Neural Circuits for Visual Working Memory

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Ivan Voitov

von Kanada

2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Dr. Georg Keller
Prof. Thomas Mrsic-Flogel

Basel, 30/03/2021

Prof. Dr. Marcel Mayor, Dekan

Contents

Abstract

Latent representations are critical for disambiguating the sensory world[1] and guiding perceptual decisions[2,3]. Visual working memory is often used to study these latent representations, but the associated neural activity patterns[4,5], their maintenance[6–8], and their distribution across the brain[9,10], remain contentious. One difficulty has come in disambiguating the neural representations underlying working memory from confounding variables introduced by the task environment. We therefore investigated visual working memory in mice alternating between performing a delayed (non)match-to-sample working memory task and a simple Pavlovian discrimination task. This experimental design isolated visual working memory engagement as the only independent variable, separable from activity associated with sensory input, movement, and reward. Transient optogenetic silencing of different cortical areas revealed a selective role of highly distributed areas of the neocortex for working memory maintenance. Neural population activity in some of these areas, namely higher visual area AM and premotor area M2, during the inter-stimulus delay period was dominated by orderly low-dimensional dynamics[11–13], which we found to be completely independent of working memory engagement. In contrast, by taking advantage of our alternating task design, we were able to decode a high-dimensional population representation of visual working memory[14,15], which was (1) present in distributed cortical areas, (2) persisted throughout the inter-stimulus delay period, and (3) predicted correct responses to the subsequent stimulus during the working memory task. Given the recruitment of such distributed neocortical representations during working memory engagement, and having observed that silencing any single area disrupted working memory, we hypothesized that these representations were instantaneously interdependent ('bound') by cortical feedback loops[16]. We tested this hypothesis directly by silencing a source cortical area while recording the feedback it received from a reciprocally connected target area. We found that transiently breaking the cortical feedback loop at the onset of the working memory delay had little effect on the low-dimensional dynamics, but selectively abolished representations of visual working memory. Our findings identify reciprocal inter-areal cortical feedback loops as key circuit motifs underlying the maintenance of distributed and high-dimensional latent representations of visual working memory.

Why working memory?

There is a long tradition of investigating working memory in neuroscience, with the focus ranging from the use of working memory as a translational marker for various neurological conditions[17–19], to constraining behavioural models of working memory[20,21], or to the physiology of persistent activity in the brain[6,22,23]. In this thesis, we will focus on working memory exclusively as an experimental handle on latent representations of the sensory world[2,24,25]. Latent representations, or internal models, are in a generously abstract sense an integral aspect of all animal behaviour – they confer the ability to flexibility associate inputs to outputs. In the process of vision, for example, an enormous cache of invariances (or *heuristics* or *assumptions*) are required to reconstruct a three-dimensional world from a sparse two-dimensional activity pattern on the retina[16,26]. For other 'basic' visual functions such as object perception, naïve template matching schemes quickly lead to combinatorial explosions and to the simplest visual tasks becoming computationally intractable[16,26]. In laboratory settings and decision making paradigms, latent representations are often operationalized as mnemonic task variables (or *contexts* or *cues*) which constrain Correct task behaviour, and thus serve as the fundamental building blocks for models of cognition[2]. Working memory, in this regard, and for the purposes of this thesis, can therefore be provisionally defined as a dynamically deployed internal representation of an external variable, which is used for a subsequent perceptual decision. This definition distinguishes working memory within the broader class of short-term memory (and functions such as motor planning), and is perhaps in its simplest form the perceptual task equivalent of an Exclusive-OR logic gate, requiring a bare minimum hierarchy of decisions (two, to be exact) to solve.

Needless to say, the validity of using working memory as an experimental handle on latent representations of the world is a difficult but necessary concept to assert. Does working memory truly underlie all aspects of animal cognition? Or is it more simply a useful, maybe even learned, faculty of the mind, on the same level as the ability to add numbers? Perhaps more importantly for neuroscience, what underlies working memory which does not elude definition across species and decision making paradigms[27,28]? For example, the phonological loop, integral to the now standard psychological models of working memory[29], is inconspicuously absent from all studies in other animals. Support for the central cognitive role of working memory largely comes from a wealth of studies on the inseparability of working memory from the earliest of perceptual processes[e.g., 30–32].

Many such studies regard working memory as both a mnemonic representation as well as an actor in perceptual processing, often either indistinguishable or a fundamental component of 'attention', which has led to puzzling questions of neural implementation. To illustrate with an example, in one early study[33], participants were asked to press a button as soon as they detected a previously cued item (e.g., an apple), while being presented a series of rapidly flashed images. There were two experimental observations; first, whether the cued item was presented as a pixel-perfect match to the target image, or simply a word describing it ('apple'), had no impact on the identification rate or reaction time, and second, the performance was stable even if the images were flashed so quickly (125 ms) that there was no conscious recollection of what the images were of. These two observations taken together posed a crucial question, how did *arbitrary* (or at least cross-modal) internal representations match their sensory targets *earlier* than the precepts were formed? Constraints on neuroscientific models for this process is the singular motive for this thesis.

The neural circuitry underlying latent working memory representations has been the focus of investigations for several decades now, but the amount of progress made has been very limited. Even a fundamental consensus regarding the potential coding schemes for working memory remains elusive[4,5]. Difficulties in specifying the neural substrates of working memory have come from the diversity of behavioural paradigms in which working memory could be engaged (with all of the associated behavioural confounds), technological constraints (non-invasive methods incompatible with single-cell recordings), and assumptions regarding the role of different brain regions and their underlying neural coding schemes (e.g., inter-areal coherence, attractor dynamics, etc.). Nevertheless, it is important to review the literature so as to at least identify why certain methodological traditions persist, and how to best make use of them going forward. As such, we will review the literature in order to assess three of the most common types of assumptions in the field; the preeminence of the neocortex, the role of working memory in perceptual decision making, and the analysis methods that have been used to identify mnemonic neural representations. Each of these topics deserves a brief background which will form Part I of this thesis. Part II will then follow with three separate experiments designed to elucidate the neural circuits underlying visual working memory, and a more wholistic review of the results, with a particular focus on their impact on interpretations of previous studies, will make up a brief Part III.

Part I – Introduction

Essay 1: The crowning achievement of evolution

A common narrative surrounding the neocortex, and the grounds for its appellation which serves as the title of this essay[34], is that the neocortex underlies the diversity and flexibility of human behaviour. As such, the neocortex is perhaps the most studied brain structure in cognitive neuroscience, and almost exclusively so in the working memory literature. Over a century of anatomy and physiology has spurred countless hypotheses regarding the functional architecture of the neocortex, and models of working memory in neuroscience have not escaped being mapped onto its biological motifs. In this essay we will outline, in roughly historical order, a small and select subset of these findings, with the goal of making the fixation on the neocortex in the study of cognition, and more specifically working memory, more tenable. As such, we will focus on three key insights of neocortical function, (1) topographic organization, (2) columnar computation, and (3) hierarchical integration.

The neocortex consists of a thin sheet of cells surrounding the cerebrum, and is itself composed of a layered structure of stratified cell sizes and densities. The global structure of such 'lamination', termed cytoarchitecture, and the concomitant parcellation of the neocortex into functional regions, is the neocortex's most defining feature. The discovery of cortical lamination[35] led to the first studies of its cytoarchitecture, and by the turn of the 20th century cortical 'organs', areas with distinct patterns of lamination, were being mapped throughout the brains of dozens of species[36,37]. These areas correlated readily with those identified by associating localized lesions with selective losses of function[38], and further discoveries of topographic organization *within* identified cytoarchitectonic areas (e.g., retinotopy[39]), firmly cemented mapping of the neocortex as a corollary of mapping behavioural diversity (e.g., prefrontal 'executive function', 'what' and 'where' visual streams).

Meanwhile on the physiology front, shortly following the advent of electrophysiological recording from neurons[40], topographical maps were reaffirmed with behaviourally localized activity patterns ('receptive fields') in visual[41], somatosensory[42], and motor[43] areas. Owing in part to the ease with which retinal stimulation could be controlled and tracked through to the cortex, the primary visual cortex emerged as a model system of study. One important observation was that although the density of retinal afferents decreased drastically in the periphery of the retina, thalamic afferents were evenly distributed across the primary visual cortex[44]. This phenomenon, termed cortical magnification, coincided precisely with the reduction of visual acuity across visual space[45], and could predict the cytoarchitectonic boundaries of the primary visual cortex[46], suggesting that each fixed surface area of cortical surface area 'processed' some minimal behavioural function (e.g., 67 μm^2 for minimum angle of resolution). Concurrently in the somatosensory cortex, topologically organized sub-modality representations, termed functional columns, were discovered[47]. Together, these studies incorporated a critical component into future frameworks of neocortical function, wherein the local vertical structure (columns) of the neocortex encapsulated the necessary machinery for the processing of its input. Finally, in one remarkable study[48], it was found that within all of the areas and species examined (excluding the primate primary visual cortex), in spite of two- to three-fold differences in cortical thickness, the number of neurons in any given column was the same (e.g., 660 neurons under a 67 μm^2).

From these observations it followed that the cytoarchitectonic differences between cortical areas, and the corresponding functional consequences, resulted from differences in the local laminar structure and connectivity motifs (microcircuits). As such, subsequent anatomical mappings of cortical microcircuits catalysed seminal insights regarding corticocortical communication[49,50], where intareal connectivity patterns were grouped into feedforward, feedback, and lateral motifs based on where in the laminar structures the source afferents terminated. This all came to a head in one seminal study[51] which mapped 25 cortical areas into a hierarchy of visual processing in the macaque brain, and revealed two primary features of corticocortical connectivity. First, despite the existence of multiple functional streams each with multiple stages of processing, the overall network was surprisingly densely interconnected (and subsequent work in mice underscores this point[52]). Second, all (but one) of the connections studied were reciprocal – if one cortical area projected to another, then it, in turn, received feedback projections from it.

The resulting view of the neocortex, as a densely and reciprocally interconnected web of functionally distinct areas, each with their own levels of abstraction (or modalities), and topographically organized columnar processing units, was readily appropriated by cognitive scientists within connectionist models for inferential processing[16,53–55]. Fundamentally, as we will discuss in the next essay, these models stressed the interdependent nature of distributed latent representations in the neocortex.

Essay 2: Perception is inference

At the heart of our perceptual abilities is the immediacy with which we disambiguate incoming sensory inputs, an idea which was first formalized by Helmholtz with the theory of unconscious inference[1]. It is easy to imagine that no two sensory inputs are ever really the same, even if they are of the same percept – our eyes are constantly moving around a three dimensional world, rotating (i.e., head tilt), and coming closer and further from the things we wish to see. From a sensory input perspective, no two objects are more different than a single object simply projected onto two different locations on the retina. Our ability to withstand such heterogeneity, at least for simpler features (e.g., translations, lighting conditions), is often termed perceptual invariance or perceptual constancy[56], and is rooted in an enormous amount of prior knowledge that we use in generating our precepts. Illusions may be thought of as the other side of the coin in the use of prior knowledge, wherein false percepts of colour, motion, size, etc., reveal the inner workings of our perceptual mechanisms when faced with unusual sensory input. The difficulty in visual processing thus lies in the need to both generalize over as many possible ways of seeing something without losing sight of what it is that is being seen.

Human-engineered solutions for visual processing exist, most often in the form of vast statistical models which learn to extract progressively more invariant features from huge repositories of sensory experience (data), and can even generalize sufficiently to achieve near-human-level object recognition performance. Natural vision extends beyond simply invariance, however, as we are constantly contextualization our perceptual processes to serve behavioural needs – we can identify things we've never seen (and recognize them as such), see compositional variations (leg of a chair vs. the chair itself), and give appropriate valence to identical input (e.g., your own hand waved in front of your face vs. someone else's), among many others. As such, dynamic contextual control of vision (often subsumed into 'attention') forms a core principle of biological models of vision[57,58]. This need to contextualize our visual processing parallels the more restrictive definition of working memory we laid our earlier.

The cognitive sciences have taken up the challenge of dissociating and studying such latent influences on perception through multiple experimental paradigms. Examples include mental imagery[59], attention[60], illusions[61], and biases[62]. Accordingly many computational (often non-mechanistic) models have been proposed for how perceptual processing organizes dynamic representations under the influence of latent variables[e.g., 54,63,64], and although many of these models do not strictly identify a form of working memory, their operations are often the focus of working memory studies which try to experimentally isolate a contextually 'primed' state of vision during an inter-stimulus delay.

A common thread within these models is some form of feedback. This is not feedback in the anatomical sense of back-projecting axons, but as a resonating signal (or message passing, or particle filtering, etc), which is iteratively updated to constrain a local sensory state within a global (latent) state (most often in the form of a fixed point or an energy minimum, if the iterations are formalized as dynamical systems). More simply, and when meshed within the biological mappings proposed by these models, what one area tells another depends on what the other area tells it in turn. If this was not true, there would be no way to communicate anything meaningful (high dimensional enough) across levels abstraction – put yet another (final) way, a combinatorial explosion occurs if you try to map everything you can see to every way that you can see it. As discussed in the previous essay, this ties in nicely within the anatomy of the neocortex, where computational modules at different levels of abstraction communicate reciprocally and form hierarchies of abstraction. The question of whether feedback representations of working memory are continuously interdependent on feedforward representations we be a key question explored in this thesis.

Many studies have sought to understand the role of working memory in perceptual decision making, a classic example being those which employ a visual delayed-match-to-sample paradigm[22,23,65,66]. Such studies have often used the observed neural correlates during the inter-stimulus delay periods to constrain the possible neural implementations of latent contextual influences on perception. However, the diversity of possible behavioural confounds and analysis methods has limited the progress which has been made. Tasks which are referred to as working memory range broadly from change detection to delayed choice to simply alternating running patterns. Perhaps the biggest obstacle is in the choice of neural representation to analyse, which is often not specified by any theoretical models (are nodes in a graph neurons? Is communication achieved by synchronization? Plasticity? Resonating activity patterns?). These issues will be discussed in the next section.

Essay 3: Attractive dynamics

The first studies which investigated correlations between short-term memory and the activities of single cells[22,23,66] identified clear persistent responses during the inter-stimulus delay periods. These responses were found in several regions, namely the parietal and prefrontal cortex of macaque monkeys, which were previously identified to support cognitive function by lesion studies. Observations of persistent delay activity have since been made in multiple task contexts, such as parametric working memory[67] and change detection, and in many distributed areas of the neocortex[68]. Such single cell neural correlations have since come under scrutiny, however, due to the limitations of sampling only the most active cells in extracellular electrophysiology experiments, and the trial-averaging approaches used to observe such trends. In particular, recent single-trial analyses have reported the mnemonic neural correlates are best explained by sharp bursts of increased activity in individual trials, with highly variable onsets during the delay, which only appear as uniform persistent activity when averaged together[14,69]. This has led to considerable debate regarding the neural implementation of working memory by persistent activity in the neocortex[4,70].

One solution to such ambiguities has been to correlate task variables with the activity of populations of many cells (i.e., state space analyses). In such approaches, with limited temporal smoothing, the stochasticity of single cell responses disappears, and clear, correlated modes of activity within the population are uncovered[71,72]. Nevertheless, such approaches when applied to the identification of mnemonic delay presentations, often employing tools and terminology from the analysis of dynamical systems, have come with their own limitations. Perhaps most importantly, whether or not the correlated modes of activity among cells are a relevant property of the recorded population (i.e., are involved with maintaining some mnemonic representation), or are simply a by-product having shared external input, is difficult to ascertain. This problem is simply due to the fact that the dynamical regime of a system is determined by the input it is receiving, and if only one part of a larger recurrent system is being measured (i.e., the local activity patterns), one has to effectively assume that no input is being received during the delay. As such, difficulties have emerged in applying these methods even during behaviours with clear dynamical readouts (movement), and have found mixed interpretations for motor planning representations[73,74]. Nevertheless, due to their analytical tractability, several low-dimensional population-level models of persistent dynamics have identified novel mechanisms for working memory maintenenace[7,75,76].

Several of the above issues have motivated studies in which mnemonic representations are contrasted across tasks[77] and across different epochs within multi-stage decision making paradigms[78]. Such studies have found evidence for distinct dynamical regimes, wherein low-dimensional dynamics are accentuated when a motor-contingency is applied to ‘entangled’ latent sensory representations. The experimental designs of this thesis may be thought of extensions of such work.

