4$ color bins, $L = 2$ locations (top or bottom), and $C = 2$ cuing conditions (attention/selection). For each of these conditions, at a given timepoint of interest, we computed the average population vector $\mathbf{r}(c_{B,L,C})$. We then computed the average correlation between each population vector and the population vectors corresponding to the same color bin at either the same or different location in the other task (Fig. S4.13):

$$\rho_{att,sel} = \frac{1}{B2L} \sum_{i=1}^{L} \sum_{j=1}^{L} \sum_{b=1}^{B} corr(\mathbf{r}(c_{b,i,1}) - \langle \mathbf{r}(c_{B,i,1}) \rangle_{\{B\}}, \mathbf{r}(c_{b,j,2}) - \langle \mathbf{r}(c_{B,j,2}) \rangle_{\{B\}}) \tag{4.15}$$

To compare the similarity of color representations on selection trials to pre-target attention color representations, we computed this correlation between (1) the response on attention trials, for all timepoints falling within the interval from -300 ms to 0 ms before the onset of the response wheel, and (2) the response on selection trials at two different timepoints: before selection (from -300 to 0 ms before the cue) and after selection (from -300 to 0 ms before the onset of the response wheel). Correlation was measured between each timepoint across windows and then averaged across all pairs of timepoints.

As above, population vectors were pseudo-populations of neurons combined across sessions, where

trials across sessions were matched according to color bin and location(Rigotti et al., 2013). This analysis only included neurons that were recorded for at least 10 trials for each conjunction of color and location. Confidence intervals for ρ_{cross}, ρ_{self}, and $\rho_{att,sel}$ were calculated with a bootstrap. On each of 1000 iterations, and for each neuron and condition (color-location-half conjunction), the entire population of trials in that condition was resampled with replacement. The average firing rates across these sampled trials provided the mean population vector for that condition on that iteration. As with principle components analyses, we divided trials based on the accuracy of the behavioral report (median split of trials for each session) and the presented results reflect analysis of trials with lower angular error, unless otherwise noted.

4.6 Acknowledgments

The authors thank Britney Morea and Hannah Weinberg-Wolf for assistance with monkeys, Sina Tafazoli for assistance with microstimulation, and Flora Bouchacourt, Caroline Jahn, Alex Libby, Camden MacDowell, Sina Tafazoli, Motoaki Uchimura, and Sarah Henrickson for their feedback. We also thank the Princeton Laboratory Animal Resources staff for their support. This work was supported by NIMH R01MH115042 (TJB) and an NDSEG Fellowship (MFP).

4.7 Collaborators

The work described in this chapter was conducted in collaboration with Tim Buschman. At the time of this writing, these results have been published on bioRxiv (Panichello and Buschman, 2020).

4.8 Supplementary figures

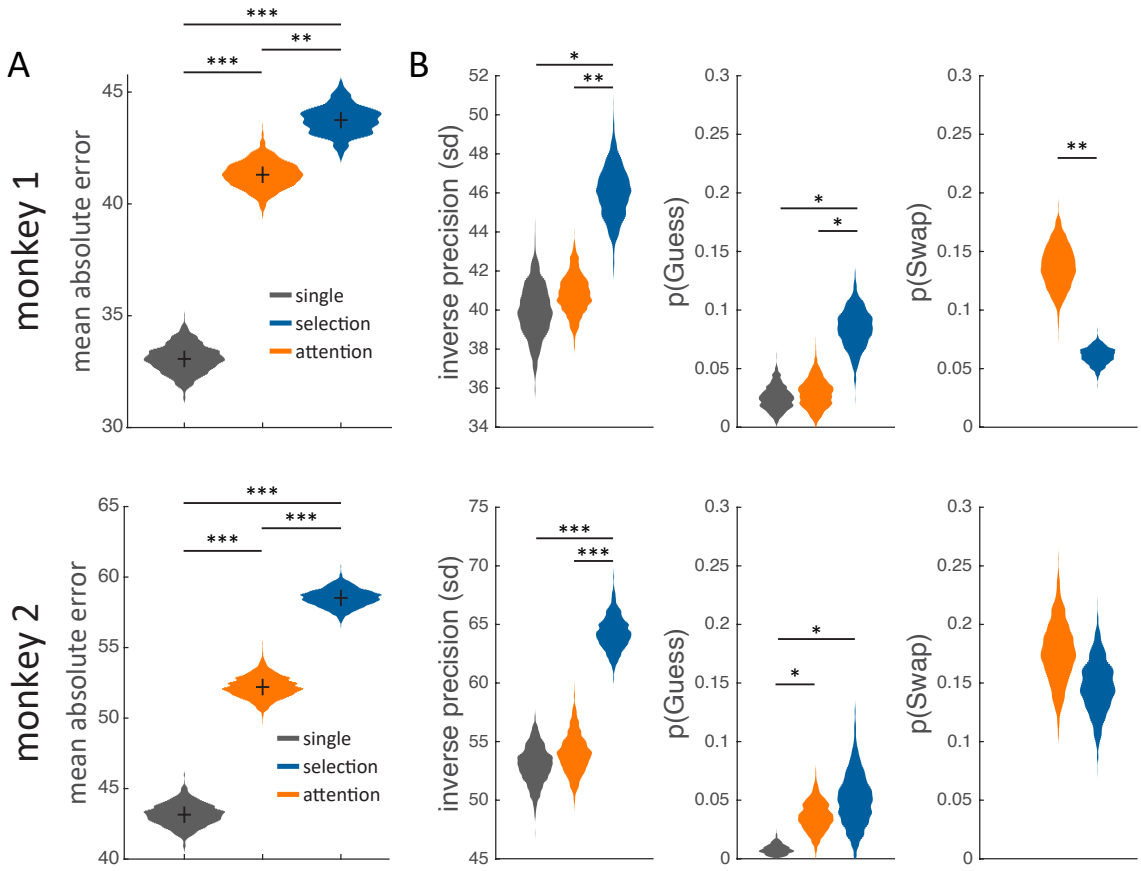


Figure S4.1: (A) Mean absolute angular error and (B) mean mixture model parameter fits for each animal. Violin plots depict bootstrapped distribution. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Although monkey 1 displayed slightly better performance than monkey 2, the animals displayed similar patterns of performance across conditions.