Part II – Experiments

Experiment 1

Maintenance of visual working memory by distributed cortical networks

Background

When working memory is studied experimentally within a decision making framework, it is never the only behavioural variable active in the task environment – at the very least, reward expectation and motor planning are necessarily involved and recruit their own neural processes. There is considerable uncertainty in the literature regarding the significance of such variables and the conclusions which may be drawn from experimental observations of delay-related activity[4], but perhaps more fundamentally there is a question whether such variables constitute experimental confounds or are simply tandem aspects of working memory. A broad range of tasks with motor preparatory confounds, such as oculomotor delayed response[66] or spatial navigation[79,80], are often used to assess working memory. The key inadequacy of such tasks is that during the delay, the animals already know the decision which needs to be made, and the solution to each task could potentially be reduced to simply delaying actions without the need to maintain internal sensory representations. Recently, a significant amount of work has gone into dissociating the neural strategies implemented during mixed motor and sensory representations[77,78,81]. This untangling process is itself complicated by the numerous working memory paradigms used across multiple species. In rodents, for example, T-maze arm-alternation tasks with no mnemonic sensory component at all have been used to evaluate working memory representations[80], and studies employing 'virtual reality' often link confounding sensory stimulation to movement readouts of behaviour[12,82]. Even in non-human primates, where tasks of sufficient complexity have been able to control for movement and reward contingencies, microsaccades during the working memory delay have been found to encode subsequent saccades[83], and even the sensory cues themselves[9]. In general, however, it has been observed that eliminating early knowledge of motor or reward contingencies results in the sparsening of representations underlying latent sensory variables both in primates[14,15], and in the few recent studies in mice which assess working memory with tasks that require mnemonic sensory representations[84–86].

We therefore designed a behavioural task which could isolate the engagement of working memory as an independent variable, and allow us to assess the representations of working memory agnostic to potential behavioural confounds. We used a Go/No-Go delayed (non)match-to-sample task to engage working memory representations of visual cues. Similar tasks have been previously used for visual working memory in non-human primates[87]. The primary innovation in this experiment was that we further contrasted this Working Memory task, in alternating blocks of several hundred trials within the same session, with a simpler Discrimination task which did not require working memory. Both tasks were performed by head-fixed mice with simple oriented bar stimuli in order to constrain as much as possible any movement or visual stimulation differences between the tasks.

The final goal of this experiment was to use this task to identify just how distributed working memory representations are in the neocortex. Given the complex association of working memory with other behavioural variables, identifying an area or network involved with working memory has been difficult. Although most studies of working memory have focused on the parietal and prefrontal cortices, recent studies mapping distributed area of the brain have found neural signatures of working memory in most regions studied[68]. In studies employing optogenetic silencing in mice the results have been less clear[81,84], however, potentially due to silencing induced disruptions of movement planning, or to the robustness of delay-related representations to silencing masking any effects of transient disruptions[88]. As such, in addition to isolating working memory engagement with our task, we developed an optogenetic silencing design which not only targeted multiple areas of the neocortex, but also at different epochs within each trial. Such a design allowed us to assess both the distribution of working memory in the neocortex as well as dissociate the movement and mnemonic effects of silencing.

Methods

Animals, ethics, and surgical procedures

All experiments were conducted in accordance with institutional animal welfare guidelines and licensed by the UK Home Office. A total of 9 PV-Cre × Ai32 (LSL-ChR2) mice were used for this experiment. Mice were of either sex and were between 8 and 16 weeks old at the start of their experiments.

Surgical procedures

Prior to all surgeries, the mice were injected with an analgesic (carprofen 5 mg kg^{-1}). General anaesthesia was induced with 3% isoflurane which was then reduced to maintain a breathing rate of around 1 Hz. A custom-designed stainless steel headplate was attached to the skull using dental cement (C&B Super Bond). In some of the older mice, the dorsal surface of the skull was carefully thinned with a dental drill. The exposed skull was then sealed with a thin layer of light-curing dental composite (Tetric EvoFlow).

Intrinsic imaging

We used intrinsic signal imaging of the dorsal cortex to identify the locations of cortical areas V1 and AM. Intrinsic imaging was performed on awake mice while they were head-fixed on top of a freely rotating Styrofoam cylinder. The visual cortex was illuminated with 700-nm light, a macroscope was focused 500 µm below the cortical surface, and the collected light was bandpass filtered centred at 700 nm (10 nm bandwidth; 67905, Edmund Optics). The images were acquired at a rate of 6.25 Hz with a 12-bit CCD camera (1300QF, VDS Vosskühler), an image acquisition board (PCI-1422, National Instruments), and custom software written in LabVIEW (National Instruments). The visual stimuli, presented on a display 22.5 cm away from the left eye, were generated using Psychophysics Toolbox running in Matlab (MathWorks), and consisted of square-wave gratings, covering 40° visual angle, 0.08 cycles/degrees, drifting at 4 Hz in 8 random directions, presented on a grey background for 2 seconds, with 18 second interstimulus intervals. The gratings were presented alternatively at two positions, at 15° elevation and either 20° or 80° azimuth. Response maps to the grating patches at either position were used to identify the centres of V1 and AM, using a previously published reference map[89].

Behavioural training and task design

Mice were trained for 2-6 weeks prior to the initiation of data acquisition. Mice were food restricted throughout the full experiment, with no scheduled breaks. The maximum weight loss was restricted to 20% of their initial body weight. Food restriction began at least 3 days following their headplate implantation surgery. The mice were trained for approximately 2 hours every day, once a day, and data acquisition days lasted approximately 3 hours each, occurring on average every 2 days. For the first few days of training the mice were handled on a cloth and iteratively fed Ensure Plus strawberry milkshake (Abbott Laboratories) through a syringe in order to acclimate them to the behavioural training environment. Over the next few days, the mice were trained to run on a freely rotating Styrofoam cylinder, while head-fixed, in front of the visual stimulation display (Dell U2415, 60 Hz), placed 22.5 cm away from their left eye.

Once the mice were running freely, they were trained to perform a simple visual detection task, where the onset of a visual stimulus was associated with reward. Visual stimuli consisted of large drifting square-wave gratings presented at 100% contrast, 0.025 cycles per degree, covering 60° degrees of the mice' visual field, centred at 15° elevation and 45° azimuth, presented on an isoluminant grey background. The luminance of the monitor was set at 0 cd/m^2, 22.5 cd/m^2, and 45 cd/m^2, at black, grey, and white values, respectively. The grating stimuli were cycling in closed loop with the mouse running for the first 1-2 weeks of training, and were then fixed at 3.5 Hz for the remainder of the experiment. The mouse running speed was recorded with a rotary encoder (05.2400.1122.1000, Kübler), and the mice had to run a specified distance between rewards. This distance was set so that the mice received roughly 1 reward per minute. A reward delivery spout was positioned under the snout of the mice from which a drop of Ensure Plus was delivered when triggered by licking of the spout during a response window of 1 second following stimulus onset. If the mice failed to lick in response to the stimulus, an automatic reward was delivered at the end of the response window. Licks were detected with a piezo disc sensor placed under the spout. The detection of licks, reward delivery, recording of data, and the presentation of visual stimuli were controlled by a custom LabVIEW code (National Instruments), the custom stimulus presentation software was written in the Unity engine (Unity Technologies), and hardware interfacing was achieved with a data acquisition board (PCIe-6321, National Instruments).

Once the animals were running and licking in response to the presentation of grating stimuli, the task parameters were changed in order to begin training either the Discrimination or Working Memory tasks (with the order varying between mice). The orientation of the stimuli now classified them as either Go or No-go (see Figure 1). If mice licked the spout during a 1 second response window from the onset of the Go stimulus, trials were classified as hit trials, otherwise they were classified as miss trials. In the miss trials, the mice received an automatic reward at the end of the response window. The same response window was used to classify No-go trials into false alarm or correct rejection trials. Licking to the No-go visual stimulus (False Alarms) was not punished. Both tasks consisted of a series of alternating stimulus (grating) and delay (grey background) periods. Delay durations were sampled from an exponential distribution with a mean of 800 ms, and then had 800 ms added (i.e., a 800 ms offset/minimum duration). Sample durations were capped at 4000 ms, and when the cap was reached the delay duration was resampled uniformly from between 3600 ms and 4000 ms in order to minimally change the average delay durations. The resulting average delay duration was 1600 ms. The stimulus duration was determined by the mouse running speed, and set to either 100 cm or 80 cm depending on their running speed, such that the stimuli took a similar time to traverse if the mouse did not stop running. The average stimulus duration was 1967 ms. This promoted constant running in mice over each sessions which in turn ensured stereotyped movement between mice and within the delay periods. The stimuli presented were either Cues (80% of trials), Probes (10%), or Targets (10%), and were sampled randomly, with the exception that after a Probe or a Target stimulus, a Cue stimulus was mandatory (100% probability). As such, Cues were only switched following Targets (in the Working Memory task; in the Discrimination task the Cues were always the same), and the majority of trials were consecutive Cues of the same stimulus/orientation. The only difference between the Discrimination and Working Memory tasks was therefore that the Cues were always 0° gratings in the Discrimination task, and of a flipped orientation to the Targets in the Working Memory task. Once the mice were performing both tasks well, the blocked task structure was introduced, which alternated blocks of 415 trials of both tasks through each session. Mice performed between 4 and 7 task blocks per session. Mice switched task blocks quickly (a few trials), as the presence or absence of the Discrimination task Cue stimulus (0° grating) was informative of the task block.

Optogenetic silencing

To silence neuronal activity, we optogenetically activated ChR2-expressing parvalbumin-positive neurons using a 473 nm laser (OBIS 473 nm LX 75 mW, Coherent) with a galvometer-scanning photostimulation macroscope[90]. Briefly, laser light was deflected off of two galvometer scanning mirrors, which targeted the light, expanded by two lenses (5x; plano-convex lenses LA1951-A and LA1384-A, Thorlabs), and then focused onto the brain with a 200 mm focal length lens (AC508-200-A, Thorlabs). A polarizing beamsplitter was placed in the light path and allowed us simultaneously image the surface of the skull in order to identify and select silencing locations. The photostimulation and image acquisition was controlled by custom LabVIEW code and a data acquisition card (PCIe 6321 ; National Instruments). The laser light was pulsed at 50 Hz, with a 50% duty cycle. The laser power was 3 mW for the first 400 ms of stimulation and then linearly tapered off to 0 mW over 200 ms in order to minimize rebound effects. The propagation of reflected light to the eye was blocked either by a cement wall around the exposed skull or a custom 3D-printed plastic lightshield implanted during the headplate surgery. Silencing occurred in 12% of trials, at onsets of either the delay onset, delay end (600 ms before stimulus onset), or stimulus onset. Because delay end silencing was difficult to interpret, as the mice could use the silencing to predict the stimulus onset and respond preemptively, we discarded those trials. The silenced areas were chosen randomly trial-to-trial, with the constraint of no two silencing trials in a row, and were either chosen by coordinates relative to bregma (areas M2 and S1), or intrinsic imaging (areas V1 and AM). A bright 473 nm masking light was flashed onto the mouse from an LED-coupled optical fibre (400 µm diameter), on each trial and at one of three onset times (chosen randomly), but always concomitant with the laser (i.e., matched onset to the laser light on silencing trials). This masking light thus had the same dynamics as the laser light, and was used to both mask the laser light and as a negative control for light onset induced behavioural changes.

Data processing

A total of 241,748 trials were collected from 9 mice. Trials during which mice stopped running during the delay, trials following either Targets or Probes (i.e., the trials which were fixed to always be Cues, see task design above), and trials when the silencing onset was at the end of the delay were excluded from the analyses.

Results

A behavioural task to isolate visual working memory in head-fixed mice.

Mice were required to switch between performing the Discrimination and Working Memory tasks in blocks of 415 trials (Fig. 1a). In both task blocks, the probability of the rewarded stimulus (Target) was 10%, and the length of the inter-stimulus grey-screen delay period was sampled from an exponential distribution ranging from 800 ms to 4,000 ms. The exponential distribution ensured a flat hazard rate for the stimulus onset and minimized expectation-based preemptive responses by the mice[91]. In order to gauge any potential experimentally uncontrolled differences in task engagement, impulsiveness, or reward expectation between the Discrimination and Working Memory tasks, we introduced a common Probe stimulus in both task blocks with the same probability as the Targets (10%). During the Working Memory task, mice performed equally well following either of the two possible Cues (n = 9 mice, p = 0.65, signed-rank test; Fig. 1b).

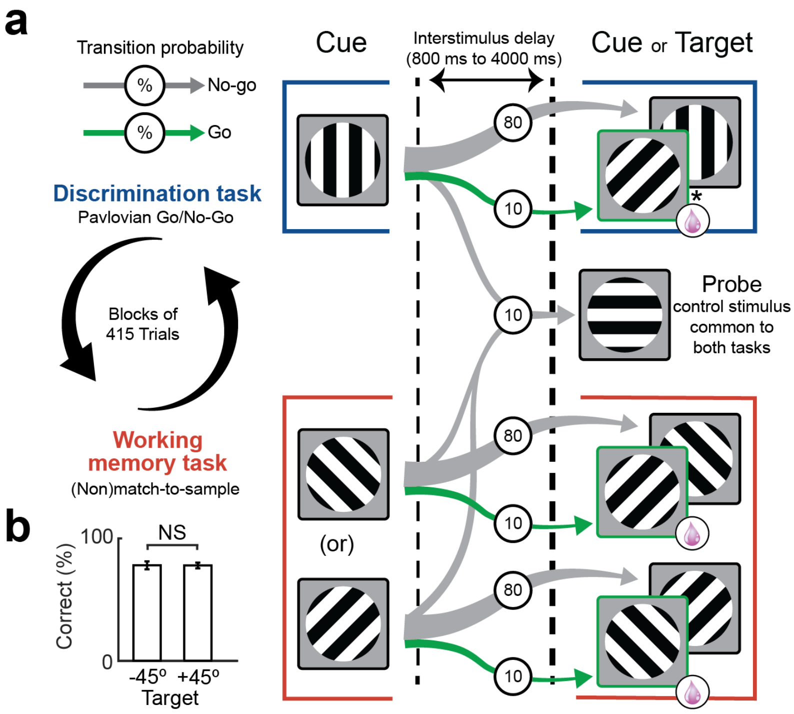


Figure 1 | Behavioural task design. a, A schematic of the task rules and transition probabilities between trial types. The probability of Cues was 80%, and the probability of Probes and Targets was 10%. The stimuli were 100% contrast square-wave gratings, covering 60° of the animal's visual field, and drifting at 3.5 Hz. The delay was an isoluminant grey screen. The Cue/Target identify of the stimulus depended on the previous stimulus in the Working Memory task, and not in the Discrimination task. Note (*) that Targets in the Discrimination task were randomly chosen -45° or +45° oriented stimuli. **b**, Percentage of correct responses in -45° Target and +45° Target trials (means are across animals, error bars are ± 95% CIs).

Mouse running speeds during the delay was stable, and the average running speed was not significantly different between the two tasks (Fig. 2a). Halting rates during the inter-stimulus delay period were not significantly different between the tasks, and early responses during the delay were very rare (fewer than 5% of trials for each mouse), and slightly higher in the Discrimination task (n = 9 mice, $p = 0.16$, $p = 0.16$, and $p < 0.01$, respectively, sign-rank tests; Fig. 2b)

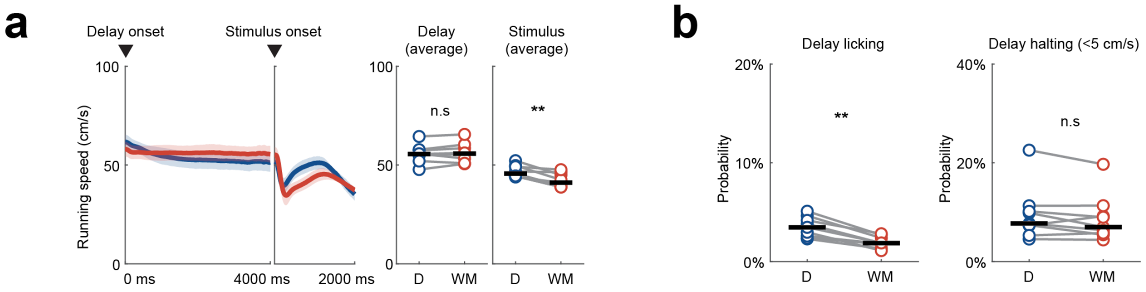


Figure 2 | Movement differences between tasks. a, Left, running speeds for the Discrimination (blue) and Working Memory (red) tasks, averaged across animals, for Experiment 1 (n = 9), during the delay and stimulus periods. Data is binned at 60 Hz, shaded regions are ± 95% CIs. Right, running speed differences between tasks, split by animal, averaged over the full duration of each delay (0 to 4,000 ms) and stimulus (0 to 2,000 ms) periods, $p = 0.25$ and $p = 0.004$, respectively, signed-rank test. Horizontal bars are the medians. **b**, Percentage of trials with early licks (left) or halts (right) during the delay period, averaged for each animal (n = 9). Early licks and halts, compared between tasks, $p < 0.01$, and $p = 0.16$, signed-rank test, respectively. Horizontal bars are the median probabilities.

The primary behavioural difference between the two tasks was that the ability for mice to correctly discriminate the correct stimulus strongly depended on the inter-stimulus delay duration only when they were performing the Working Memory task (Fig. 3a). Specifically, the false alarm rate to the Cue stimulus in the Working Memory task increased as a function of the preceding delay length (n = 9 mice, $p < 0.05$, t-test for the significance of the response-delay length slopes; Fig. 3c). Note that chance performance corresponds to when the false alarm rate curves and hit rate curves cross. The specificity of this delay length dependent impairment of performance to false alarms, not hits, is indicative of a positive bias for responding, likely due to the a lack of punishment for false alarms. Individual mouse performances, as measured by *d'*, and the variations between binned delay length intervals, is shown in Fig. 3d.

Responses to the Probe stimuli were very rare in both tasks, and did not show any delay length dependence in either task (Fig. 3b). As any task-agnostic influence of delay length on response probability (‘fidgetiness’) would have resulted in increased responses to the Probe stimuli, but the delay duration effects on performance were restricted to the working memory dependant stimuli (i.e., Cue-Target discrimination in the Working Memory task), we concluded that this delay length effect reflects the time-dependent disruption (e.g., decay[23] or interference[92]) of a working memory trace.

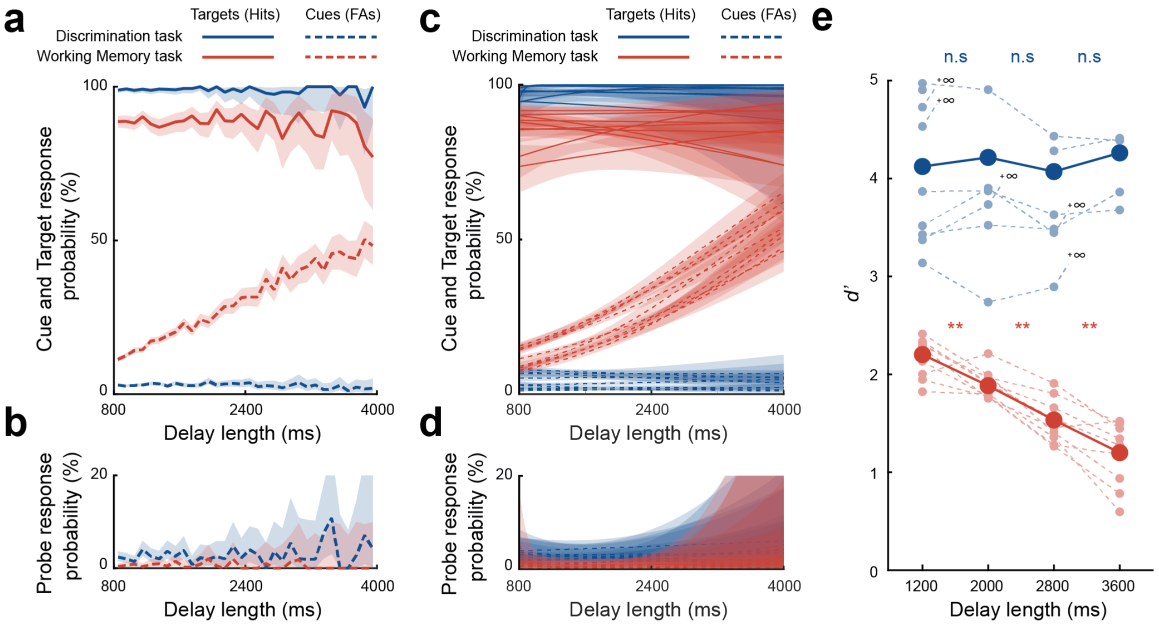


Figure 3 | Dependence of the visual working memory trace on delay length. a, Performance as a function of delay length, as measured by responses to Cues and Targets. Data is shown as mean response probabilities across trials (n = 150,381), ± 95% CIs (shaded regions), in 100 ms bins. Responses to Cues are False Alarms (FAs) and responses to Targets are Hits. **b**, Responses to task-irrelevant Probe stimuli, as in **a**. Responses are FAs. **c-d**, Same as **a-b**, with data from individual animals (n = 9) fit with a linear delay length dependence and a logit link function to the response probabilities. **e**, Performance as a function of delay length as measured by d', split by animal. Positive infinities (i.e., when no misses occurred) were treated as non-existent data points for statistical analysis. Dashed lines are the means of individual animals. Thick lines are averages across trials from all animals pooled together. Statistical tests are between quartiles of delay length (n = 9; p = 0.96, p = 0.23, and p = 0.98 for the Discrimination task, and p = 0.010, p = 0.001, and p = 0.003 for the Working Memory task; signed-rank tests).

We next investigated whether using latent representations of the Cue to infer the meaning of the stimulus during the Working Memory task corresponded to an added cost in decision making time. In agreement with this hypothesis, reaction times to the stimuli were longer in the Working Memory task to both Targets (Hits; n = 9; 304 ms and 374 ms inflection points for responses in the Discrimination and Memory tasks, respectively; $p < 0.05$, signed-rank test; Fig. 4b) and Cues (False Alarms; 135 ms and 298 ms; $p < 0.001$, signed-rank test, Fig. 4a). This was also true for the respective modes of the reaction times (Fig. 4c-d). Importantly, responses were delayed by the Working Memory task irrespective of whether they were more frequent (False Alarms) or less frequent (Hits) than in the Discrimination task, indicating that these delayed response times were resulting from systematic delays in the decision making process.

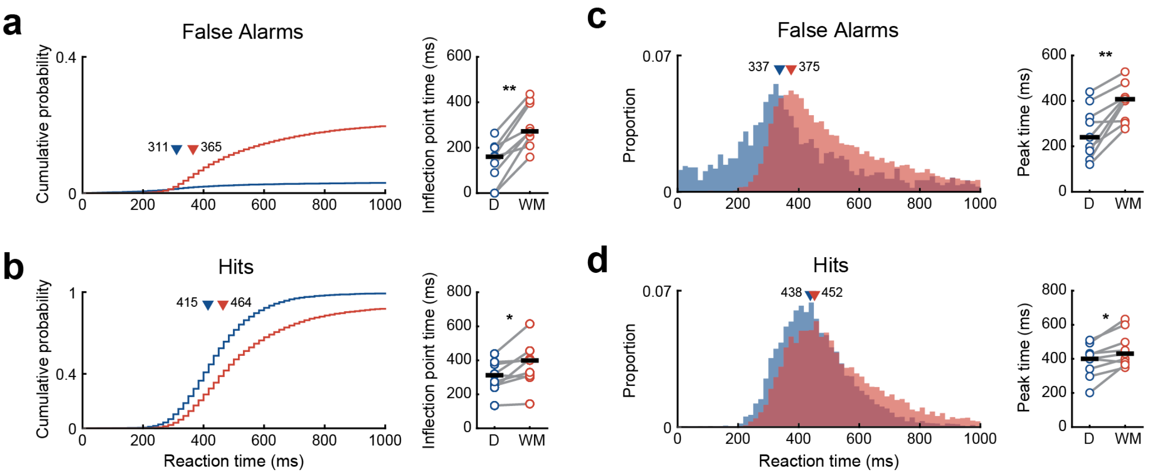


Figure 4 | Effects of working memory engagement on response times. a, Left, cumulative probability of False Alarms (FAs) to Cues at different times from stimulus onset for the Discrimination (blue) and Working Memory tasks (red). Data from all animals pooled together, and binned at 60 Hz. Arrows represent inflection point times (maxima of bin-to-bin differences). Right, differences between the two tasks for inflection point times, calculating separately for each animal ($n = 9$, $p < 0.01$, signed-rank test). Horizontal bars represent median inflection point times. **b**, Same as **a**, for Hits to Targets. $p < 0.05$, signed-rank test. **c**, Left, proportion of FAs to Cues at different times from stimulus onset for the Discrimination (blue) and Working Memory tasks (Red). Data from all animals pooled together, and binned at 60 Hz. Arrows represent the times with the largest proportion of responses (peaks of histogram). Right, differences between the two tasks for peaks, calculating separately for each animal ($n = 9$, $p < 0.01$, signed-rank test). Horizontal bars represent median peak times. **d**, Same as **c**, for Hits to Targets. $p < 0.05$, signed-rank test.

In our task design the non-Target stimuli automatically served as Cues for the subsequent trials, effectively removing inter-trial intervals and allowing us to analyse continuous sequences of delay-stimulus pairs (i.e., trials) under the conditions of differential working memory engagement. An analysis of inter-trial history effects found an impairing effect of the Target switch (i.e., in trials immediately following a Target) which was only present in the Working Memory task (although mouse performance was above chance in both tasks even immediately after a Target; Fig. 5a). Over the course of longer sequences of Cues, performance remained stable, and this was equally true when removing all trials following at least one false alarm subsequent to the most recent Target (Fig. 5b). This indicated that (1) the mice were constantly updating their working memory with each Cue, and (2) the mice were not using response contingency to do this (i.e, not inferring the stimulus meaning from a lack of reward to a False Alarm) – the mice were relying solely on visual input to update their working memory representations over minutes-long timescales.

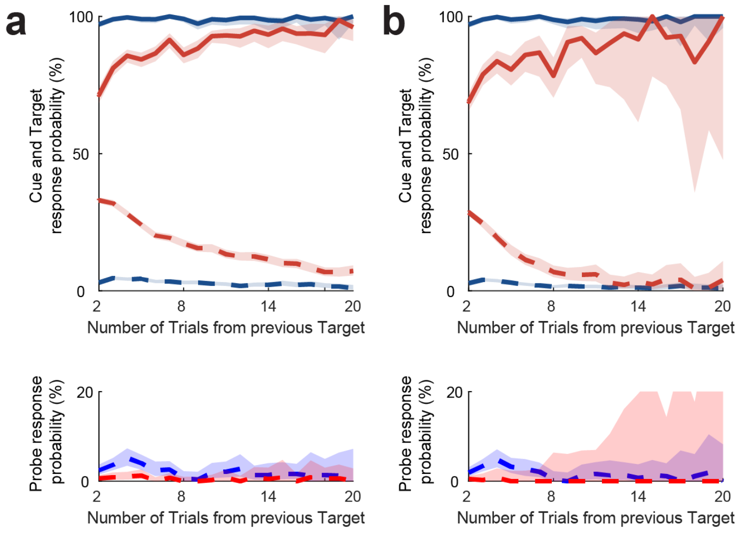


Figure 5 | Trial history effects on working memory maintenance. a, The relationship between task performance, measured by responses to Cues (False Alarms, top panel) and Probes (False Alarms, bottom panel), and to Targets (Hits, top panel), and the number of preceding trials from a previous Target. Trials were pooled from all animals ($n = 150{,}381$), solid lines represent Hits, dashed lines represent False Alarms, colour denotes task, shaded regions are ±95% CIs. Note that the animals are always well above chance performance, which would correspond to the Hit and False Alarm rates being equivalent. **b**, Same as in **a**, but only considering trials which followed an uninterrupted chain of Correct Rejections to Cues or Probes. Note the similarity to **a**, indicating that response contingency did not play a significant role in Working Memory maintanance.

Optogenetic dissociation of cortical substrates of visual working memory.

We next investigated which cortical areas supported visual working memory by contrasting the effects that optogenetic silencing had on the Discrimination and Working Memory tasks. To address both the necessity of cortical areas as well as to dissociate their role in working memory (i.e., delay maintenance or decision), we silenced one of six different areas of the dorsal cortex at either the onset of the delay or at the onset of the stimulus. The areas silenced were, on the hemisphere contralateral to the visual stimulus, V1, AM, S1, and M2, and ipsilateral AM and M2 (Fig. 6a). Area AM corresponds to a higher visual area putatively homologous to the primate parietal cortex and area M2 to either the premotor or prefrontal cortex[93–95]. The silencing itself was transient (400 ms followed by a 200 ms ramp down) and occurred in only 12% of trials. A bright masking light of matched wavelength near the mice was concomitant with the optogenetic silencing light. The masking light alone, with the same onset windows as the silencing light, had no effect on running speed or responses (Fig. 6b). Although cortical silencing slightly slowed the mouse running speeds, there was no interaction of this effect with the task that the mice were performing (Fig. 6b). Importantly, the delay onset silencing occurred outside of the earliest possible stimulus onset and response time (the delay duration minimum was 800 ms).

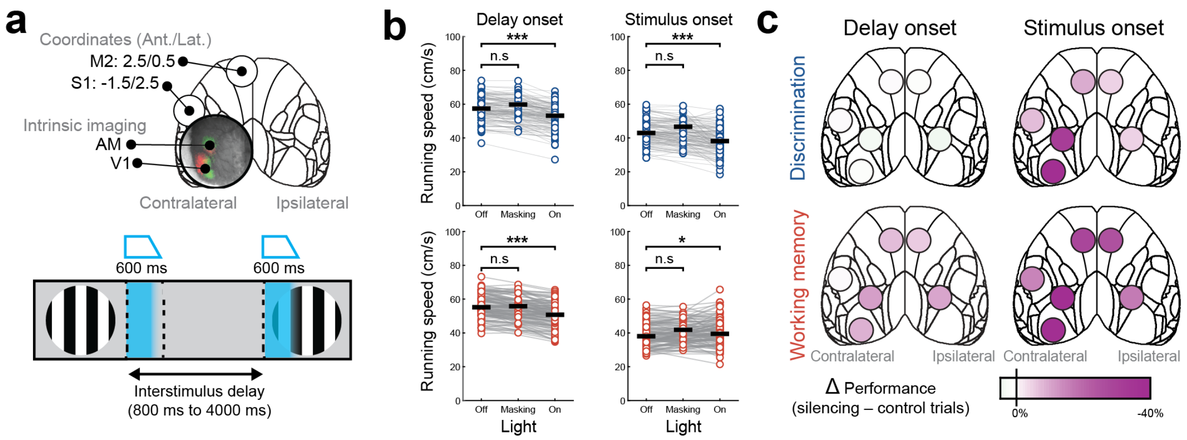
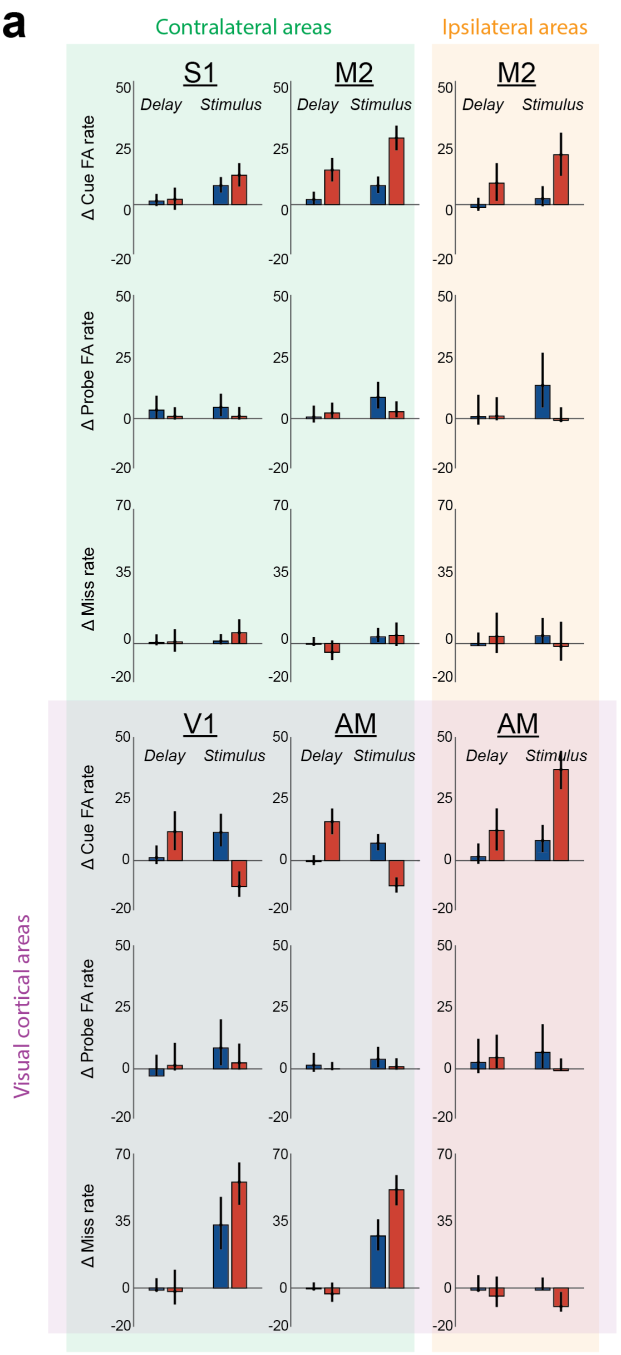


Figure 6 | Overview of optogenetic silencing. a, A schematic of cortical areas targeted for optogenetic silencing (top), and optogenetic silencing design (bottom). Areas V1 and AM were identified by intrinsic imaging of retinotopy (see methods). Areas M2 and S1 were identified with shown coordinates (millimetres anterior/lateral of bregma). The optogenetic silencing light was flashed for 400 ms at 3 mW, followed by a linear ramp down to 0 mW for 200 ms, at either the onset of the delay or stimulus periods. **b**, Effect of masking light and silencing light stimulation on running speed in the Discrimination (top) and Memory (bottom) tasks. Circles are individual sessions, horizontal bars are median running speeds. The masking light was flashed in an equivalent manner to the silencing light (see methods), and showed no significant effect on running speed. Silencing light had a small but significant effect on running speed in both tasks ($n = 124$ sessions from $\boldsymbol{n} = 9$ animals, $p < 0.01$ and $p = 0.01$, for Discrimination and Working Memory, respectively, signed-rank test). A three-way ANOVA with task, onset, and silencing light conditions found a significant effect of silencing light ($p < 0.01$), and onset ($p < 0.01$), but no significant effect of task, ($p = 0.28$) or interaction of task with light ($p = 0.13$), or onset ($p = 0.53$). Note that not all sessions had a masking light. **c**, An overview of optogenetic effects on performance split by silencing onset (left delay onset and right stimulus onset), task (top Discrimination task and bottom Working Memory task), and area (shaded circles). Performance was measured as [100% - False Alarm rate (%) - Miss rate (%)]. Colour coding is the differences in performance between control and silencing trials.