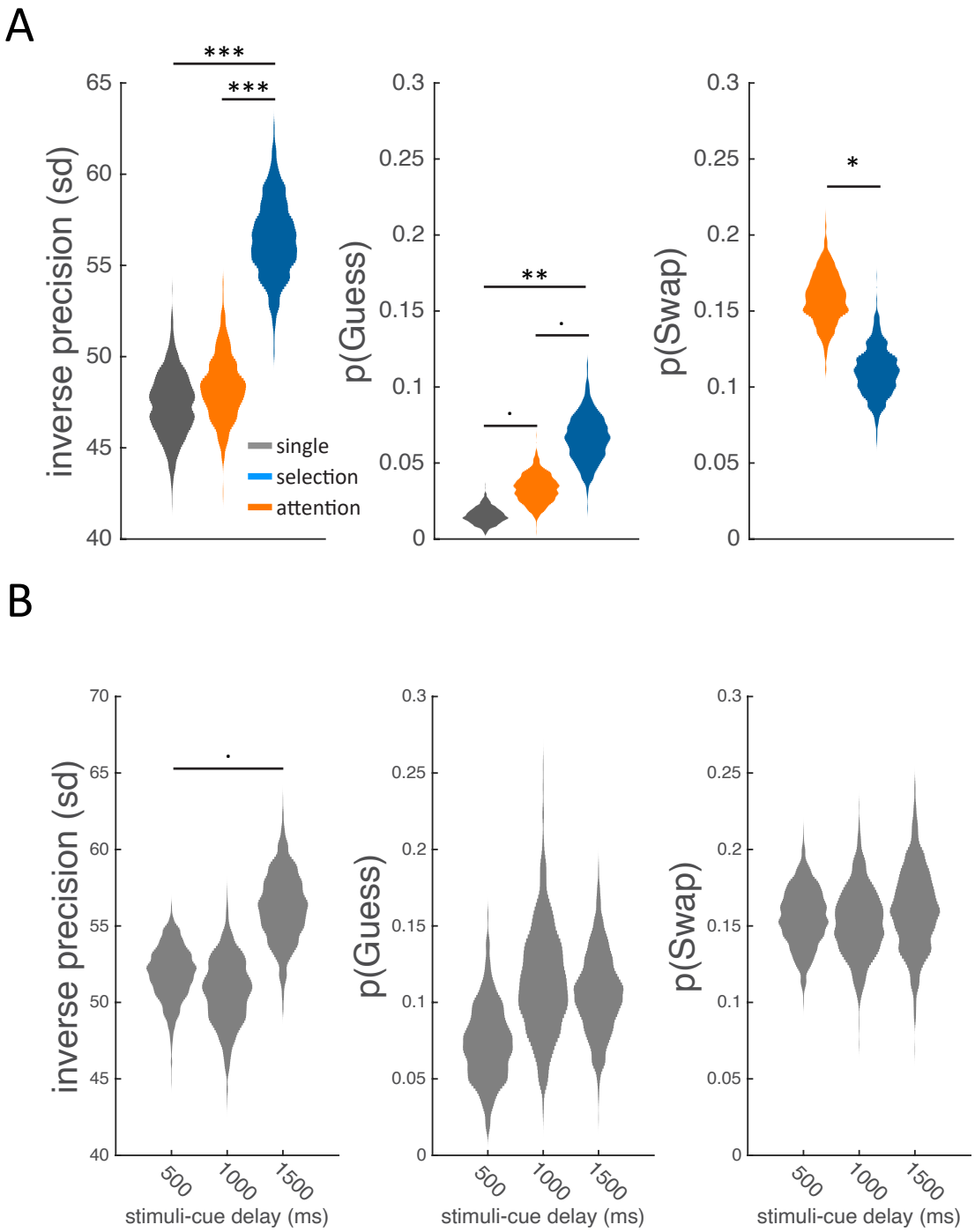


Figure S4.2: Mixture model parameter fits of behavior pooled across animals for the (A) main task shown in Fig. 4.1a and (B) the retro-cue timing manipulation shown in Fig. 4.1d. Earlier cues improved the precision of memory reports ($\beta = 3.95 = / - 1.88$ SEM, $p = 0.012$, bootstrap) but did not significantly change the probability of forgetting (i.e. random responses; $\beta = 0.03 + / - 0.03$ STE, $p = 0.126$, bootstrap). $\cdot$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

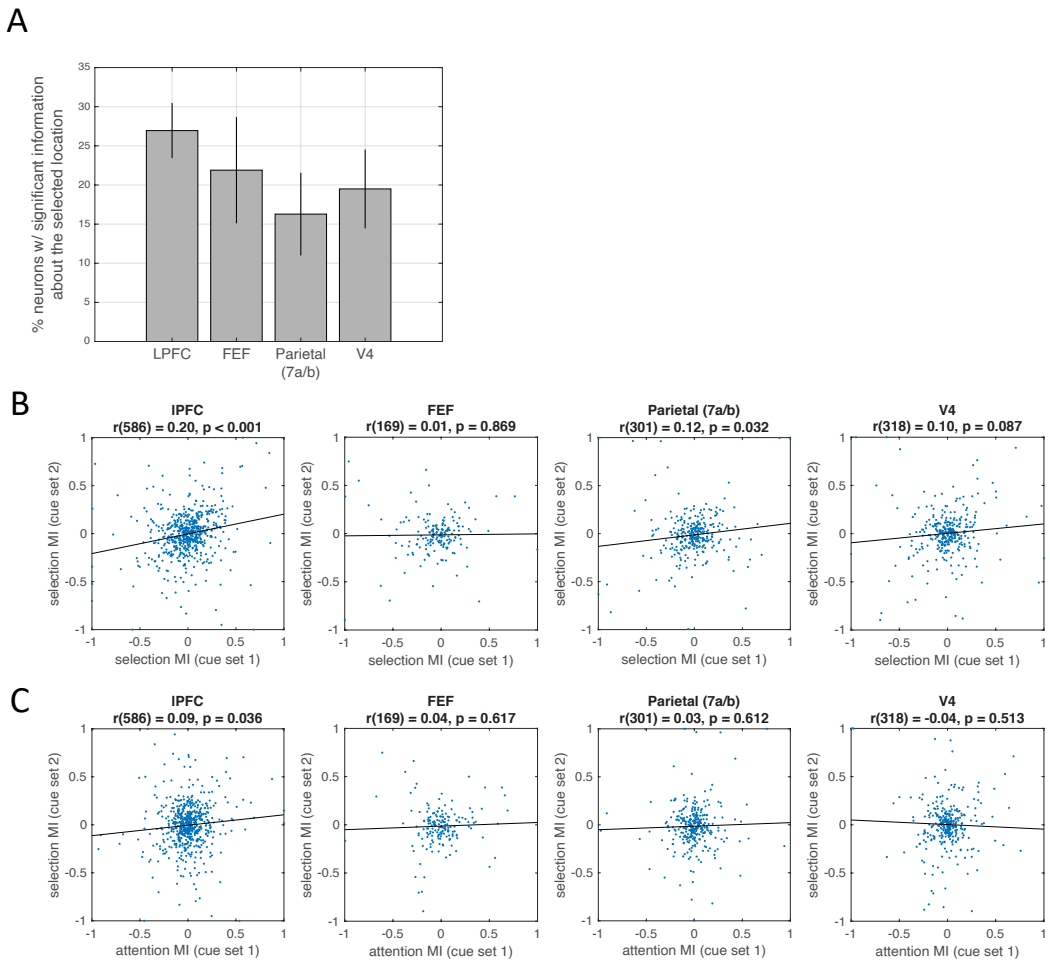


Figure S4.3: (A) Percent of neurons in each region of interest. with firing rates that were significantly modulated by the selected location after cue onset (trials pooled across cue set 1 and 2). For each neuron, we quantified location selectivity using a modulation index (see methods) and compared this value to a null distribution by permuting location labels across trials. All four regions showed strong selectivity: LPFC had 159 out of 590 neurons selective; FEF: 37/169, 7a/b: 49/301, V4: 62/318, all $p < 0.001$ for binomial test against chance of 5%. (B) Correlation of modulation indices for selection cue set 1 and 2. Positive modulation indices indicate that a neuron had a higher firing rate in response to the "upper" cue within a particular cue set, and negative modulation indices indicate that a neuron had a higher firing rate in response to the "lower" cue within a particular cue set. Positive correlations indicate that neurons responded to the selected location in a consistent fashion across cue sets. (C) Correlation of modulation indices for selection cue set 2 and attention cue set 1. Positive correlations indicate that neurons coded for the selected location in a consistent fashion across cue sets and conditions (attention/selection).

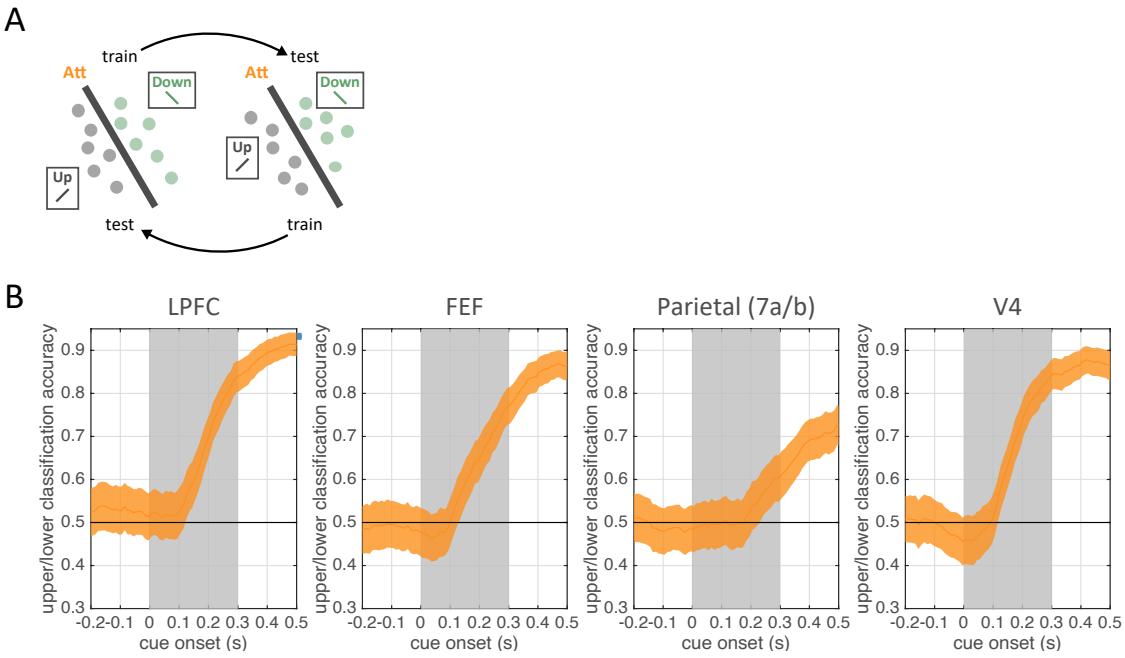


Figure S4.4: (A) To confirm that the neural responses to the two cue conditions were discriminable on attention trials, we calculated the cross-validated classification accuracy (10-fold, one example fold shown for clarity). (B) Mean classification accuracy for each brain region, relative to cue onset. Error bars reflect the standard deviation across 1,000 iterations, each with a different partition of trials into folds. Note that this analysis captures a mixture of information about the control of attention (up or down) and information about the visual appearance of the cue itself (a line oriented +/- 45 degrees). Importantly, these results show these two conditions are separable in all brain regions, and so failures in cross-classification performance (Fig. 4.2d, purple traces) are not due to poor separability of these two conditions.

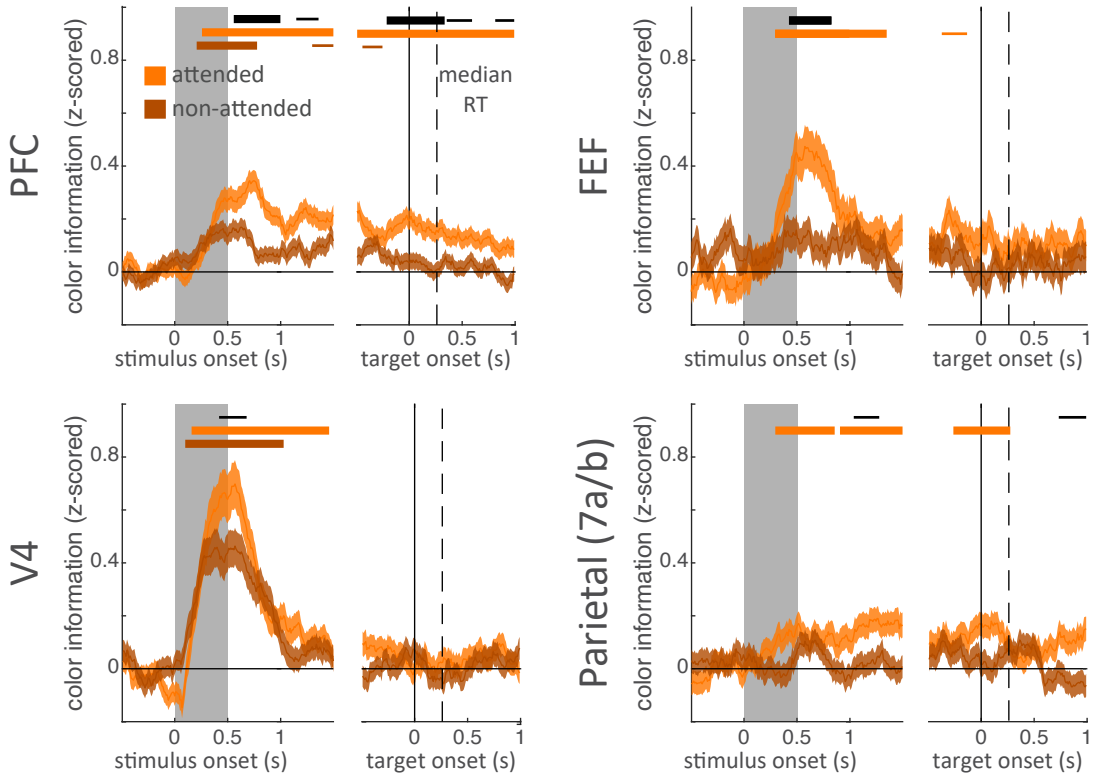


Figure S4.5: Mean z-scored color information for the attended and non-attended color on attention trials. Error bars are standard error of the mean. Horizontal bars indicate significant information for the attended item (light orange), the non-attended item (dark orange), and significant differences in information about the attended and non-attended items (black). Bar width indicate significance: $p < 0.05$, 0.01, and 0.001 for thin, medium, and thick, respectively.

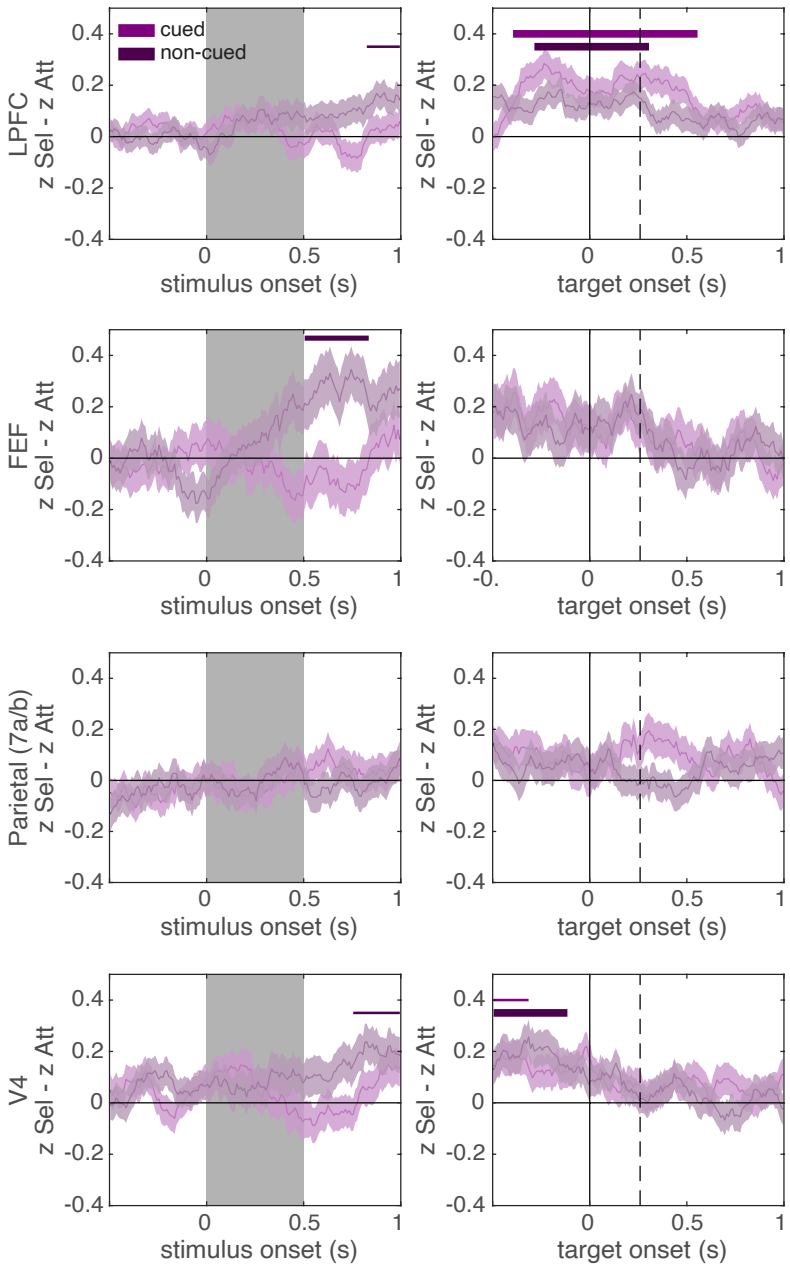


Figure S4.6: Difference in z-scored color information between selection and attention trials for the cued item (selected/attended; light purple) and uncued item (non-selected/non-attended; dark purple). Positive values indicate there was more information about an item on selection trials. Error bars are standard error of the mean, created by bootstrapping across cells (see methods for details). Horizontal bars indicate significant differences from zero (i.e. differences between selection and attention) for the cued item (light purple) and the non-cued item (dark purple). Bar width indicate significance: $p < 0.05$, 0.01, and 0.001 for thin, medium, and thick, respectively.