In all six cortical areas that we tested, and in both of the onset windows, silencing during the Working Memory task reduced the performance more so than in the Discrimination task (Fig. 6c). Silencing at the onset of the delay, in particular, had no significance effect whatsoever on the Discrimination task performance, in any area. We summarize all of the silencing effects in Fig. 7, simply by averaging the proportion of incorrect responses during the silencing trials relative to trials with no silencing, split by area silenced, silencing epoch, task, and the stimulus type (i.e., false alarms to Cues and Probes, and misses to Targets).

Looking further into the types of errors introduced by silencing, we found that for all of the areas we tested except area S1, silencing at the onset of the delay led to a significant increase in incorrect responses only when the mice were performing the Working Memory task (all $p <$ adjusted α, see Fig. 7 legend; fisher exact test; no effect during the Discrimination task, all $p <$ adjusted α). Importantly, only the incorrect responses which depended on the delay duration, i.e., false alarms to Cues in the Working Memory task, were increased by the silencing, and there was no effect on responses to the Probe stimuli in either task (all $p >$ adjusted α). We therefore interpreted these effects as highly selective disruptions of the visual working memory trace.



b

	V1	AM	iAM
Cue	4.0514e-01	8.6605e-01	5.8927e-01
	1.0066e-03	**2.6577e-11**	1.1697e-02
	3.6901e-08	**1.8000e-09**	1.0581e-01
	1.2607e-03	**1.1922e-07**	**1.1941e-07**
Probe	6.2991e-01	2.4852e-01	6.7294e-01
	3.0300e-01	6.0793e-01	3.5408e-01
	4.8436e-03	6.6179e-03	**1.0253e-04**
	8.6788e-02	2.0587e-01	1.0000e+00
Target	1.0000e+00	1.0000e+00	1.0000e+00
	8.3932e-01	3.1538e-01	4.1371e-01
	1.5886e-19	**3.0898e-36**	3.0295e-02
	1.7443e-27	**3.2992e-48**	1.0000e+00

	S1	M2	iM2
Cue	1.3726e-01	5.7798e-02	2.5333e-01
	3.4089e-01	**8.0258e-11**	**1.3565e-03**
	1.8688e-12	**3.5280e-11**	**7.3194e-06**
	3.7251e-08	**3.3839e-33**	**3.8278e-25**
Probe	3.0195e-02	5.6249e-01	2.1739e-01
	1.8933e-01	1.0489e-02	1.0690e-02
	3.3052e-03	**1.0186e-06**	1.8712e-02
	1.8317e-01	2.2133e-03	1.0000e+00
Target	6.5335e-01	1.0000e+00	1.0000e+00
	7.9839e-01	1.3547e-01	4.3037e-01
	1.3798e-01	2.6914e-03	1.0000e+00
	4.9371e-02	1.4203e-01	1.3584e-02

Figure 7 | Optogenetic silencing effects on performance. **a**, Average optogenetic silencing effect on responses to Cues (FAs), Probes (FAs), and Targets (Hits) for all areas silenced (labels), during both tasks (colours), and for the two silencing onsets (left bars, right bars). Bars represent the difference in FA or Miss rate between silenced and control trials (pooled from all animals, n = 173,432). Error bars represent ±95% CIs. Shaded background group the areas silenced into contralateral, ipsilateral, and visual cortical areas. **b**, Statistical significance of optogenetic silencing effect (p values) for each effect shown in **a**, Fisher exact test, with significantly difference effects highlighted. Significance thresholds were adjusted for multiple comparisons (α = 0.0014, 36 comparisons, Bonferroni correction).

The effects of silencing at the stimulus onset were dissociated by cortical area and, surprisingly, by the task. Silencing the visual cortex (areas AM and V1) contralateral to the stimulus led to (1) an increase in misses to the Targets in both tasks, with a stronger effect in the Working Memory task, (2) an increase in incorrect responses (false alarms) to the Cues in the Discrimination task, and (3) a paradoxical *decrease* in incorrect responses to the Cues in the Working Memory task. The overall performance reduction was nevertheless higher in the Working Memory task, but this result suggested a strong divergence of the role of cortical sensory processing as a function of working memory engagement. In contrast, silencing anterior and ipsilateral cortical areas (contra- and ipsi-lateral M2, ipsilateral AM, and contralateral S1) did not significantly affect the miss rates but simply increased incorrect responses to the Cues, more so when the mice were performing the Working Memory task compared to the Discrimination task. Was the task-mediated dissociation of the effect of silencing the contralateral visual cortex during the stimulus a property of the different baseline response rates in the Working Memory task? We analysed this by looking at the temporal structure of responses in the two tasks (i.e., cumulative reaction times) over the course of the stimulus with and without silencing. Surprisingly, we found that not only were responses inhibited during the Working Memory task, but they were actually lower than the responses during the Discrimination task. This indicates that the contralateral visual cortex controls the motor component of responses selectively when the mice are engaging in a visual working memory task.

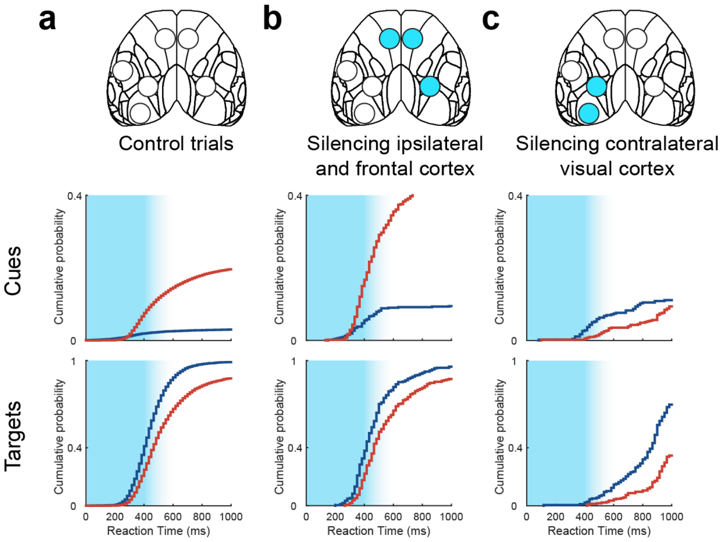


Figure 8 | Cortical dissociation of memory-guided visual processing. a, Cumulative response probabilities over time, in control trials, to Cue stimuli (top, False Alarms), and to Target stimuli (bottom, Hits). Data is equivalent to **Figure 4a-b**; pooled from all animals (n = 9, n = 173,432 trials), and binned at 60 Hz. Background shading represents the optogenetic silencing time window during stimulus onset silencing trials. **b**, Cumulative response probabilities as in **a**, from trials with optogenetic silencing at the stimulus onset of either contralateral or ipsilateral area M2, or ipsilateral area AM. **c**, Cumulative response probabilities as in **a**, from trials with optogenetic silencing at the stimulus onset of either contralateral area AM or area V1. Note the inverse relationship between the Working Memory and Discrimination Tasks for the effect of silencing on response probabilities to Cues.

In summary, we were able to identify clear psychometric differences between mice when they were engaging in a Working Memory task versus a working memory independent Discrimination task, namely the dependence of performance on the inter-stimulus delay duration, reaction times, and trial history effects – all processes which are expected to underlie working memory maintenance. Using this task, we were then able to identify a distributed role of multiple cortical regions, at the onset of the delay period, which was highly selective to working memory maintenance, and a more complex, dissociated role of the neocortex during stimulus processing.

Experiment 2

High-dimensional neural representations of visual working memory

Background

We next set out to identify the neural activity patterns underlying visual working memory. Taking a lead from the optogenetic silencing experiments, we chose to record from multiple cortical areas, specifically focusing on areas AM and M2, contralateral to the visual stimulus presentation, hypothesizing that they may have contrasting functions in supporting visual working memory. Of particular interest was the representation of visual working memory during the inter-stimulus delay period, as multiple previous studies have described persistent modes of activity during the delays of working memory tasks in the primate homologue regions[e.g., 67,76].

Although the blocked task design of our experiments allowed us to simply contrast the activity patterns of the same populations of cells between the Discrimination and Working Memory tasks during the delay periods, the fact that different stimuli served as Cues in the two tasks could potentially introduce confounds of sensory input or sensory history to the neural data. To circumvent this problem we introduced a second blocked trial structure, which ran in parallel to the task blocks used for Experiment 1, which simply involved rotating the stimuli back and forth (details will be described in the methods and results sections). This final control allowed us to evaluate the neural activity differences between the two tasks under the same sensory input, and roughly corresponds to the use of spatial location in some visual working memory studies in non-human primates to differentiate mnemonic and purely sensory responses in neurons with spatially localized receptive fields (often termed *attention-in* and *attention-out* trials)[e.g., 96].

We used two separate approaches to identify the neural representations underlying visual working memory. First, we looked for low-dimensional dynamical modes of activity within the neural populations of areas AM and M2, either through the selection of delay- or stimulus-responsive cells[23,67] or through dimensionality reduction methods[76,78]. Such methods are common, often characterize neural activity patterns as isolated dynamical systems, and have generated significant experimental evidence for the use of persistent delay activity in the mechanistic modelling of working memory[5–7]. As an alternative, we used a decoding approach in which we trained a linear model to predict the current task that the mouse was performing from the population activity, which would identify a population subspace that represents the activity added (or removed) by working memory engagement. Importantly, this latter method is not available without the two-task approach used in our experiments, and is therefore more limited in identifying working memory dynamics purely from the neuronal data (i.e., if only given the Working Memory task activity).

Methods

Animals and ethics

All experiments were carried out in accordance with institutional animal welfare guidelines and licensed by the UK Home Office. A total of 3 Ai-148 × Cux-creER (GCaMP6f expressed in most excitatory layer 2/3 cells) and 3 Ai-148 (cre-dependant GCaMP6f in all cortical cells) were used for these experiments. Mice were of either sex and were between 8 and 16 weeks old at the start of their experiments.

Surgical procedures

Prior to all surgeries, mice were injected with dexamethasone (2–3 mg kg^{-1}) and an analgesic (carprofen 5 mg kg^{-1}). General anaesthesia was induced with 3% isoflurane which was then reduced to maintain a breathing rate of around 1 Hz. A first surgery to implant a custom-designed stainless steel headplate was performed. The headplate was attached to the skull using dental cement (C&B Super Bond). The exposed skull was then sealed with a thin layer of light-curing dental composite (Tetric EvoFlow). Following a minimum recovery time of 3 days and intrinsic imaging (as in Experiment 1) to identify area AM, a second surgery was performed to make a cranial window over areas AM and M2. A 5 mm craniotomy was made over the dorsal surface of the skull and a 300 µm thick, 5 mm diameter glass window was implanted. In the 3 Ai-148 mice, a 50 nl viral injection of [AAV9.hSyn.Cre.WPRE.hGH] diluted to a low titre (5E11 vg/ml) in cortex buffer was made into AM and M2.

Behavioural task

The behavioural training and task were similar to that described in Experiment 1. The key difference was that the oriented Cue/Target gratings were ±30° instead of ±45°, and the rotation blocks were introduced. The rotation blocks did not require training, and consisted of blocks of trials during which all gratings (except for the 90° oriented Probes) were rotated 15° clockwise or counter-clockwise. In the transitions between rotation blocks, the grating angles were changed slowly (averaging 10 minutes for a full 30° rotation). The final stimulus orientations were -45°, -15°, and +15°, and -15°, +15°, and +45°, for the Cues and Targets in the clockwise and counter-clockwise rotation blocks, respectively. A typical session, therefore, involved alternating task block switches and rotation block switches such that both of the -15° and +15° oriented stimuli had paired Discrimination and Working Memory task identities across rotation blocks. All other aspects of visual stimulation were the same as in Experiment 1.

Imaging

We imaged calcium transients in layer 2/3 cells of areas AM and M2 simultaneously using a wide field of view two-photon microscope[97]. The surface blood vessel pattern above the imaging sites was compared with the blood vessel pattern from the intrinsic signal imaging maps to confirm the location of area AM. Field of views over each area were 600 µm × 600 µm and spread over four axial planes 50 µm apart. Frames from all 8 fields of view were acquired at 4.68 Hz. The image acquisition software was ScanImage[98]. Two small cameras (22BUC03, ImagingSource) were positioned to acquire greyscale videos of the body and left pupil at 30 Hz.

Data analysis

The imaging data was pre-processed using modified CaImAn software[99]. Briefly, cell masks were identified as point-seeds at individual cell locations by the experimenter, using the registered mean frame image as well as a pixel-surround correlation image. The CaImAn cell segmentation and neuropil demixing algorithms (based on constrained non-negative matrix factorization) were the applied to the seeds to define the mask boundaries and extract the calcium time series. A second round of experimenter-mediated curation was performed on these masks. The calcium traces were then detrended, normalized ($\Delta F/F_0$), and deconvolved using the standard CaImAn algorithms (FOOPSI[100]).

For all data where $\Delta F/F_0$ activity is shown (Figures 11-13), the underlying statistical analyses (e.g., estimating the latencies of delay responses) were done on the deconvolved timeseries. For all subsequent statistical modelling and population analyses (e.g., PCA, WMCD), only deconvolved calcium activity was used. Imaging frames with low correlations to the average image (putative movement artefacts), or significant pupil movements (greater than 5 standard deviations from the mean), were discarded. Individual cells were further curated following timeseries extraction using (1) spatial neuropil correlations, (2) SNR (see CaImAn documentation), and number of event (activity) restrictions. For analyses limited to delay or stimulus responsive cells (Figure 11-13), we defined responsiveness with an effect size threshold (0.2 deconvolved $\Delta F/F_0$ difference post-pre delay or stimulus) and a paired sample t-test ($\alpha = 0.01$). Although the inter-stimulus delay periods range from 0 to 4,000 ms as in Experiment 1, the fewer number of trials available for analysis within each individual imaging session led to too few long duration trials (due to the exponential distribution of delay durations), and as such all analyses were limited to trials with delay lengths ranging from 0 to 3,200 ms. For all decoding analyses (i.e., the WMCD), the model was 5-fold cross validated and all reported classification accuracies are of the left-out (test) data. The model was identified by linear discriminant analysis, but other linear methods achieved very similar results (e.g., logistic regression).

Results

Low dimensional dynamics did not discriminate working memory engagement.

In order to examine differences between the Discrimination and Working Memory tasks in the activities of individual cells within the same session without stimulus orientation specific response and adaptation confounds, we introduced a second block structure into both tasks. The orientations of Cue and Target stimuli were rotated in blocks of several hundred trials, in parallel but out of phase to blocks of the Discrimination and Working Memory task trials, with at least 2 blocks being completed per session. The rotations were of +15° and -15°, such that the Cue stimuli in the Discrimination task were of the identical orientation to one of the Cue stimuli in the Working Memory task. Accordingly, for all further analyses we were able to contrast the neural activity with and without working memory engagement in response to sequences of visually identical delay-stimulus pairs (Fig. 9a). We further constrained our analysis to trials (i.e., delay-stimulus pairs) following a common Cue stimulus (i.e., not following a Target or Probe). In order to simplify the analyses, if a single session consisted of at least one Working Memory block and one Discrimination block in both Rotation blocks, we treated the two available pairs of matched Cue task blocks as separate experiments.