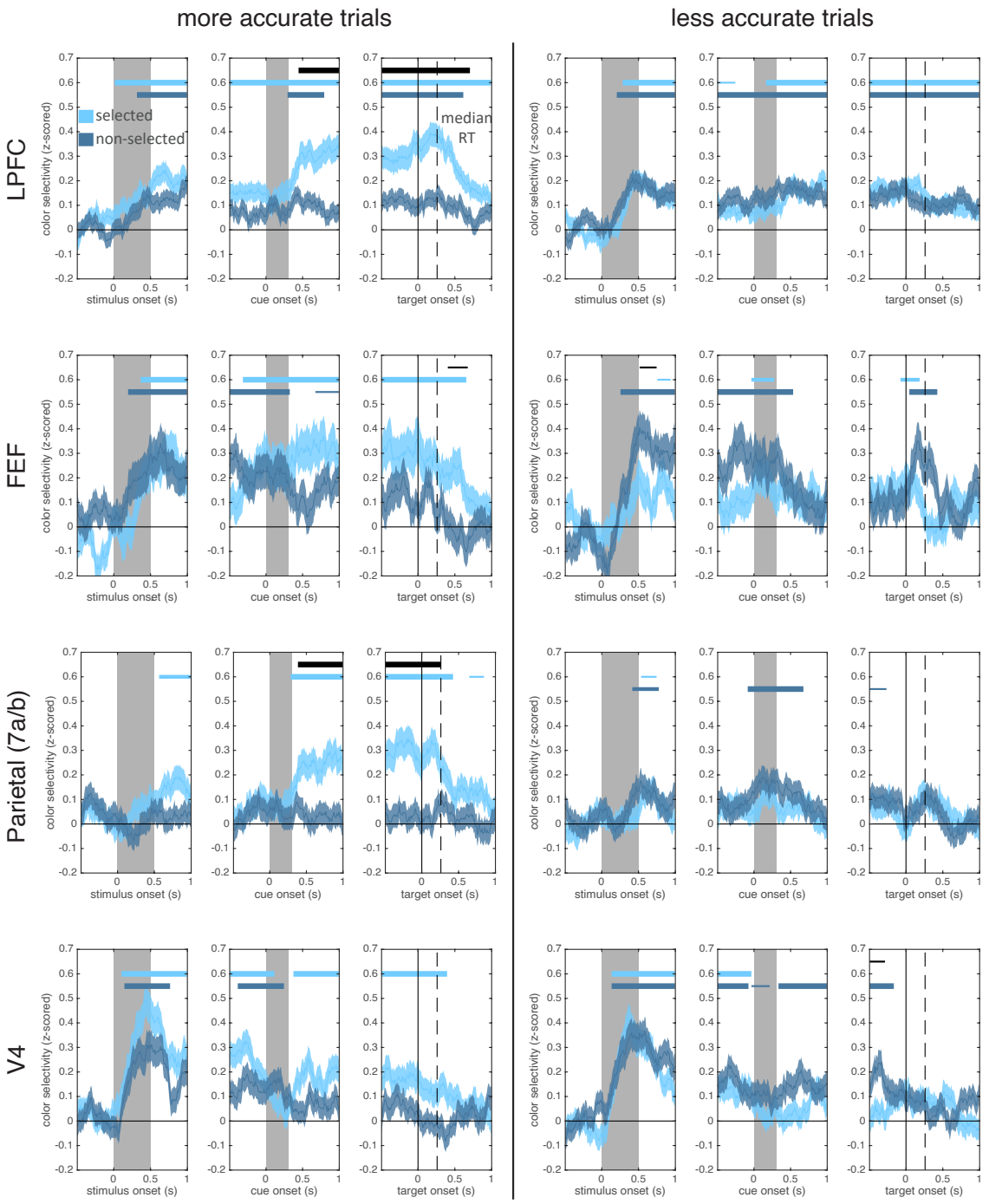


Figure S4.7: Mean z-scored color information for the selected (light blue) and non-selected color (dark blue), separated into trials with more accurate behavioral responses (left column; error was less than median error) and less accurate behavioral responses (right column; error was greater than median error). Plots follow Fig. 4.4b. Horizontal bars indicate significant information for the selected item (light blue), the non-selected item (dark blue), and significant differences in information about the selected and non-selected items (black). Bar width indicate significance: $p < 0.05$, 0.01, and 0.001 for thin, medium, and thick, respectively.

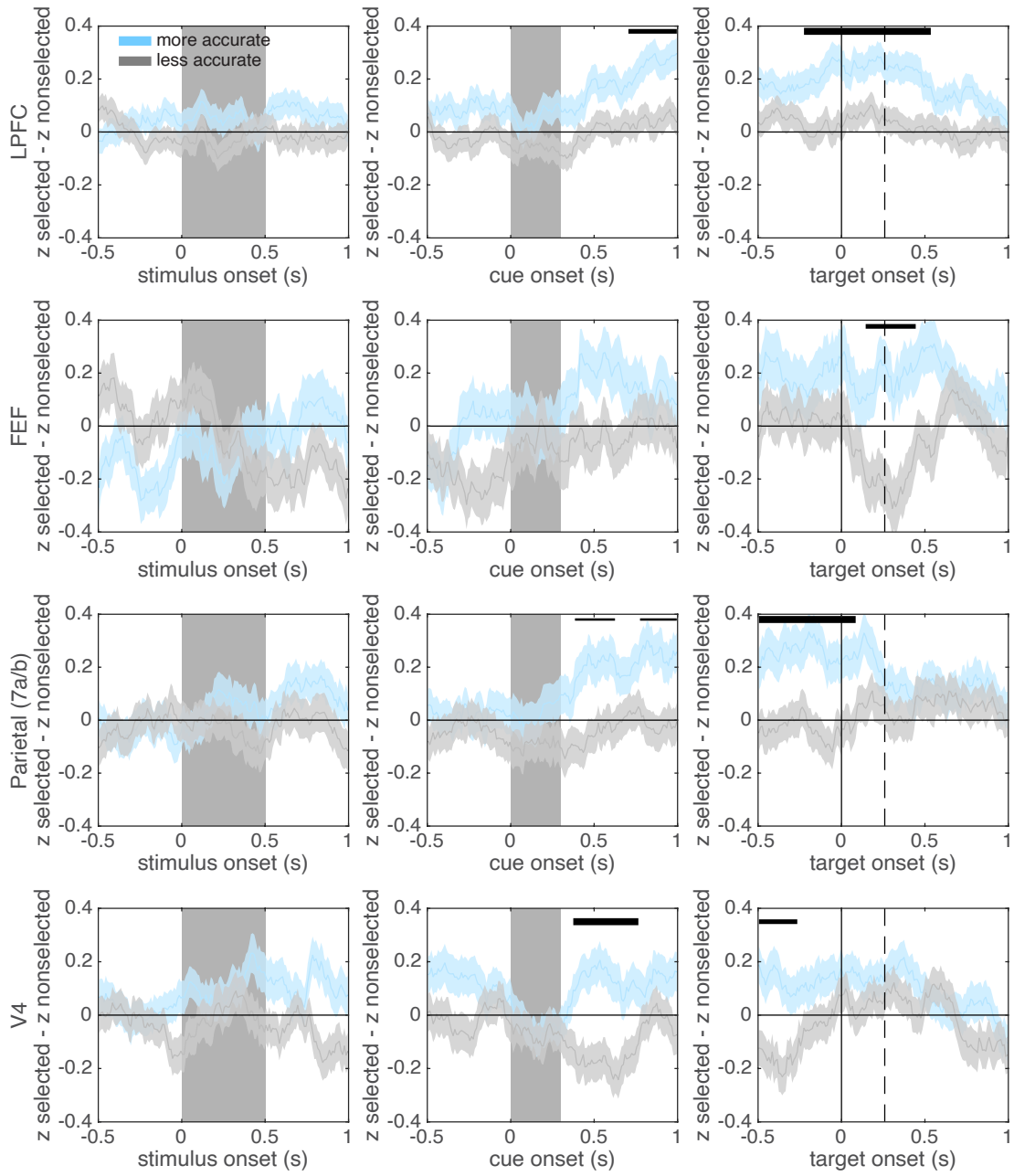


Figure S4.8: Difference in z-scored color information between the selected and non-selected item for more accurate and less accurate trials. As in Fig. S4.7, trials were split based on angular error (relative to median error). Positive values indicate there was more information about the selected item than the non-selected item. Error bars indicate standard error of the mean. Horizontal bars indicate significant differences between more and less accurate trials. Bar width indicate significance: $p < 0.05$, 0.01, and 0.001 for thin, medium, and thick, respectively.

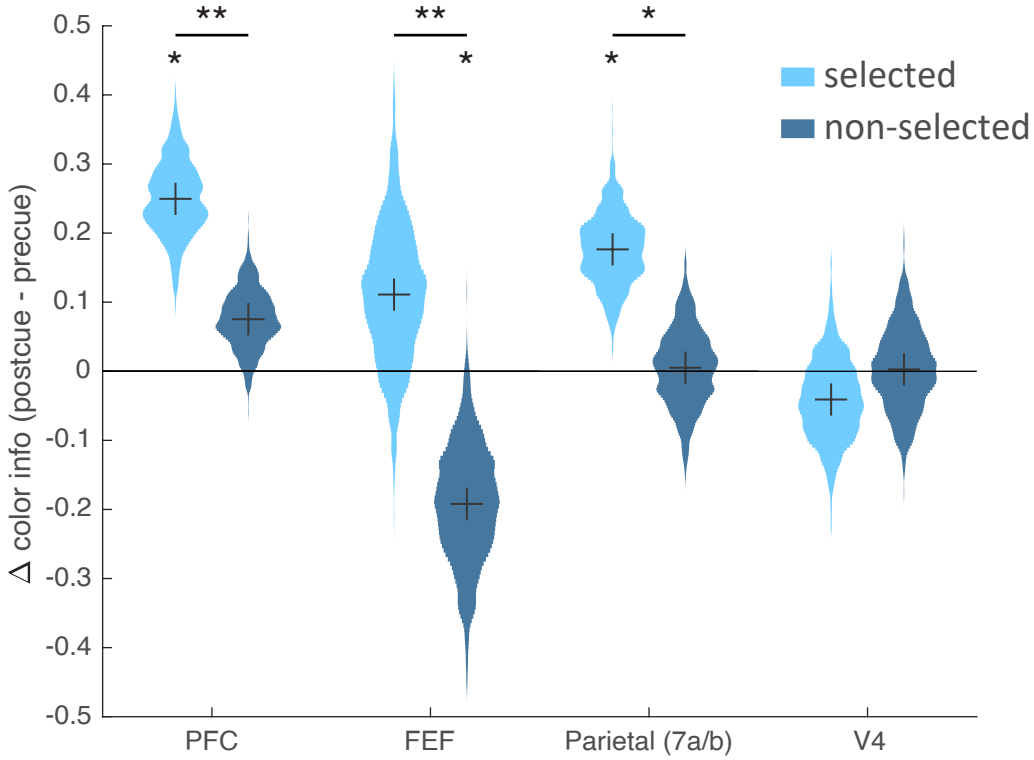


Figure S4.9: Selection enhanced the representation of the selected item in frontal and parietal regions and reduced the representation of the un-selected item in FEF. Y-axis shows the increase in color information after selection (post-cue period: 200 to 500 ms after cue offset), relative to information before selection (pre-cue period: -300 to 0 ms before cue onset). Violin plots show the distribution of this difference (bootstrapped across neurons). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

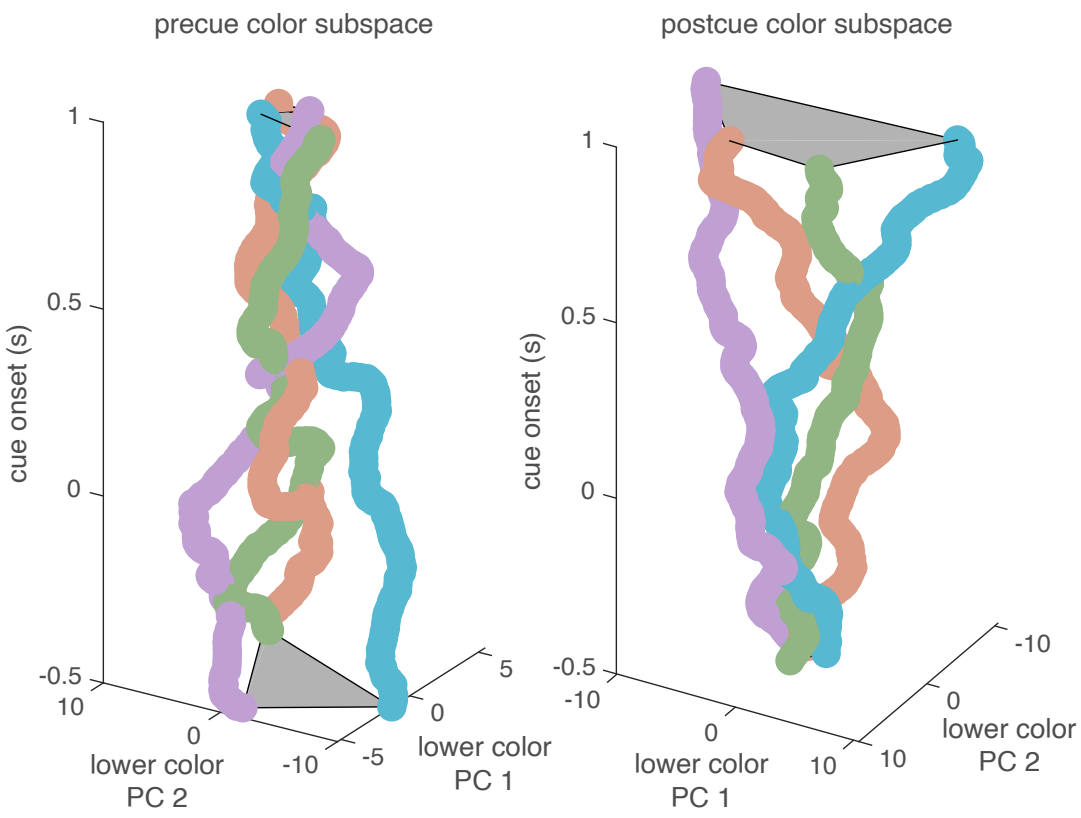


Figure S4.10: Population trajectories for 'lower' colors, over time, as projected into the 'lower' color subspace. Lower color subspace was defined as a 2D space that maximally explained variance across the four 'lower' colors (see methods for details). Subspace was defined either before or after selection, as in Figure 4E. As for the 'upper' color (Fig. 4.4e), temporal cross generalization was poor, suggesting the color information was represented in a different subspace by the end of the trial.

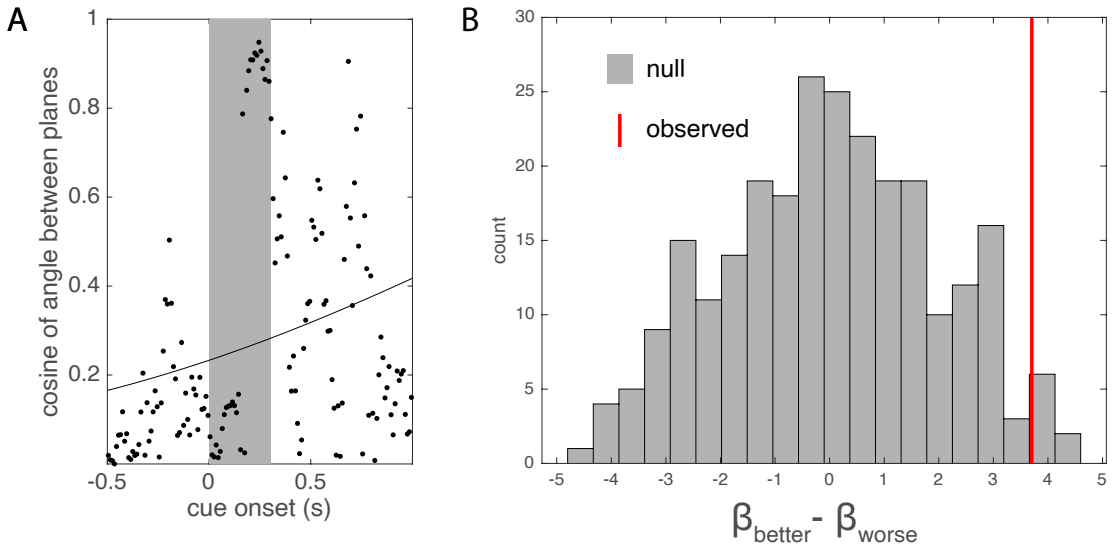


Figure S4.11: (A) The upper and lower color planes do not align on inaccurate trials. Figure follows Fig. 4.4c, but shows data for trials in which absolute angular error was greater than the median error. Black markers show the cosine of the angle between the two color planes around the time of cue onset. Black line shows best-fitting logistic function. (B) Difference in the slope of the logistic fit (i.e., the coefficient for time) between trials in which angular error was 'better' or 'worse' than the median (red line). Null distribution (gray histogram) was generated by randomly permuting the accuracy label ('better' or 'worse') between population vectors for the same color-location conjunction. Trials in which the animal was more accurate were associated with a greater increase in the cosine of the plane angle (i.e. a greater increase in alignment) around cue onset.

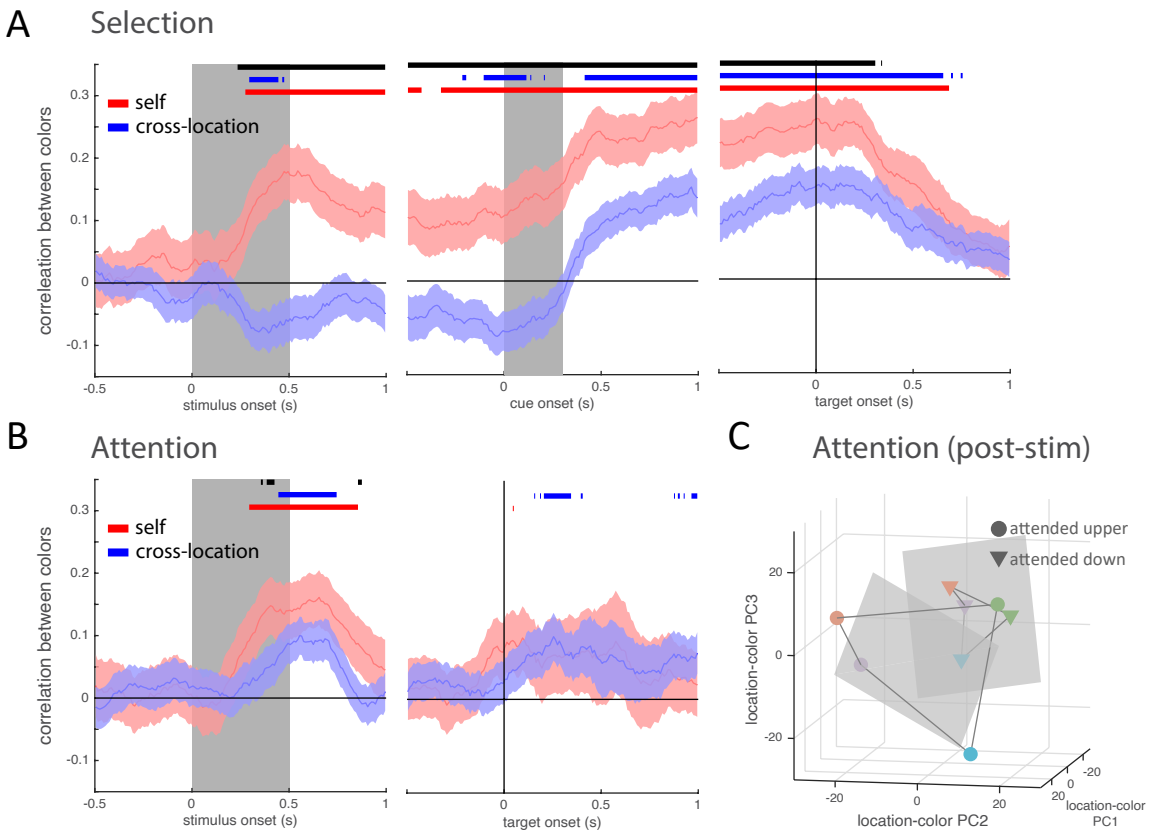


Figure S4.12: (A) Correlation of population vectors representing colors at the same location (self; red line) or between locations (cross-location; blue line) on selection trials. Correlations were measured after subtracting the mean vector at each location (as in Fig. 4.4b). Self-correlation was computed on held-out data and provides an upper-bound on the between-location correlation values, given the noise level. Bars reflect uncorrected t-tests ($p < 0.05$) for each correlation type vs zero (red and blue) and between each other (black). (B) Same as in A, but for attention trials. (C) Population responses 200 ms after stimulus offset on attention trials (projected into a reduced dimensional space for visualization). Markers indicate mean position of population activity for each condition (binned by the color and location of the attended item) in a subspace spanned by the first three principle components that explain the most variance across all 8 conditions.