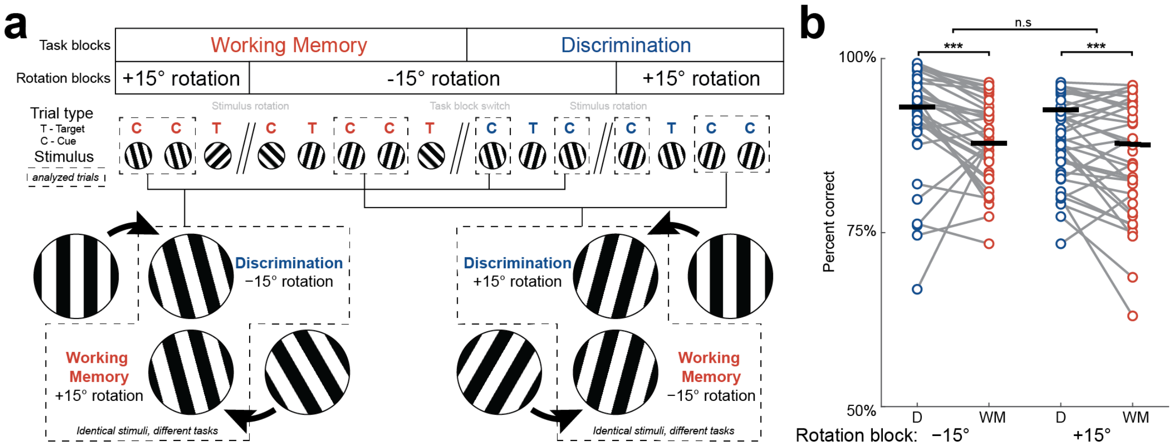


Figure 9 | Stimulus rotation. a, Schematic of the rotation blocks for the control of Cue orientation differences between the two tasks (see methods). Over the course of each experiment, Cue and Target stimuli were rotated 15° clockwise and 15° counter-clockwise in blocks of trials which were staggered relative to task trial blocks. As Cue stimuli in the Working memory task were rotated 30° from Cue stimuli in the Discrimination task, Cue stimuli from opposing rotation blocks and opposing tasks were identical in orientation (and accordingly identical in all visual stimulation parameters). **b**, Percentage of correct responses in the two rotation blocks (left -15°, right +15°). There was a significant difference in performance between tasks in both rotations ($p < 0.001$), but no difference between rotations ($p = 0.78$), and no interaction between rotations and task ($p = 0.33$) as identified by a two-way ANOVA ($n = 45$ sessions). Horizontal bars are medians.

Rotations of the stimuli did not interact with performance, which was similar across rotation blocks for both the Working Memory and Discrimination tasks (Fig. 9b). Importantly, running speed

(Fig. 10a) and other task-extrinsic arousal related variables, measured experimentally by pupil diameter[101], were not significantly different between the tasks (Fig. 10b).

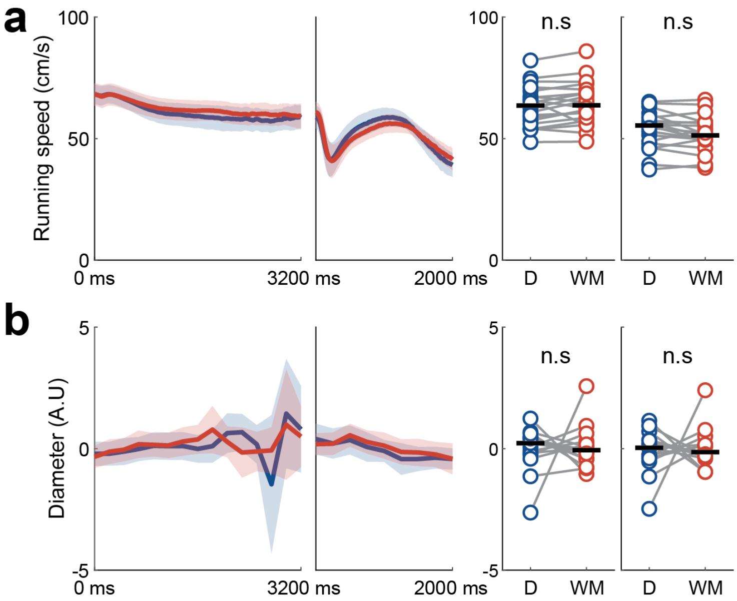


Figure 10 | Movement and arousal controls. a, Left, running speeds for the Discrimination (blue) and Working memory (red) tasks, averaged across sessions, for Experiment 2 ($n = 20$), during the delay and stimulus periods. Binned at 4.68 Hz, shaded regions are ± 95% CI. Right, running speed differences between tasks, averaged over the full duration of each delay (0 to 3,200 ms) and stimulus (0 to 2,000 ms) periods, split by experiment, p = 0.05 and $p = 0.07$, for the delay and stimulus periods, respectively, signed-rank test. Horizontal bars are medians. **b**, Left, pupil diameters for the Discrimination (blue) and Working memory (red) tasks, averaged across sessions, for Experiment 2 ($n = 15$), during the delay and stimulus periods. Binned at 4.68 Hz, shaded regions are ± 95% CI. Measured diameters were mean-subtracted prior to comparison between animals in order to better detect diameter changes. Right, pupil diameter differences between tasks, averaged over the full duration of each delay (0 to 3,200 ms) and stimulus (0 to 2,000 ms) period, split by animal, $p = 0.56$ and $p = 0.85$, for the delay and stimulus periods, respectively, signed-rank test. Horizontal bars are medians.

In both areas AM and M2, single cell responses during to the delay were at least as common as responses to the stimuli, but had more variable phases of onset and dynamics (Fig. 11a). Interestingly, none of the delay-responsive cells were clearly responsive to the stimulus, and they were present with and without working memory engagement. In fact, although the average activity of single cells was often slightly biased to one of the tasks, on average, the trial-to-trial variability in their peak activity levels far too great to be able to clearly distinguish the tasks from single cells. Both of these observations have been described in non-human primate visual delayed-match-to-sample tasks[14,15], and contrast to observations from vibrotactile parametric working memory tasks[67].

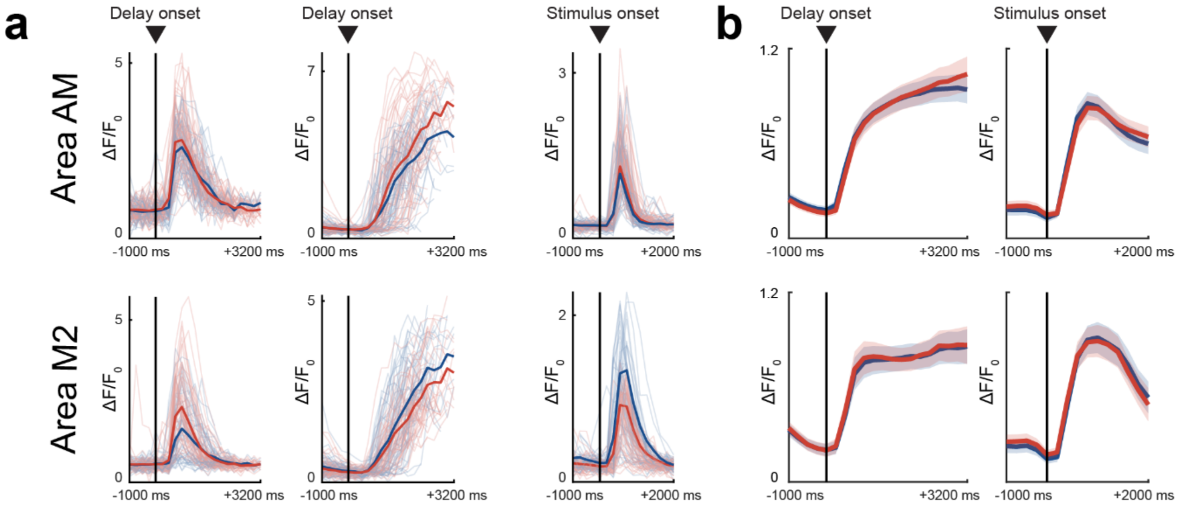


Figure 11 | Neural activity in areas AM and M2. a, Example single cell responses recorded from areas AM (top row) and M2 (bottom row). Each plot is of a seperate cell. The first two columns show cells with responses triggered off of the onset of the delay period, and the third column shows responses triggered off of the onset of the stimulus. Thick lines are the mean responses during either the Discrimination (blue) or Working Memory (red) tasks. Dim lines are responses during individual trials, selected from 50 of the longest delay duration trials per task. **b**, Cell- and trial-averaged responses ($\Delta F/F_0$) of the delay- and stimulus-responsive cells to the delay and stimulus onsets, respectively, split by task. Shaded regions are the ± 95% CIs of the means across cells. Top, area AM, bottom, area M2.

Strikingly, averaging the activity of all delay- or stimulus- responsive cells together revealed that the overall average neural activity, with and without working memory engagement, was nearly identical (Fig. 11b). Furthermore, although delay-responsive cells exhibited diverse dynamics, primarily in the latency at which they fired from the delay onset, on the population level the delay activity was persistent throughout the delay, including during the Discrimination task.

To get a better handle on how the dynamics of single cells changed as a function of working memory engagement, we fit gaussian distributions to their delay-triggered responses, sorted the cells by the means of these distributions (i.e., the latencies of the fit responses), and then plotted the sorted cells with their trial-averaged activities split by task (Fig 12). We found that the activities of delay-responsive cells had an almost evenly staggered distribution of onset times, effectively leading to a tiling the full delay period as a population. The latencies were fit and plotted from separate trials (see figure legend), and were therefore reliable from trial-to-trial, and not simply a by-product of overfitting.

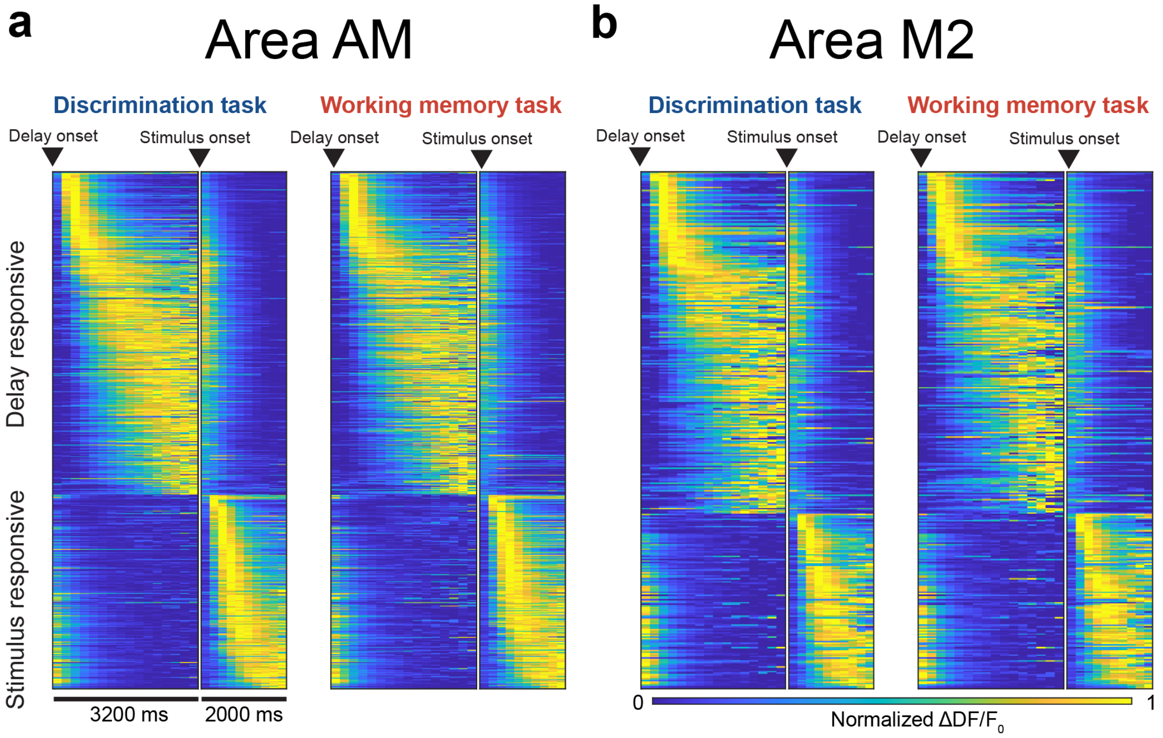


Figure 12 | Sequential activity patterns during the delay. **a**, Trial-averaged single cells responses for all delay-responsive and stimulus-responsive cells (individual rows), sorted by the latency of their trial-averaged responses from trial onset. Responses were ΔF/F$_0$, normalized independently for each cell to range from 0 and 1. All odd trials were taken out to estimate the latencies. All remaining (even) trials were split by task (Discrimination, left, and Working Memory, right) and are displayed sorted by the odd trials' latencies (i.e., rows across task plots are the same cells). **b**, Same as **a**, for area M2.

The calculated onset latencies were not significantly different between the Discrimination and Working Memory tasks (Fig. 13a). To visualize this, we plotted all cells separately for both tasks, this time sorting them by their responses in the opposing task (Fig. 13c), and found a similar pattern of sequential activity as before. Areas AM and M2 exhibited similar but significantly different proportions of ramping responses (Fig. 13b). These analysis demonstrated that at least among delay-responsive cells, the average activities and dynamics were agnostic of task, and that persistent delay activity was being driven by visual working memory. Such delay activity could nevertheless be related to variables such as motor planning or reward expectation, which are essential to decision making tasks.

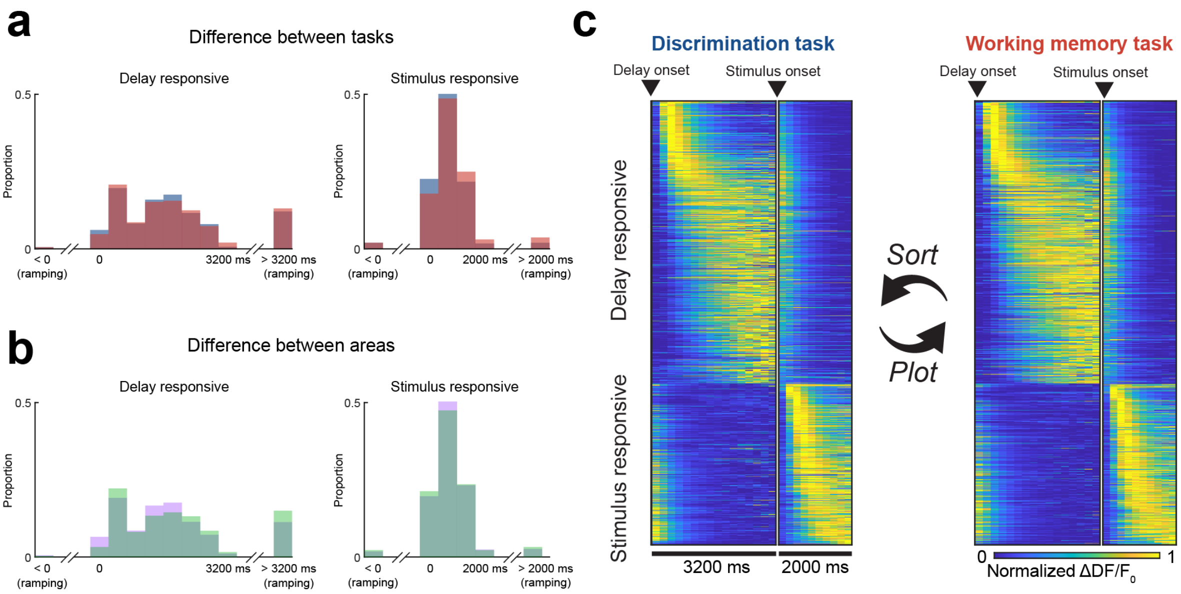


Figure 13 | Similar response latencies between tasks and areas. a, The Discrimination task (blue) and the Working Memory task (red) trial-averaged single cell response latencies for the delay (left) and stimulus (right) evoked responses. Individual bars are the proportion of cells with mean response latencies at each of the onset lags from the delay (left) and stimulus (right) onsets, binned at 400 ms. Mean latencies earlier or later than the delay or stimulus durations were representative of ramping responses and separately binned together. **b**, Same as in a, but comparing the response latencies of areas AM (violet) and M2 (green). **a-b**, A two-way ANOVA with task and area conditions did not identify a significant difference in response latencies across tasks in either the delay responses ($p = 0.43$) or stimulus responses ($p = 0.07$), a small but significantly later response latency for area M2 for the delay responses ($p = 0.01$), no significant difference between areas for stimulus responses ($p = 0.66$), and no significant interaction effects between areas and tasks in either the delay ($p = 0.86$) or stimulus responses ($p = 0.50$). **c**, A visualization of the similarity of response latencies between tasks. Data is presented in the same manner as in **Figure 12**, but pooled from both area AM and M2. The sorting of cells used for plotting each task was calculated from lthe atencies estimated from the opposing task. Note the similar patterns of sequential activation.

A common, more wholistic approach to neural population analysis is to reduce the dimensionality of the full population of recorded cells while maintaining as much of the variance as possible. To do this, we performed PCA on the trial-averaged activity patterns identified during the delay and stimulus epochs. This effectively leads to 'pseudo-simultaneous' neural activity patterns which can be pooled across experiments, and sacrifices the ability to capture trial-to-trial variability in favour of a more robust measure of trial-averaged neural dynamics[75,102]. Consistent with previous reports of low-dimensional activity modes in cortical populations[72], we captured a large amount of trial-averaged activity variance within only the first three PCs (83% in AM and 78% in M2). To visualize the data we projected different groupings of trials onto the identified PCs, grouped by task and three equal bins of the delay durations (Figure 14). The resulting trajectories were strikingly similar between tasks, with longer delays simply extending a 'rotational' mode of activity. We generated a null distributed of the Euclidean distance between these trajectories by simply shuffling the task identities of individual trials, and found that the true Euclidean distances between the Working Memory and Discrimination task trajectories were not significantly different from each other ($p > 0.01$ for all times during the delay and stimulus, adjusted $\alpha = 0.002$). These results indicate that the strongest dynamical modes (e.g., the tiling of offsets throughout the delay) of the full populations, in both areas AM and M2, were not related to whether or not the mice were maintaining a working memory trace. On the whole, these observations surprised us, as the removal of these activity patterns by optogenetic inhibition had such drastically divergent effects on behaviour depending on working memory engagement, in some cases even in the opposite directions (i.e., when silencing area AM during the Cue).