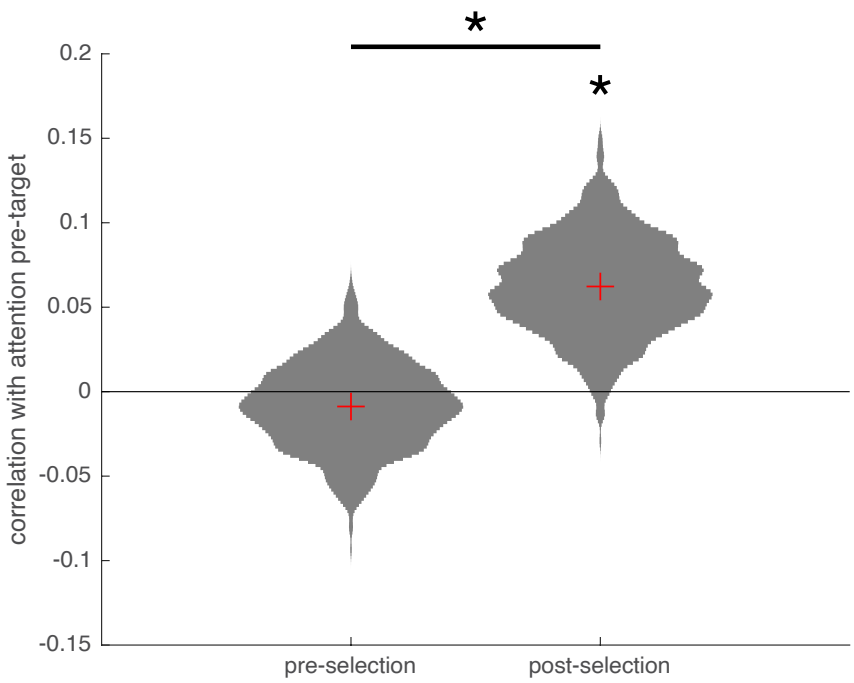


Figure S4.13: Mean correlation between the population representations for each color across attention and selection tasks (after subtracting the mean vector at each location, as in Fig. 4.4b and Fig. S4.12, see methods for details). Violin plots reflect the bootstrapped estimate of the distribution about the mean. The mean correlation was computed between the color representations taken from the 300 ms before the onset of the response wheel on attention trials and the color representations taken from either a pre-selection period (left distribution; -300 to 0 ms pre-cue) or post-selection period (right distribution; -300 to 0 ms before response wheel onset).

Chapter 5

General Discussion

The work in this thesis examines how two cognitive processes, expectation and attention, transform perception, and if these processes similarly impact memory. We found that expectations transform perception in a manner consistent with Bayesian inference; perceptual representations, as inferred via behavioral reports and neural decoding, reflected a weighted average of sensory and prior information. Behavioral analyses revealed that visual memory is governed by similar principles, with the relative weight of prior information increasing with the amount of time that a representations is held in memory. These results suggest that a process akin to Bayesian inference is applied continuously in time to compensate for the accumulation of noise in perception and memory. We found that these dynamics were well described by a model in which memories drift towards attractor states reflecting expected stimuli, suggesting neural architectures that could support this process (Brody et al., 2003; Chaudhuri and Fiete, 2016).

In the realm of attention, we found a surprising correspondence between perception and memory. In both cases, attention transformed neural representations such that attended items were coded using completely distinct patterns of neural activity (i.e., attended representations were coded in a distinct subspace). These results suggest that transformations in state space provide a means by which representations can be 'gated' from a passive state to an active state in which they can affect cognitive or motor preparatory activity, in contrast to accounts postulating gating via toggling of cortical excitability (e.g., Cohen et al., 1990; Hazy et al., 2007).

Overall, these studies suggest that while we can expect cognitive transforms observed in visual perception to generalize to other domains, we should not expect these computations to be implemented by monolithic mechanisms. For example, we observed that recently changed environmental

statistics were manifest in memory but not perception (Fig. 3.7f), suggesting that expectations affect perception and memory via unique mechanisms that adapt over different timescales. And, unlike perceptual attention, mnemonic attention does not bias competitive dynamics (Fig. 4.3b). Thus, mechanisms underlying canonical computations will have to be interrogated in a case-by-case basis across perception, memory, and other domains such as planning, long-term memory, and motor control.

These studies suggest some immediate lines of follow up. First, how is the accumulation of expectation across perception and memory reflected in neural systems? The electrophysiological recordings described in Chapter 4 can be used to address this question by quantifying the extent to which neural activity tracks the latent color memories inferred from the dynamical model. For example, activity in V4 may track the accumulation of biases during perception and activity PFC may track the accumulation of biases in memory (despite the fact that both regions carry some information about the stimulus throughout the entire trial).

Additionally, future work should further specify the extent to which expectations bias perception and memory in an optimal manner. While the dynamical model described in Chapter 3 quantifies the strength of expectations during perception and memory, a precise mapping between this model and classic Bayesian decision models (e.g., Ma, 2019) is unclear. Specifying this mapping would reveal whether expectations during perception and/or memory are weighted in an optimal manner and would facilitate comparison with the rest of the literature.

Recent work has revealed that attention fluctuates rhythmically during perception (Buschman and Miller, 2009; Fiebelkorn et al., 2018; Fiebelkorn et al., 2013). It is unclear whether attention is rhythmically allocated to items in memory in a similar manner. This could be tested using the electrophysiological recordings described in Chapter 4 by quantifying the amount of information about each memory item over time on a trial-by-trial basis and looking for significant peaks in the Fourier transform of these traces.

Finally, at the time of this writing there an active debate about the role of prefrontal cortex in working memory (Christophel et al., 2017). Prominent competing accounts include: (1) prefrontal cortex does not carry information about the content of working memory (2) prefrontal cortex maintains categorical information in working memory (3) prefrontal cortex maintains categorical as well as fine sensory information in working memory. The results presented in chapter 4 are inconsistent with the first account; prefrontal cortex clearly carried information about the remembered color. However, it is unclear from these data whether PFC carries fine sensory information because biases in perception and memory result in partially discretized color representations (Chapter 3). Indeed,

the pervasive presence of biases across different perceptual spaces (Wei and Stocker, 2017) suggests that partial categorical encoding of fine sensory information may be pervasive. The fine sensory / categorical distinction should itself be understood in continuous rather than dichotomous terms. Nevertheless, a new experiment could address this question by presenting a memory stimulus and retrospectively cuing the subject to render a relatively categorical or fine sensory judgment. If prefrontal cortex is important for the maintenance of categorical but not fine sensory information, one might expect a double dissociation where PFC carries more memory information in the categorical condition and sensory regions more memory information in the fine sensory condition.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Akrami, A., Kopec, C. D., Diamond, M. E., & Brody, C. D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour [Number: 7692 Publisher: Nature Publishing Group]. *Nature*, *554*(7692), 368–372. https://doi.org/10.1038/nature25510

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology: CB*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Almeida, R., Barbosa, J., & Compte, A. (2015). Neural circuit basis of visuo-spatial working memory precision: A computational and behavioral study. *Journal of Neurophysiology*, *114*(3), 1806–1818. https://doi.org/10.1152/jn.00362.2015

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations [Publisher: National Academy of Sciences Section: PNAS Plus]. *Proceedings of the National Academy of Sciences*, *113*(4), E420–E429. https://doi.org/10.1073/pnas.1518931113

Baddeley, A. (2003). Working memory: Looking back and looking forward [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744. Retrieved June 5, 2017, from http://psycnet.apa.org/journals/xge/144/4/744/

Bae, G.-Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4), 7–7. Retrieved June 5, 2017, from http://jov.arvojournals.org/article.aspx?articleid=2121591

Ban, H., Preston, T. J., Meeson, A., & Welchman, A. E. (2012). The integration of motion and disparity cues to depth in dorsal visual cortex. *Nature neuroscience*, *15*(4), 636–643. https://doi.org/10.1038/nn.3046

Bar, M. (2004). Visual objects in context. *Nature reviews. Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438. https://doi.org/10.1016/j.tics.2015.06.004

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7. https://doi.org/10.1167/9.10.7

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), 6–6. https://doi.org/10.1167/11.10.6

Bays, P. M., & Taylor, R. (2018). A neural model of retrospective attention in visual working memory. *Cognitive Psychology*, *100*, 43–52. https://doi.org/10.1016/j.cogpsych.2017.12.001

Biederman, I. (1972). Perceiving Real-World Scenes [Publisher: American Association for the Advancement of Science Section: Reports]. *Science*, *177*(4043), 77–80. https://doi.org/10.1126/science.177.4043.77

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Bosman, C. A., Schoffelen, J.-M., Brunet, N., Oostenveld, R., Bastos, A. M., Womelsdorf, T., Rubehn, B., Stieglitz, T., De Weerd, P., & Fries, P. (2012). Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron*, *75*(5), 875–888. https://doi.org/10.1016/j.neuron.2012.06.037

Bouchacourt, F., & Buschman, T. J. (2019). A Flexible Model of Working Memory. *Neuron*, *103*(1), 147–160.e8. https://doi.org/10.1016/j.neuron.2019.04.020

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations [Publisher: The Association for Research in Vision and Ophthalmology]. *Journal of Vision*, *15*(15), 6–6. https://doi.org/10.1167/15.15.6

Brandman, T., & Peelen, M. V. (2017). Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding [Publisher: Society for Neuroscience Section: Research Articles]. *Journal of Neuroscience*, *37*(32), 7700–7710. https://doi.org/10.1523/JNEUROSCI.0582-17.2017

Brody, C. D., Romo, R., & Kepecs, A. (2003). Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations. *Current Opinion in Neurobiology*, *13*(2), 204–211.