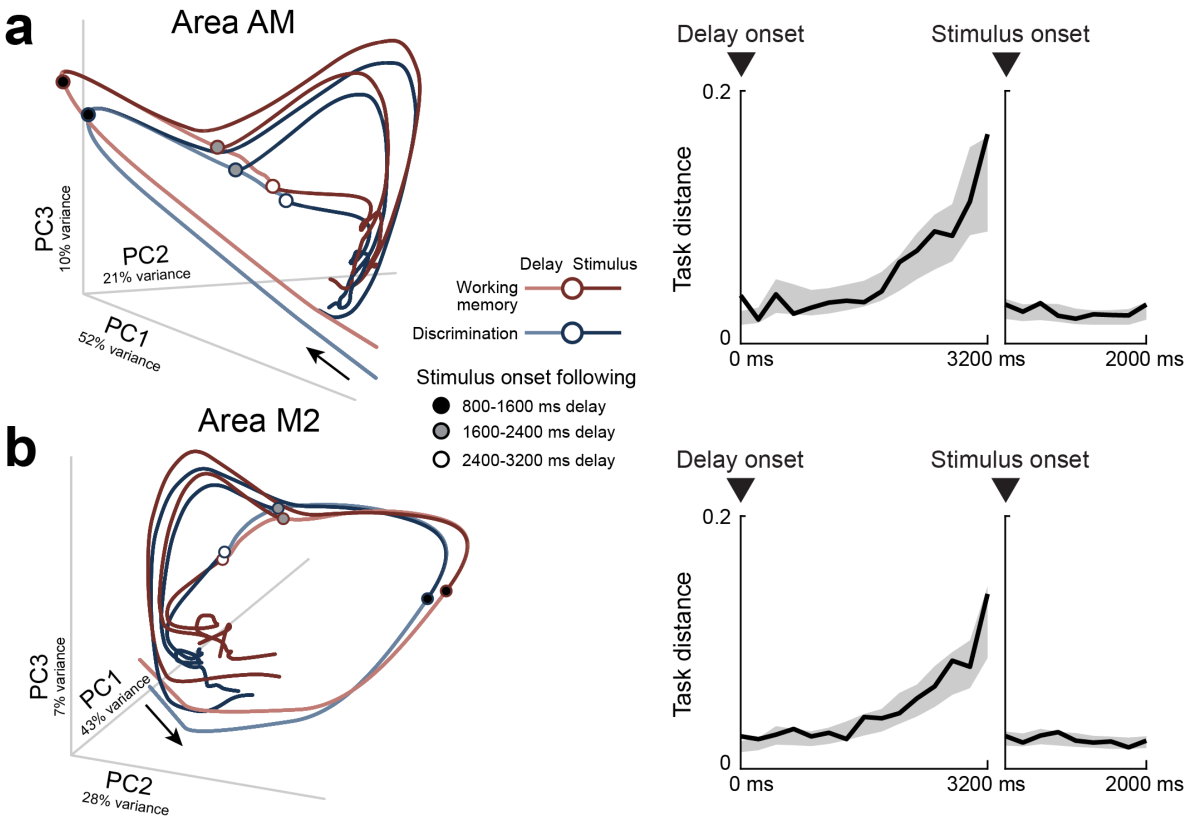


Figure 14 | Low-dimensional dynamics are unaffected by visual working memory. a, The trial-averaged responses of all active cells in area AM from all experiments (n = 18), triggered off of the delay and stimulus onsets, were used to identify the principle components accounting for the majority of all variance in the data (proportion explained (%) shown on the respective axes). Left, trials from either task (Discrimination, blue, and Working memory, red), and with delay durations of 800-1600 ms, 1600-2400 ms, and 2400-3200 ms, were averaged independently and plotted as trajectories in the principle component space. Arrows represent delay onsets and direction of time, circles represent minimum stimulus onsets for the three trajectories. Shaded lines are preceding to the stimulus onsets (representing delay-evoked activity), saturated lines are following the stimulus onsets (representing stimulus-evoked activity). Right, the average Euclidean distance between the Working Memory and Discrimination task trajectories (see methods) is plotted over the course of the delay and stimulus. Shaded regions are the ± 95% CIs of the null distribution of trajectory distances acquired by randomly shuffling the task identifies of each trial. **b**, Same as in **a**, for data from area M2 (n = 13 experiments).

Decoding a high-dimensional representation of working memory

Although single cells were not reliably more active with or without working memory engagement (Fig. 11), they often did show a slight bias in the trial-averaged peak (i.e., not phase) that their activity reached during the delay period. We therefore hypothesized that these slight fluctuations around the mean activity within each delay period could result in clear separation of the representations of the Discrimination and Working Memory tasks if summed across the full population of recorded cells. First, we analysed whether these single cell biases were significantly more distributed then would be expected by chance by plotting trial and delay averaged activities of all cells in both tasks against each other (Fig 15). Although there were no clear subpopulations of cells which were task modulated, their spread was greater than would be expected from a null distribution (r = 0.85 vs. r_{noise} = 0.97 for area AM, and r = 0.86 vs. r_{noise} = 0.96 for area M2). Importantly, we generated the null distributions by shifting the task identities of each trial half-way into the next task block, which allowed us to preserve in the null distribution any slow temporal activity drift (e.g., fluorophore bleaching or axial brain movement) which could confound measures of activity across trial blocks[103]. Having observed that the population activity was significantly biased, we next trained a linear decoder using the delay activity of individual trials to predict which task the animal was performing on any given trial, and were able to achieve a surprisingly high decoding accuracy (92% of trials had their task correctly classified for area AM, and 88% for area M2).

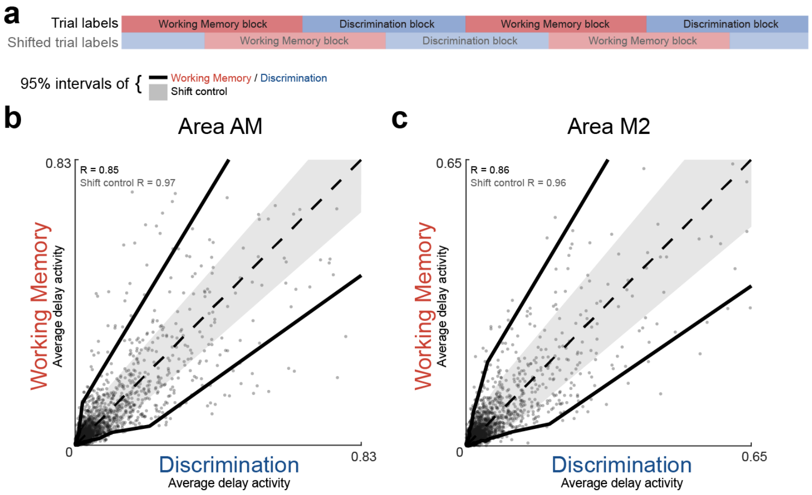


Figure 15 | Encoding of visual working memory by distributed populations of cells. a, Schematic for generating a null distribution trial-averaged delay period activity differences between the Discrimination and Working memory tasks. All trial task identities were shifted halfway through the next task block. Data from such shifted task identities maintained the long temporal structures of blocked delay period responses but abolished task-related changes. **b**, Single cell trial-averaged delay activities from area AM for both tasks plotted against each other. Solid lines encompass 95% of the data (binned into 20 equal portions of cells). Shaded regions encompass 95% of the null distribution of cells with shifted trial identities. **c**, Same as in **b**, for area M2.

How is this representation of visual working memory embedded in the neural state space? We approached this question by titrating the neural activity available to the decoder and observing the effects on task classification accuracy. Importantly, as classification accuracy was of data left out for training (i.e., the model was cross-validated), it is not necessarily true that classification accuracy would always increase with more data[75,102], both due to overfitting as well as the possibility that trial-to-trial variability was unrelated to the decoded variable of interest ('noise'). First, we randomly sampled cells and added them to the decoder, and found that roughly 100 cells were necessary to approach the full population decoding accuracy (Figure 16a). Then, we did the same procedure with the top PCs of the raw population activity (i.e., not the trial-averaged PCs as in Figure 14), and found that as many as 20 principle components were needed to approach the same classification accuracy (Fig. 16b). The fact that the variance explained by the first 20 principle components was significantly greater than the variance explained by 100 randomly sampled cells, but the classification accuracy was similar, suggested that the low-dimensional modes of activity in the neural population were not selectively modified by working memory engagement, and that the trial-to-trial variability of single cells, at least with respect to working memory encoding, was limiting the information available in the population (i.e., was not correlated[104]). To confirm this observation, we titrated the *exclusion* of the top (highest variance explained) PCs, and found that the decoding accuracy was strikingly robust to the removal of the majority of the variance of the population activity, as long as a sufficiently high dimensional representation was preserved (Fig 16c). Taken together these results indicated that the representation of visual working memory was embedded in a high-dimensional neural state space – many neurons encode visual working memory unreliably, but do not share a single (or some low number of) correlated modes for this encoding. We termed the high-dimensional subspace identified by the decoder as the Working Memory Coding Dimension (WMCD), and focus on it for later analyses.

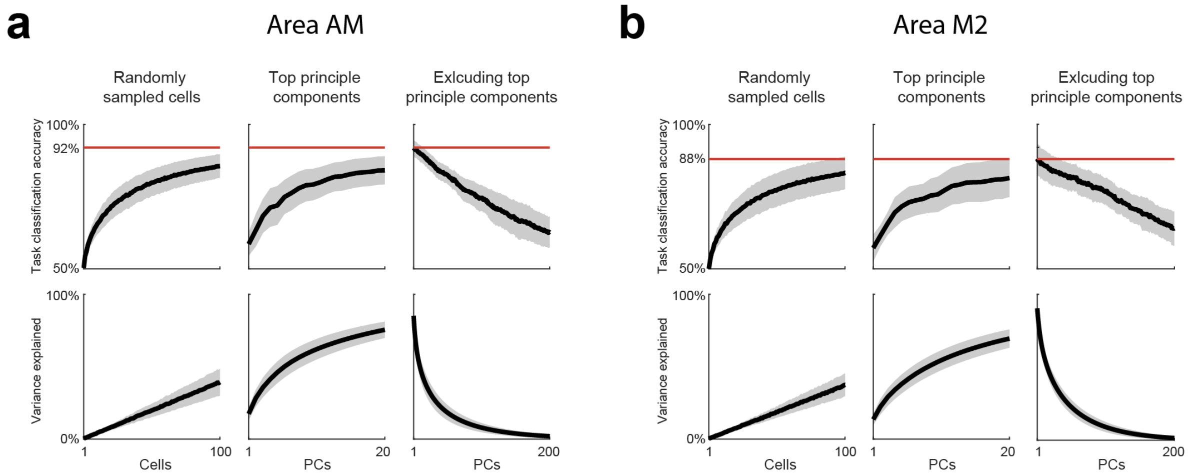


Figure 16 | A high-dimensional representation of visual working memory. a, Delay period population embedding of the working memory coding dimension (WMCD) in area AM. Top row, task decoder test classification accuracy from observing only a limited number of randomly sampled cells (left), observing only the top principle components (centre), and observing the data with the top principle components removed (right). Bottom row, the respective proportion of total variance explained in the data observed by the decoder. Black lines and shaded regions are the means and their ± 95% CIs across experiments (n = 18). Red lines represent average test accuracy when not limiting the data observed by the decoder. **b**, Same as in **a**, for area M2 (n = 13).

It is important to note that the WMCD is related to the coding dimension that separates Cue identify (i.e., the preceding stimulus orientation) during the Working Memory task, which is a subspace commonly used in the delayed-match-to-sample literature. Essentially, instead of contrasting the +15° or -15° Cues with the -45° and +45° Cues, we contrast them with the working memory independent representations of the 'neutral' Cue in the Discrimination task, but the direction (in perceptual orientation space) of contrast remains the same. The WMCD has the advantage that the Cue identify coding dimension comes with potential sensory confounds of different feedforward activity (e.g., orientation specific offset responses at the onset of the delay period). As the WMCD is defined across blocks of trials at different times within each recording session, and could therefore identify confounding slow temporal drifts of activity (e.g., fluorophore bleaching or axial movement), we performed another control (in line with the shifted-trial control in Fig. 15) in which we used the WMCD to try to classify trials which occurred either early or late in any given task block (Fig. 17a). We found that the WMCD was able to discriminate between trials early and late in each block above chance only in a small fraction of blocks, indicating that slow temporal drifts n activity were not a significant factor in defining the WMCD.

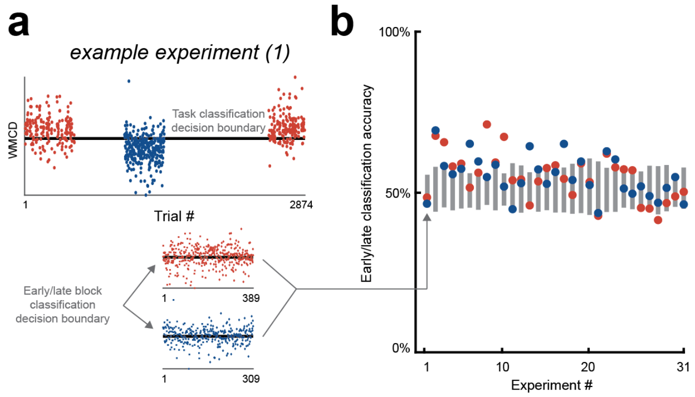


Figure 17 | Influence of time on the WMCD. a, An example analysis of the relationship between the WMCD and the slow temporal structure in population activity. The WMCD is calculated from all admissible trials (top). Empty stretches of trials represent the opposite rotation blocks in this experiment (see methods). The WMCD was then used to identify the optimal decision boundary for distinguishing between the first and second half of trials in either task (bottom). **b**, The ability of the WMCD to distinguish the first half and the second half of either the Discrimination task trials (blue) or Working Memory task trials (red) is plotted. Grey bars represent the null classification accuracy of using the WMCD to distinguish randomly sampled trials from each other. Note that few experiments were able to distinguish time above chance.

We next investigated the single-trial dynamics of the neural population activity projected onto the WMCD. Generally, we observed that on single trials the WMCD activity persisted throughout the delay. We quantified this by first limiting our analysis to trials with sufficiently long delays (greater than 3,200 ms), and then comparing the WMCD activity in the first and second half of this period.

The relationship between the two halves of the delay was significantly more robust when the mice were performing the Working Memory task as compared to when they were performing the Discrimination task, both in measures of the slope of the relationship and the regression coefficients (Figure 18). These differences were completely masked by the dominant trial-averaged modes of neural population activity, highlighting the importance of single-trial analysis[70]. Such an increase in the robustness of delay related activity along the WMCD during working memory engagement is indicative with line attractor dynamics previously suggested by recurrent neural network modelling[7,75].

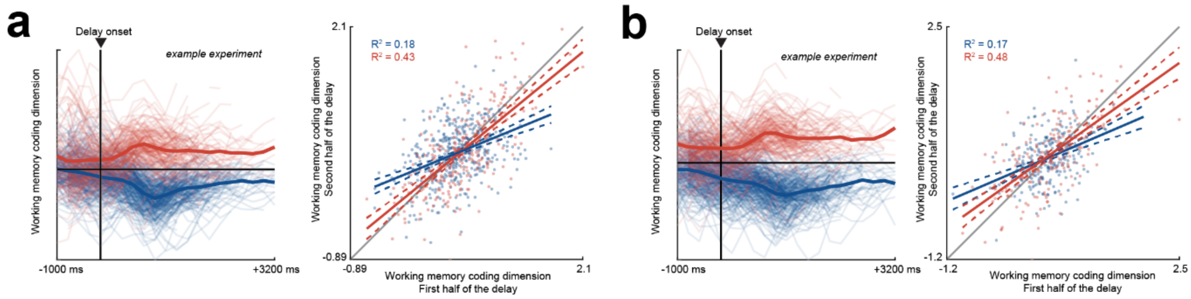


Figure 18 | Working memory coding dimension dynamics. a, Left, delay onset triggered population activity, from an example experiment recorded from area AM, projected onto the WMCD. Saturated lines are means of Discrimination task (blue) and Working Memory task (red) trials, dim lines are projections of individual trials. Right, individual trials' WMCD projected activity split into first half of the delay (<1,600 ms) and second of the delay (>1,600 ms) components. Only trials with a sufficiently long delay (>3,200 ms) durations were included ($n = 719$ trials collected from all experiments). Projections were *z*-scored within each experiment before pooling across experiments ($n = 18$). Solid lines are fitted regression lines for Discrimination task trials (blue) and Working Memory task trials (red). Regression coefficients of determination are printed in the top left. Dashed lines represent ± 95% CIs of the regression line slope. **b**, Same as in **a**, for area M2 ($n = 13$ experiments, $n = 450$ trials).