Bruce, C. J., & Goldberg, M. E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. *Journal of Neurophysiology*, *53*(3), 603–635. https://doi.org/10.1152/jn.1985.53.3.603

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, *109*(43), 17645–17650. https://doi.org/10.1073/pnas.1117386109

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. *Neuron*, *76*(4), 838–846. https://doi.org/10.1016/j.neuron.2012.09.029

Buschman, T. J., & Kastner, S. (2015). From behavior to neural dynamics: An integrated theory of attention. *Neuron*, *88*(1), 127–144. https://doi.org/10.1016/j.neuron.2015.09.017

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science (New York, N.Y.)*, *315*(5820), 1860–1862. https://doi.org/10.1126/science.1138071

Buschman, T. J., & Miller, E. K. (2009). Serial, Covert Shifts of Attention during Visual Search Are Reflected by the Frontal Eye Fields and Correlated with Population Oscillations. *Neuron*, *63*(3), 386–396. https://doi.org/10.1016/j.neuron.2009.06.020

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, *108*(27), 11252–11255. https://doi.org/10.1073/pnas.1104666108

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213–2218. https://doi.org/10.1097/00001756-200312020-00016

Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience*, *19*(3), 394–403. https://doi.org/10.1038/nn.4237

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814. https://doi.org/10.1016/j.neuron.2005.01.027

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race. *PLOS ONE*, *8*(7), e69684. https://doi.org/10.1371/journal.pone.0069684

Córdova, N. I., Tompary, A., & Turk-Browne, N. B. (2016). Attentional modulation of background connectivity between ventral visual cortex and the medial temporal lobe. *Neurobiology of Learning and Memory*, *134*, 115–122. https://doi.org/10.1016/j.nlm.2016.06.011

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and Strong Priors [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

Dekker, T. M., Ban, H., van der Velde, B., Sereno, M. I., Welchman, A. E., & Nardini, M. (2015). Late Development of Cue Integration Is Linked to Sensory Fusion in Cortex. *Current Biology*, *25*(21), 2856–2861. https://doi.org/10.1016/j.cub.2015.09.043

Dekowska, M., Kuniecki, M., & Jaskowski, P. (2008). Facing facts: Neuronal mechanisms of face perception [WOS:000258083200011]. *Acta Neurobiologiae Experimentalis*, *68*(2), 229–252.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

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

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Eger, E., Henson, R. N., Driver, J., & Dolan, R. J. (2007). Mechanisms of Top-Down Facilitation in Perception of Visual Objects Studied by fMRI [Publisher: Oxford Academic]. *Cerebral Cortex*, *17*(9), 2123–2133. https://doi.org/10.1093/cercor/bhl119

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Ester, E. F., Nouri, A., & Rodriguez, L. (2018). Retrospective Cues Mitigate Information Loss in Human Cortex during Working Memory Storage. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(40), 8538–8548. https://doi.org/10.1523/JNEUROSCI.1566-18.2018

Esterman, M., & Yantis, S. (2010). Perceptual Expectation Evokes Category-Selective Cortical Activity. *Cerebral Cortex*, *20*(5), 1245–1253. https://doi.org/10.1093/cercor/bhp188

Everling, S., Tinsley, C. J., Gaffan, D., & Duncan, J. (2002). Filtering of neural signals by focused attention in the monkey prefrontal cortex. *Nature Neuroscience*, *5*(7), 671–676. https://doi.org/10.1038/nn874

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A Dynamic Interplay within the Frontoparietal Network Underlies Rhythmic Spatial Attention. *Neuron*, *99*(4), 842–853.e8. https://doi.org/10.1016/j.neuron.2018.07.038

Fiebelkorn, I. C., Saalmann, Y. B., & Kastner, S. (2013). Rhythmic Sampling within and between Objects despite Sustained Attention at a Cued Location. *Current Biology*, *23*(24), 2553–2558. https://doi.org/10.1016/j.cub.2013.10.063

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

Freiwald, W. A., & Tsao, D. Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *330*(6005), 845–851. https://doi.org/10.1126/science.1194908

Fries, P. (2015). Rhythms For Cognition: Communication Through Coherence. *Neuron*, *88*(1), 220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *291*(5508), 1560–1563. https://doi.org/10.1126/science.1055465

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in cognitive sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *290*(1038), 181–197. https://doi.org/10.1098/rstb.1980.0090

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Gronau, N., Neta, M., & Bar, M. (2007). Integrated Contextual Representation for Objects' Identities and Their Locations [Publisher: MIT Press]. *Journal of Cognitive Neuroscience*, *20*(3), 371–388. https://doi.org/10.1162/jocn.2008.20027

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(1), 30–54. https://doi.org/10.1037/xhp0000290

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613. https://doi.org/10.1098/rstb.2007.2055

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration

areas. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(30), 7881–7887. https://doi.org/10.1523/JNEUROSCI.1740-07.2007

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), 665–667. https://doi.org/10.1038/nn.4284

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(10), 3088–3092.

Hutchinson, J. B., & Turk-Browne, N. B. (2012). Memory-guided attention: Control from multiple memory systems. *Trends in Cognitive Sciences*, *16*(12), 576–579. https://doi.org/10.1016/j.tics.2012.10.003

Inagaki, H., Fontolan, L., Romani, S., & Svoboda, K. (2017). Discrete attractor dynamics underlying selective persistent activity in frontal cortex. *bioRxiv*, 203448. https://doi.org/10.1101/203448

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased Activity in Human Visual Cortex during Directed Attention in the Absence of Visual Stimulation. *Neuron*, *22*(4), 751–761. https://doi.org/10.1016/S0896-6273(00)80734-5

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: Permitting preparation without movement. *Nature Neuroscience*, *17*(3), 440–448. https://doi.org/10.1038/nn.3643

Kaul, C., Rees, G., & Ishai, A. (2011). The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain. *Frontiers in Human Neuroscience*, *4*, 238. https://doi.org/10.3389/fnhum.2010.00238

Kilpatrick, Z. P. (2018). Synaptic mechanisms of interference in working memory. *Scientific Reports*, *8*. https://doi.org/10.1038/s41598-018-25958-9

Kilpatrick, Z. P., Ermentrout, B., & Doiron, B. (2013). Optimizing working memory with heterogeneity of recurrent cortical excitation. *The Journal of Neuroscience: The Official Journal of the*

Society for Neuroscience, *33*(48), 18999–19011. https://doi.org/10.1523/JNEUROSCI.1641-13.2013

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., & Sterling, P. (2006). How Much the Eye Tells the Brain. *Current biology : CB*, *16*(14), 1428–1434. https://doi.org/10.1016/j.cub.2006.05.056

Kok, P., Brouwer, G. J., Gerven, M. A. J. v., & Lange, F. P. d. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *33*(41), 16275–16284. https://doi.org/10.1523/JNEUROSCI.0742-13.2013

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kok, P., & Turk-Browne, N. B. (2018). Associative Prediction of Visual Shape in the Hippocampus [Publisher: Society for Neuroscience Section: Research Articles]. *Journal of Neuroscience*, *38*(31), 6888–6899. https://doi.org/10.1523/JNEUROSCI.0163-18.2018

Koyluoglu, O. O., Pertzov, Y., Manohar, S., Husain, M., & Fiete, I. R. (2017). Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *eLife*, *6*, e22225. https://doi.org/10.7554/eLife.22225

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *NeuroImage*, *37*(4), 1445–1456. https://doi.org/10.1016/j.neuroimage.2007.06.020

LaBar, K. S., Gitelman, D. R., Parrish, T. B., & Mesulam, M. (1999). Neuroanatomic overlap of working memory and spatial attention networks: A functional MRI comparison within subjects. *NeuroImage*, *10*(6), 695–704. https://doi.org/10.1006/nimg.1999.0503

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164.

Lenartowicz, A., Escobedo-Quiroz, R., & Cohen, J. D. (2010). Updating of context in working memory: An event-related potential study. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 298–315. https://doi.org/10.3758/CABN.10.2.298

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Ma, W. J. (2019). Bayesian Decision Models: A Primer [Publisher: Elsevier]. *Neuron*, *104*(1), 164–175. https://doi.org/10.1016/j.neuron.2019.09.037

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current biology: CB*, *14*(9), 744–751. https://doi.org/10.1016/j.cub.2004.04.028

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *19*(1), 431–441.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations Change the Signatures and Timing of Electrophysiological Correlates of Perceptual Awareness [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *31*(4), 1386–1396. https://doi.org/10.1523/JNEUROSCI.4570-10.2011

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, L. M., & D'Esposito, M. (2005). Perceptual Fusion and Stimulus Coincidence in the Cross-Modal Integration of Speech [Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive]. *Journal of Neuroscience*, *25*(25), 5884–5893. https://doi.org/10.1523/JNEUROSCI.0896-05.2005

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370–373. https://doi.org/10.1038/nature01341

Moore, T., & Fallah, M. (2001). Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences*, *98*(3), 1273–1276. https://doi.org/10.1073/pnas.98.3.1273

Murphy, A. P., Ban, H., & Welchman, A. E. (2013). Integration of texture and disparity cues to surface slant in dorsal visual cortex. *Journal of Neurophysiology*, *110*(1), 190–203. https://doi.org/10.1152/jn.01055.2012

Murray, A. M., Nobre, A. C., Clark, I. A., Cravo, A. M., & Stokes, M. G. (2013). Attention restores discrete items to visual short-term memory. *Psychological Science*, *24*(4), 550–556. https://doi.org/10.1177/0956797612457782

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *114*(2), 394–399. https://doi.org/10.1073/pnas.1619449114

Musslick, S., Saxe, A. M., Ozcimder, K., Dey, B., Henselman, G., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

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

Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(39), 17041–17046. https://doi.org/10.1073/pnas.1001699107

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*(4), 486–511. https://doi.org/10.1037/rev0000101

Nee, D. E., & Jonides, J. (2009). Common and Distinct Neural Correlates of Perceptual and Memorial Selection. *NeuroImage*, *45*(3), 963–975. https://doi.org/10.1016/j.neuroimage.2009.01.005

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–373. https://doi.org/10.1162/089892904322926700

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

O'Reilly, R. C., Mozer, M., Munakata, Y., & Miyake, A. (1999). Discrete Representations in Working Memory: A Hypothesis and Computational Investigations. *Proceedings of the Second International Conference on Cognitive Science Tokyo, Japan*, 183–188.

Palmer, t. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. https://doi.org/10.3758/BF03197524

Panichello, M. F., & Buschman, T. J. (2020). Selective control of working memory in prefrontal, parietal, and visual cortex [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.04.07.030718. https://doi.org/10.1101/2020.04.07.030718

Panichello, M. F., Cheung, O. S., & Bar, M. (2013). Predictive Feedback and Conscious Visual Experience. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00620

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nature Communications*, *10*(1), 3366. https://doi.org/10.1038/s41467-019-11298-3

Panichello, M. F., & Turk-Browne, N. B. (2020). Behavioral and neural fusion of expectation with sensation [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.07.03.187146. https://doi.org/10.1101/2020.07.03.187146

Papadimitriou, C., Ferdoash, A., & Snyder, L. H. (2015). Ghosts in the machine: Memory interference from the previous trial. *Journal of Neurophysiology*, *113*(2), 567–577. https://doi.org/10.1152/jn.00402.2014

Papadimitriou, C., White, R. L., & Snyder, L. H. (2017). Ghosts in the Machine II: Neural Correlates of Memory Interference from the Previous Trial. *Cerebral Cortex (New York, N.Y.: 1991)*, *27*(4), 2513–2527. https://doi.org/10.1093/cercor/bhw106

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–1231. https://doi.org/10.1037/a0030947

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid Forgetting Results From Competition Over Time Between Items in Visual Working Memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(4), 528–536. https://doi.org/10.1037/xlm0000328

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, *16*(2), 283–290. https://doi.org/10.3758/BF03203943

Piet, A. T., Erlich, J. C., Kopec, C. D., & Brody, C. D. (2017). Rat Prefrontal Cortex Inactivations during Decision Making Are Explained by Bistable Attractor Dynamics. *Neural Computation*, *29*(11), 2861–2886. https://doi.org/10.1162/neco_a_01005

Posner, M. I. (1980). Orienting of attention [Place: United Kingdom Publisher: Taylor & Francis]. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

Rademaker, R. L., Park, Y. E., Sack, A. T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*. https://doi.org/10.1037/xhp0000491

Renart, A., Song, P., & Wang, X.-J. (2003). Robust Spatial Working Memory through Homeostatic Synaptic Scaling in Heterogeneous Cortical Networks. *Neuron*, *38*(3), 473–485. https://doi.org/10.1016/S0896-6273(03)00255-1

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *19*(5), 1736–1753.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714. https://doi.org/10.1016/s0896-6273(00)81206-4

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002

Reynolds, R. I. (1985). The Role of Object-Hypotheses in the Organization of Fragmented Figures [Publisher: SAGE Publications Ltd STM]. *Perception*, *14*(1), 49–52. https://doi.org/10.1068/p140049

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Rolston, J. D., Gross, R. E., & Potter, S. M. (2009). Common median referencing for improved action potential detection with multielectrode arrays. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2009*, 1604–1607. https://doi.org/10.1109/IEMBS.2009.5333230

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., & Kastner, S. (2012). The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. *Science*, *337*(6095), 753–756. https://doi.org/10.1126/science.1223082

Sachs, O., Weis, S., Zellagui, N., Sass, K., Huber, W., Zvyagintsev, M., Mathiak, K., & Kircher, T. (2011). How different types of conceptual relations modulate brain activation during

semantic priming. *Journal of Cognitive Neuroscience*, *23*(5), 1263–1273. https://doi.org/10.1162/jocn.2010.21483

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

Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience*, 3440–17. https://doi.org/10.1523/JNEUROSCI.3440-17.2018

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, *249*(4971), 892–895. https://doi.org/10.1126/science.2392679

Shin, H., Zou, Q., & Ma, W. J. (2017). The effects of delay duration on visual working memory for orientation. *Journal of Vision*, *17*(14), 10–10. https://doi.org/10.1167/17.14.10

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20. https://doi.org/10.1016/j.tics.2004.11.006

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are There Multiple Visual Short-Term Memory Stores? [Publisher: Public Library of Science]. *PLOS ONE*, *3*(2), e1699. https://doi.org/10.1371/journal.pone.0001699

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. *Neuron*, *103*(5), 934–947.e5. https://doi.org/10.1016/j.neuron.2019.06.012

Sperling, G. (1960). The information available in brief visual presentations [Place: US Publisher: American Psychological Association]. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance [Publisher: American Association for the Advancement of Science Section: Reports]. *Science*, *240*(4850), 338–340. https://doi.org/10.1126/science.3353728

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, *24*(18), 2174–2180. https://doi.org/10.1016/j.cub.2014.07.066

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. https://doi.org/10.1038/nn1669

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Tort, A. B. L., Komorowski, R., Eichenbaum, H., & Kopell, N. (2010). Measuring Phase-Amplitude Coupling Between Neuronal Oscillations of Different Frequencies. *Journal of Neurophysiology*, *104*(2), 1195–1210. https://doi.org/10.1152/jn.00106.2010

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579. https://doi.org/10.1038/21176

Treue, S., & Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*(6591), 539–541. https://doi.org/10.1038/382539a0

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit Perceptual Anticipation Triggered by Statistical Learning. *Journal of Neuroscience*, *30*(33), 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010

Vallar, G. (1998). Spatial hemineglect in humans. *Trends in Cognitive Sciences*, *2*(3), 87–97. https://doi.org/10.1016/S1364-6613(98)01145-0

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

Von Helmholtz, H. (1866). *Handbuch der physiologischen Optik*. voss.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, J., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience*, *21*(1), 102–110. https://doi.org/10.1038/s41593-017-0028-6

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249. https://doi.org/10.1073/pnas.1619153114

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. https://doi.org/10.1167/4.12.11

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439. https://doi.org/10.1038/nn.3645

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory [Publisher: Public Library of Science]. *PLOS Biology*, *18*(3), e3000625. https://doi.org/10.1371/journal.pbio.3000625

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yendrikhovskij, S. N. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology*, *45*(5), 409–417. Retrieved May 7, 2018, from https://www.researchgate.net/publication/243786913_Computing_color_categories_from_statistics_of_natural_images

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhang, W., & Luck, S. J. (2009). Sudden Death and Gradual Decay in Visual Working Memory. *Psychological science*, *20*(4), 423–428. https://doi.org/10.1111/j.1467-9280.2009.02322.x

Zhao, C., Seriès, P., Hancock, P. J. B., & Bednar, J. A. (2011). Similar neural adaptation mechanisms underlying face gender and tilt aftereffects. *Vision Research*, *51*(18), 2021–2030. https://doi.org/10.1016/j.visres.2011.07.014