A key property that a population code which captures the latent representations of working memory should have is the ability to predict lapses or errors in working memory encoding. We found that the delay activity prior to an incorrect response to the subsequent Cue had an on average weaker projection onto the WMCD (Fig 19). Although the ability of the WMCD to predict correct behavioural responses was present in almost all experiments, it was not able to predict the majority of incorrect responses, suggesting that mechanisms other than working memory encoding could also underlie incorrect responses (e.g., exploration).

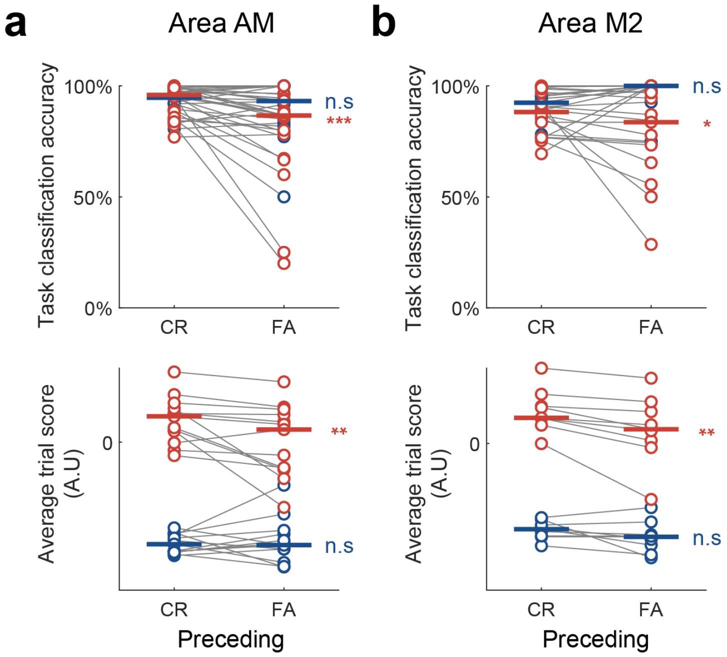


Figure 19 | Prediction of behavioural response from the WMCD. a, Top, task classification test accuracy from area AM delay activity projected onto the WMCD, for trials preceding a Correct Rejection (CR) response to the Cue subsequent to the delay, and for trials preceding a False Alarm (FA) to the Cue, split by task ($n = 18$ experiments; $p < 0.001$ for the Working Memory task, $p = 0.65$ for Discrimination task, signed-rank tests). Horizontal bars are medians. Bottom, raw projection scores for each experiment, positive values represent delay activities corresponding to the Working Memory task, and negative values corresponding to the Discrimination task ($p < 0.01$ for the Working Memory task, $p = 0.84$ for Discrimination task, signed-rank tests). **b**, As in **a**, but for area M2 ($n = 13$ experiments; task classification differences, $p < 0.001$ for the Working Memory task, $p = 0.07$ for the Discrimination task, signed-rank tests; average score differences, $p < 0.001$ for the Working Memory task, $p = 0.16$ for the Discrimination task, signed-rank tests).

Our results from Experiment 2 demonstrated that although the low-dimensional modes of delay activity were omni-present in both areas AM and M2 and exhibited strong persistent dynamics, they were completely agnostic of working memory engagement, which was instead supported by representations which were embedded in a highly variable, high-dimensional subspace of the neural population. This high-dimensional representation had the necessary properties of a latent mnemonic representation, it persisted throughout the delay it predicted the behaviour. Our observations regarding the trial-to-trial variability and dynamical structure of the WMCD has significant implications for mechanistic models of working memory[4,70], which will be discussed in Part III of this thesis.

Experiment 3

Maintenance of visual working memory by cortical feedback loops

Background

In the first two experiments of this thesis we observed that distributed regions of the neocortex support visual working memory, and that in at least two of these regions, AM and M2, there exists a high-dimensional population code underlying working memory. If these representations are as distributed as they are (i.e., redundant), why does silencing any single region lead to a disruption of the working memory trace? Theoretical work regarding how latent representations are maintained in the brain in cohesive states (recall Essay 2) has shown that they need not only to be distributed, but more importantly interdependent. Intuitively, any meaningful (e.g., invariant) influence that one area exerts on another, requires that they are constantly aware of what the other is doing. This is perhaps more a principle of feedback control in general, but is omnipresent in most instantiations of biologically plausible inference algorithms[58,105]. We therefore hypothesized that silencing one area locally (the 'feedforward' area) while imaging the feedback it receives from a reciprocally connected area, would selectively abolish the representation of working memory in the feedback. An alternative hypothesis exists, wherein the distributed brain regions operate completely independently, and contribute individual parts of the 'correct' task behaviour in parallel, and therefore have no influence on the other areas' representation of working memory.

Previous literature regarding the influence of feedback on representations of working memory is divergent, and largely suffers from the indetermination of working memory representations, as discussed in the introduction. For example, feedback has been implicated in the synchronization of distributed areas[106], modulatory influences on lateral intarecations[107], predictive processing[108], and learning[109,110]. Notable recent studies in mouse motor planning have identified a role for trans-colossal feedback in the maintenance of persistent activity during the delay[88], but whether such findings translate to working memory and other corticocortical connections in unknown.

Method

Animals and ethics

All experiments were carried out in accordance with institutional animal welfare guidelines and licensed by the UK Home Office. A total of 7 PV-cre mice of either sex were used. Mice were between 8 and 16 weeks old at the start of their experiments.

Surgical procedures

Prior to all surgeries, the mice were injected with dexamethasone (2–3 mg kg^{-1}) and an analgesic (carprofen 5 mg kg^{-1}). General anaesthesia was induced with 3% isoflurane which was then reduced to maintain a breathing rate of around 1 Hz. A first surgery to implant a custom stainless steel headplate was performed. The headplate was attached to the skull using dental cement (C&B Super Bond). The exposed skull was then sealed with a thin layer of light-curing dental composite (Tetric EvoFlow). Following a minimum recovery time of 3 days and intrinsic imaging to identify area AM (same as in Experiment 1), a second surgery was performed to make a cranial window over either area AM or M2 and perform viral injections. In 3 mice, a 3 mm diameter craniotomy was made centred around area AM, and a smaller <1 mm diameter craniotomy was made over area M2 (identified with coordinates relative to bregma; 0.5 mm lateral, 2.5 mm anterior). 100 nl viral injections of [rAAV1/Syn-Flex-ChrimsonR-tdT] and [AAV1/Syn-jGCaMP7b-WPRE], diluted in cortex buffer, were then made into areas AM and M2, respectively, with a Nanoject III microinjector (Drummond Scientific). Immediately afterwards, the larger area AM craniotomy was sealed with a 3 mm glass window. In the other 4 mice, the same procedure was done but with areas AM and M2 reversed.

Optogenetic silencing

The behavioural training and task methods were identical to those described in Experiment 2. In 15% of trials, we optogenetically silenced either area AM ($n = 4$) or M2 ($n = 3$). Optogenetic silencing was achieved by stimulating the PV/ChrimsonR+ cells immediately underneath the imaging site. A 637 nm laser (OBIS, Coherent) was relayed via a 400 µm diameter optical fibre to a 100 mm focal length lens which then focused the light onto the back aperture of the objective. The laser power was 6 mW for 400 ms immediately following the onset of the silencing delay period and then ramped down linearly to 0 mW over the next 200 ms. The stimulation light was pulsed at 60 Hz. The optogenetic laser and the visual stimulation display were blanked (turned off) in counter-phase with the resonant scanner turnaround times (12 kHz), so as to avoid light spill-through during ongoing imaging frame acquisition.

Imaging

The 2-photon imaging of axonal calcium signals was done on a custom-built microscope (SpectraPhysics MaiTai DeepSee at 930 nm, Nikon 16x objective). We acquired two planes 25 µm apart in layer 1, with a field of view of 400 × 400 µm at a frame rate of 22.78 Hz. Two cameras (22BUC03, ImagingSource) were used to record the pupil and body positions at 30 Hz. For each imaging site, we also recorded a volumetric image stack to confirm the location of tdTomato-ChrimsonR transfected PV+ cells immediately underneath the recorded axons.

Data analysis

The imaging data was first registered and pre-processed using a modified Suite2p pipeline[111]. The data was registered, boutons masks were extracted, and their calcium traces were baseline-subtracted. F_0 normalization was not performed due to the very low baseline fluorescence levels. Frames with low correlations to the registered average image or frames with significant eye movements were discarded as in Experiment 2. The boutons' time series data was then clustered into putative axons using custom scripts written in Matlab (MathWorks). Briefly, we used independent component analysis to extract a 40-dimensional *temporal* feature space from the full dimensional timeseries. The activity of all boutons, projected into this feature space, was then clustered using a Gaussian mixture model. The number of clusters was chosen by minimizing an adjusted Akaike information criterion error. Boutons with significant distances from their allocated cluster centre (less than 0.95 of their posterior probability given the cluster) were not clustered, and all others were clustered together by simply averaging their signals. This clustering procedure was then iterated a second time for all of the unclustered boutons. This clustering analysis returned the timeseries of putative axons, each averaged from roughly 10 boutons. We restricted all further analyses to axons which had a significant amount of delay-evoked activity, defined as 0.2 z-scored ΔF more in any one second of the delay than the last second of the preceding stimulus (i.e., post-pre) with a 0.01 α significance difference.

Results

In order to examine the role of feedback in maintaining working memory representations, we imaged the axons of area AM or M2, that projected to the reciprocally connected area (M2 or AM), while silencing the target area. This experiment identifies the representational component within the feedback connection which is contributed to by the target area's activity, although it does not distinguish whether the target area's influence arrives through corticocortical connections or not. We chose to silence at the onset of the delay period, for a brief window of time (400 ms plus a 200 ms ramp down, as in Experiment 1), because it had a highly selective effect on the Working Memory task (i.e., no effect on behaviour in the Discrimination task in Experiment 1), and would allow us to examine the dynamics of persistent activity following the silencing. Importantly, running speeds and pupil diameters were similar between the Discrimination and Working Memory tasks (as in Experiments 1 and 2; Fig. 20a-b), and the optogenetic silencing had no significant effects on either running speed or pupil diameter in either task, with no significant interaction effects (Fig. 20c-d).

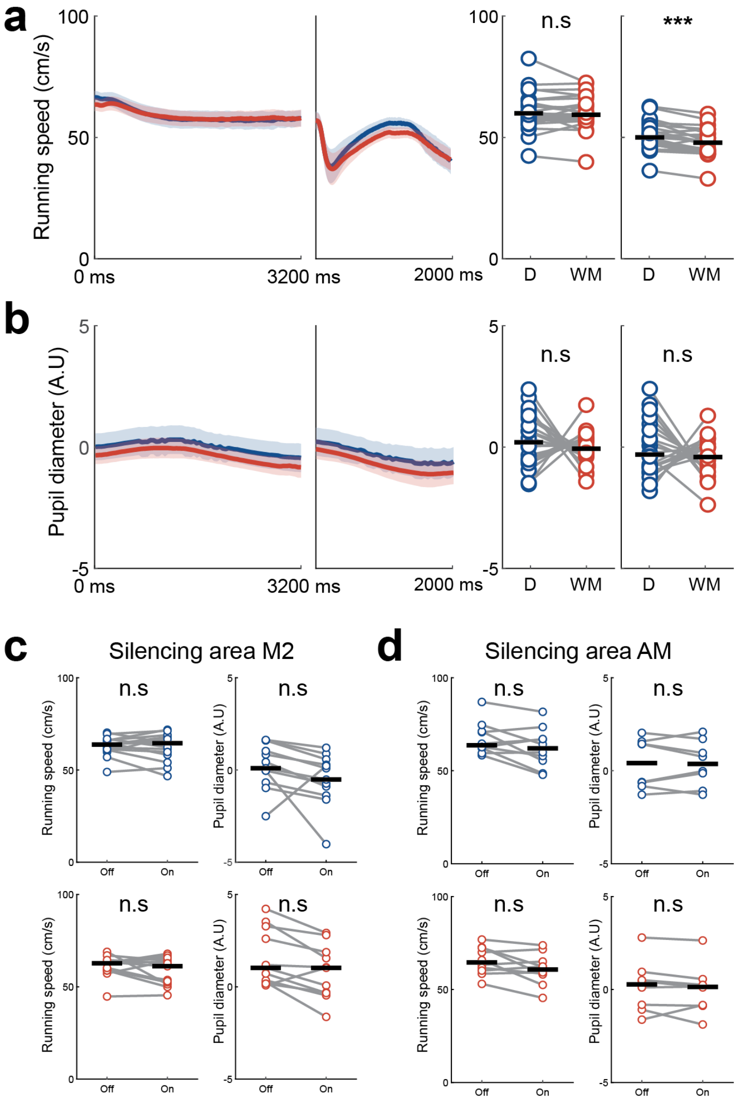


Figure 20 | Movement and arousal controls for optogenetic silencing. a, Left, running speeds for the Discrimination (blue) and Working Memory (red) tasks, averaged across sessions, for Experiment 3 (*n* = 25 sessions), during the delay and stimulus periods. Binned at 22.39 Hz, shaded regions are ± 95% CI. Right, running speed differences between tasks, averaged over the full duration of each delay (0 to 3,200 ms) and stimulus (0 to 2,000 ms) periods, split by experiment, $p = 0.72$ and $p < 0.001$, for the delay and stimulus periods, respectively, signed-rank test. Horizontal bars are medians. **b**, Left, pupil diameters for the Discrimination (blue) and Working memory (red) tasks, averaged across sessions, for Experiment 3 (*n* = 21 sessions), during the delay and stimulus periods. Binned at 22.39 Hz, shaded regions are ± 95% CI. Measured diameters were mean-subtracted prior to comparison between animals in order to better ascertain diameter changes (see methods). Right, pupil diameter differences between tasks, averaged over the full duration of each delay (0 to 3200 ms) and stimulus (0 to 2,000 ms) period, split by animal, $p = 0.48$ and $p = 0.39$, for the delay and stimulus periods, respectively, signed-rank test. Horizontal bars are medians. **c**, Effect of optogenetic silencing of area M2 at the onset of the delay on running speed (left) and pupil diameter (right) in the Discrimination (top) and Working Memory (bottom) tasks. A two-way ANOVA with task and silencing conditions found no significant effect of silencing, task, or interaction, for running speed and pupil diameter (*n* = 17 experiments; $p = 0.56$, 0.10, and 0.93, and $p = 0.53$, 0.86, and 0.89, respectively). Significance threshold was adjusted with Bonferroni correction (α = 0.006, 8 comparisons, Bonferroni correction). **d**, same as in **c**, but silencing area AM. A two-way ANOVA with task and silencing conditions found no significant effect of silencing, task, or interaction, for running speed and pupil diameter (n = 28 experiments; $p =$ 0.07 0.35, and 0.51, and $p = 0.006$, 0.36, and 0.89, respectively).

Agnostic of the task that the mice were engaged in, our silencing paradigm had diverse effects on individual delay responsive axons, with roughly equal amounts of inhibition and excitation (Fig. 21a). Compared to a null distribution identified by bootstrapping the control trials, we found that 26% of area AM axons were significantly influenced by silencing area M2, and 28% of area M2 axons by silencing area AM. The total net influence was neither inhibitory nor excitatory (Fig. 21b), as perhaps expected from previous reports of ‘modulatory’ feedback influences and *in vitro* physiological characterizations[112]. Furthermore, the magnitudes of these effects were similar across tasks, suggesting at the ‘functional connectivity’ level there was no strong influence of working memory engagement.

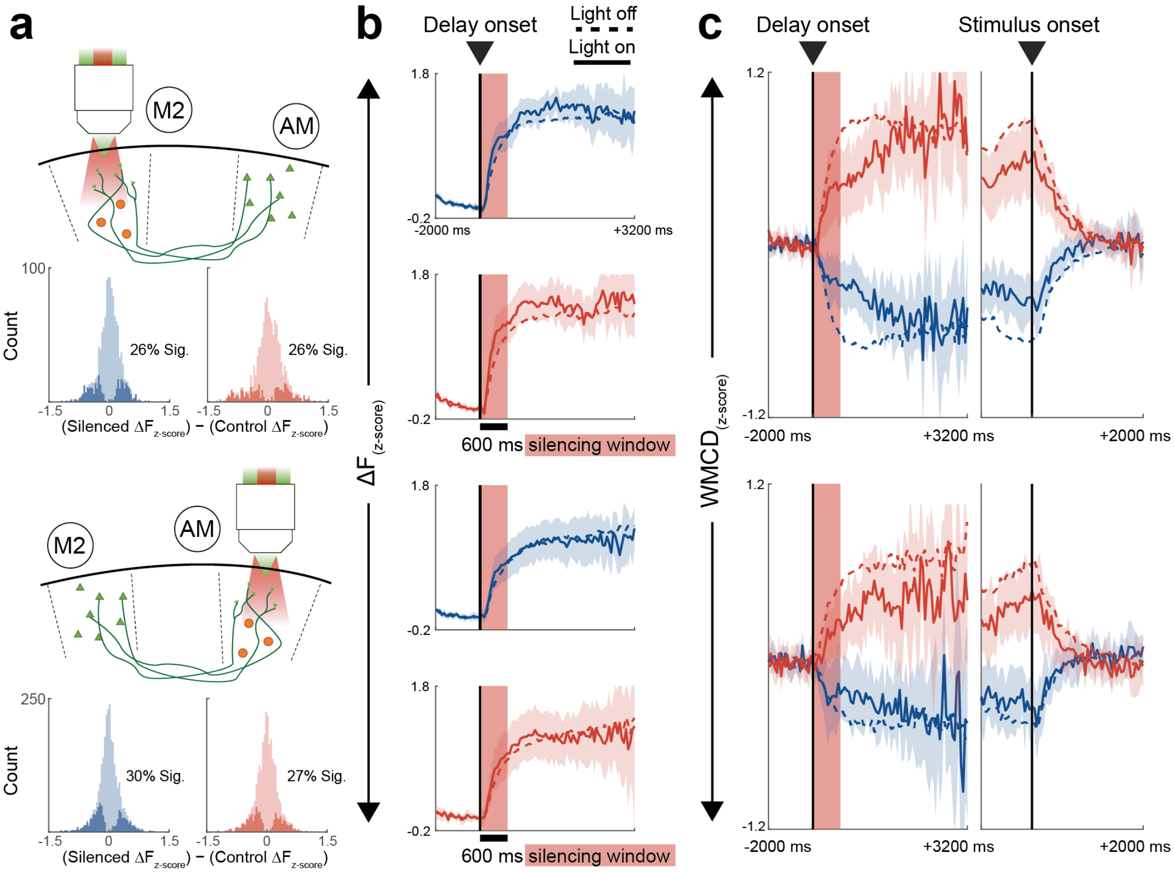


Figure 21 | Effect of breaking the cortical feedback loop on visual working memory. a, Schematics of simultaneous axonal imaging and silencing experiments, and overview of corresponding neural activity effects. Top two rows, experiments where area M2 was silenced, and axons of area AM were imaged in M2 (*n* = 21), and histograms of the difference in trial-averaged and z-scored activity during the delay of delay-responsive axons (see methods) in the control and silenced trials of the Discrimination (left, blue) and Working Memory (right, red) tasks. Percentage of significantly influenced axons shown with no multiple comparisons correction (t-tests across trials). Bottom two rows, same as above but experiments where area AM was silenced and axons of area M2 were imaged in AM (*n* = 12). Note that the silencing effects on the delay-responsive axons are centred around zero, symmetric, and similar between the two tasks. **b**, Trial-averaged activity of all delay-responsive axons, z-scored and averaged across experiments, in the Discrimination and Working Memory tasks, over the duration of the delay. Dashed lines correspond to delay activity in control trials, solid lines to activity in silenced trials. Data binned to 22.78 Hz, red background shading at the onset of the delay represents the silencing time window (600 ms), shaded regions around solid lines are the ±95% CIs of activity in silenced trials. **c**. Activity shown as in **b**, but projected onto the WMCD for each experiment, z-scored, and averaged across experiments. Statistical analyses follow in Figure 22.

We then identified the WMCD of the population activity in the feedback axons in the same manner as in Experiment 2, and projected the delay responses onto it. In accordance to our hypothesis, this high-dimensional representation of working memory engagement was selectively inhibited by disrupting the cortical feedback (Fig. 21c). For the WMCD activity in area M2 axons, this selective silencing effect occurred both during the Discrimination and Working Memory tasks, but in the area AM axons it was significantly stronger in the Working Memory task. This suggested that the two areas, although sharing very similar representations (as identified in Experiment 2), may have functionally distinct roles in working memory maintenance.

We next analysed the ability of these feedback population representations of working memory to recover, over the course of the delay, following the optogenetic silencing of their targets, by comparing the activity in the first and second halves of the delay period of sufficiently long delay duration trials. We found that area AM silencing induced a reduction in the WMCD activity which remained relatively unchanged over the duration of the delay (did not recover), while area M2 silencing recovered within roughly 1 second following the offset of the silencing (Fig. 22a,c).

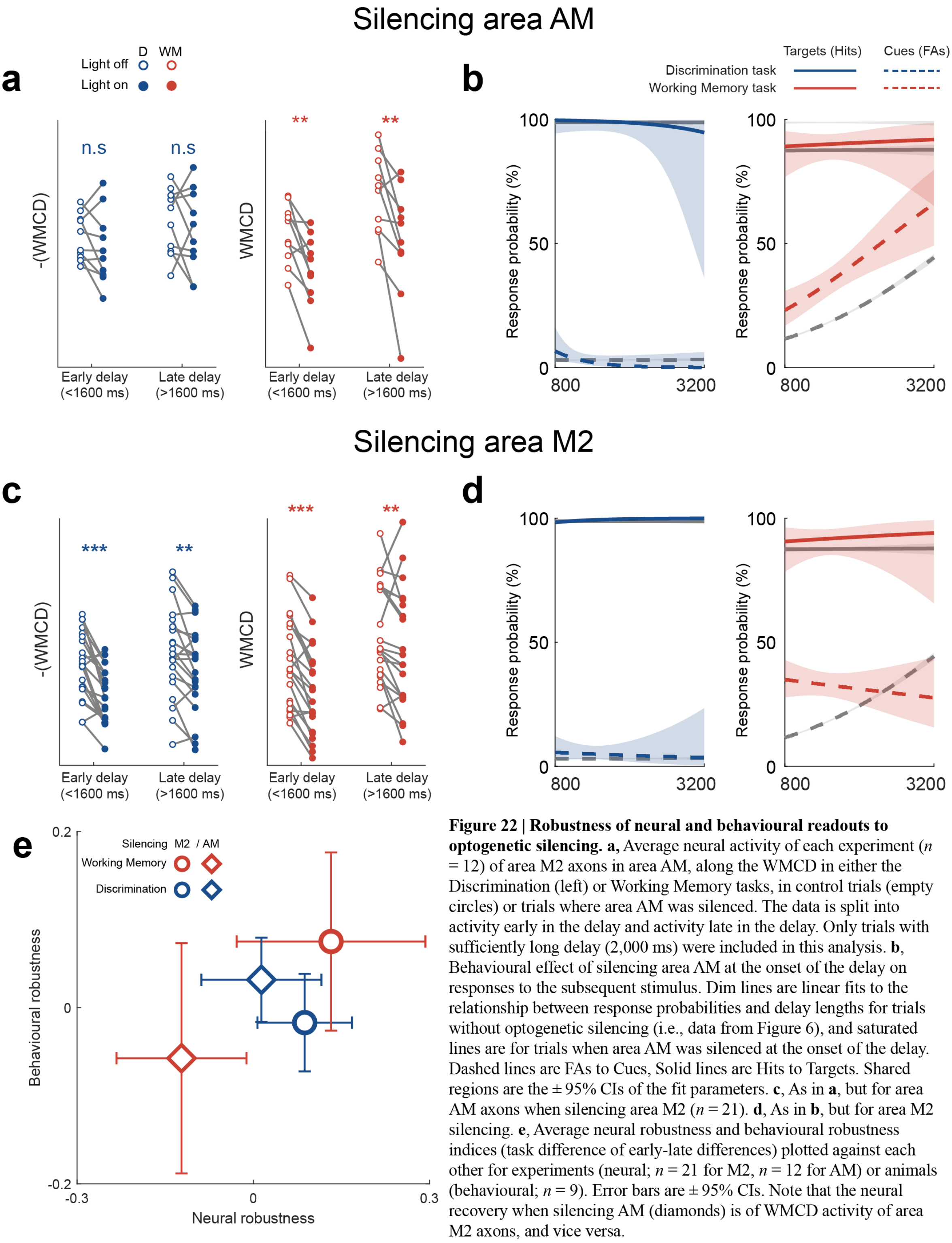


Figure 22 | Robustness of neural and behavioural readouts to optogenetic silencing. a, Average neural activity of each experiment (*n* = 12) of area M2 axons in area AM, along the WMCD in either the Discrimination (left) or Working Memory tasks, in control trials (empty circles) or trials where area AM was silenced. The data is split into activity early in the delay and activity late in the delay. Only trials with sufficiently long delay (2,000 ms) were included in this analysis. **b**, Behavioural effect of silencing area AM at the onset of the delay on responses to the subsequent stimulus. Dim lines are linear fits to the relationship between response probabilities and delay lengths for trials without optogenetic silencing (i.e., data from Figure 6), and saturated lines are for trials when area AM was silenced at the onset of the delay. Dashed lines are FAs to Cues, Solid lines are Hits to Targets. Shared regions are the ± 95% CIs of the fit parameters. **c**, As in **a**, but for area AM axons when silencing area M2 (*n* = 21). **d**, As in **b**, but for area M2 silencing. **e**, Average neural robustness and behavioural robustness indices (task difference of early-late differences) plotted against each other for experiments (neural; *n* = 21 for M2, *n* = 12 for AM) or animals (behavioural; *n* = 9). Error bars are ± 95% CIs. Note that the neural recovery when silencing AM (diamonds) is of WMCD activity of area M2 axons, and vice versa.

We hypothesized that this difference between areas AM and M2 in the robustness of the WMCD would be reflected in *behavioural* recovery to such perturbation over the course of the delay. For this analysis, due to limitations of statistical power for assessing response differences following silencing, we pooled together silencing data from Experiment 1, and found that such differences existed. While silencing area AM at the onset of the delay led to an increase in incorrect responses to the subsequent Cue (Fig. 22c), even if the delay duration lasted for up to 3,200 ms, silencing area M2 had a more transient behavioural effect, with the performance going back to baseline levels if the delay duration was sufficiently long (Fig. 22d). Note that the behavioural effect of area M2 silencing not a motor-related effect, the increase in incorrect responses was still restricted to Cues in the Working Memory task, and responses to the Probes in the Working Memory task likewise did not increase (Experiment 1). We defined a recovery index for both the behavioural response differences induces by silencing as well as the neural representations of the WMCD, and summarized these results in Fig 22e – feedback representations of working memory and behavioural readouts of working memory were not robust to (i.e., did not recover from) area AM silencing, and the opposite was true for area M2 silencing.

Part III – Discussion

Cognition in mice

We found that mice could readily switch between performing a simple visual discrimination task and a working memory dependent delayed-(non)match-to-sample task. A very important observation which augments the interpretation of our experiments was that several task-extrinsic variables, such as movement and arousal (proxied by pupil dilation), were not significantly different between the two tasks. This demonstrates that the relatively straightforward controls of having equal reward probabilities and delay-stimulus timings (specifically flat hazard rate delay durations) are sufficient to constrain behaviours such as motor planning and reward expectation. This is critical, as such factors have previously been shown to have enormous impacts on neural activity[113,114], and could potentially mask any finer psychometric readouts of behaviour[91]. Once these variables have been controlled, we were able to identify clear psychometric differences between the two tasks, which had clear interpretations. First and foremost, there was a linear inter-stimulus delay length effect for memory-guided responses (to Cues in the Working Memory task), which was completely absent for Probes in the same task. Previous literature of visual working memory retention duration for stimulus orientation has reported similar linear relationships and times[115,116], suggesting that a common mechanism may support working memory traces in mice and humans. We also observed small but systematic reaction time increases as a result of working memory engagement. Unfortunately similar dual-task studies have not been carried out in humans for comparison. There are likely a large number of psychometric differences which we have not yet analysed, specifically regarding the upkeep of working memory by the presentation of variable numbers of repeated Cues, and the disrupting effects of Probes, which are beyond the scope of this thesis. A final interesting behavioural observation in Experiments 2 and 3 was the ease with which mice could adjust to rotations of the stimuli with no prior training. This underscores the fact that sensory systems do not rely on simple stimulus-response associations to carry out perceptual decisions.

Necessity of distributed cortical areas for visual working memory

Transient optogenetic silencing at the onset of the delay revealed a very selective role of distributed cortical regions in maintaining the working memory trace, as the silencing had no effect when the mouse was doing the simple discrimination task, and had no effect on the responses to the Probe stimulus. This is likely due to the fact that the silencing window was outside of any possible response window (i.e., earlier than the minimum delay duration), and therefore did not interfere with neural processes associated with motor planning. Silencing during the stimulus found relatively straightforward effects when the mice were performing the simple discrimination task, contralateral visual cortex silencing increased misses and false alarms, as expected from producing some form of visual scotoma. When the mice were engaged in a working memory task, however, we not only observed a greater increase in the miss rate, but also a paradoxical *decrease* in false alarm rate. This observation does not lead to simple explanations, and has not been reported previously – it simply seems that the functional role of the contralateral visual cortex changes to be more motor oriented during working memory engagement. Or, conversely, subcortical visual areas, such as the optic tectum, may support motor responses *only* during the simple discrimination task. This finding should be investigated further with simultaneous recordings of cortical and subcortical visual areas.

Representation of visual working memory

One immediate result from our imaging experiments was that the neural activity patterns in areas AM and M2 were strikingly similar. Both regions had substantial populations of delay onset and stimulus onset locked single cell responses, and neither region had cells which were clearly locked to both onsets. This is different to studies in the primate prefrontal or parietal cortices which often find mixed tuning cells that respond to both the delay and stimulus onsets, at least in parametric working memory tasks[67], but could potentially be reconciled by the fact the majority of *all* cells we recorded were neither clearly delay nor stimulus locked. Such cells were no less useful for decoding the task, and were analysed through dimensionality reduction or decoding. Other properties of this 'non-triggerable' population will require significantly more sophisticated analyses in the future in order to identify their dynamics.

The most surprising result from our experiments was that for many measures of low-dimensional neural activity patterns, in both areas AM and M2, there was no detectable relationship with working memory engagement, even during the delay period. These measures included average activity levels, sequential activity patterns, and the trial-averaged dynamics explaining roughly 80% of the variance. These activity patterns could still be related to motor planning or reward expectation, which were present in both of our tasks (and, importantly, *all other* decision making tasks). This result warrants a critical reevaluation of three types of studies. First, many earlier studies in head-fixed mice which studied short term memory, in Go/No-go or two-alternative-choice tasks, that didn't *necessitate* mnemonic stimulus representations (i.e., could be solved by motor planning), have often ascribed to sequential or low-dimensional dynamics mnemonic roles[11,12,81,117,118]. These functions should be re-evaluated as movement or reward related. Furthermore, several studies have used such 'delayed-response' designs in mice as models of working memory[13,80]; the validity of their results for working memory are questionable. Second, in population analyses of neural activity it is sometimes assumed that low-dimensional projections of neural activity, specifically when capturing dynamical modes which carry the most variance, aid in 'de-noising' the underlying activity patterns[75,102], and that these modes carry the relevant task information[76]. Such methods, when applied to our data, led to the selective elimination of working memory information, and should therefore be more substantiated. Finally, optogenetic silencing is often used to assert that the neural representations under question are necessary for the behavioural effects of silencing. Our results from Experiments 1 and 3, at least for the delay onset silencing, are clear examples of when the removal of the dominant dynamical mode of activity is *not* the effect leading to any behavioural changes. At least in our case, off-target (i.e., inter-areal) effects on high-dimensional representations were driving the changes we saw in behaviour (see next section).

The distributed high-dimensional representations we observed for working memory are consistent with some recent findings in the primate literature of sparse, highly variable activity patterns underlying working memory[14,15]. These studies have led to significant debates on the role of persistent low-dimensional activity for working memory[4,5]. Although our findings clearly support one side of this debate, our observation of more robust delay activity patterns along the working memory coding dimension *during* the Working Memory task does imply that some dynamical modes (suggested from our data to be line attractors aligned to the working memory coding dimension) do carry working memory information, and are simply masked by much more dominant, working memory independent, 'rotational' modes of activity.

The key limitation of our experiments and analyses relative to those in the primate delayed-match-to-sample literature is that we used a decoding approach to identify our working memory representations – we needed 'labelled' data of task-isolated working memory (i.e., the Discrimination task as a control) in order to identify our working memory representations. The development of future analyses to identify whether there are any differences in the dynamics of the population activity with and without working memory engagement, potentially building off of our observations of the difference in robustness of the coding dimension delay activity, would enable clearer characterization of working memory representations in tasks without such careful motor planning and reward controls.

Robustness of neural representations to feedback silencing

Using a combination of optical tools available in mice, namely optogenetic silencing and axonal recordings, we provided some of the first evidence for a previously hypothesized circuit motif underlying the maintenance of latent visual working memory representations. The fact that dominant, cell-averaged persistent activity modes were not clearly inhibited or excited suggests that their function could be more modular within the neocortex. One of the few differences between areas AM and M2 which we observed was in their robustness to neural perturbation. Recent studies have investigated the robustness of premotor circuits to unilateral optogenetic inhibition and found that trans-colossal input allows for such population dynamics to recover. This is a perfectly reasonable hypothesis for our results as well, and our data further suggests that such a mechanism is weaker, or different, in posterior areas of the neocortex. Importantly, because we were recording axonal signals of a distal cortical area, we can further identify that it is not a property of the dynamics of the area being silenced that lead to robustness per se (i.e., it was area AM representations which were robust to area M2 silencing), but solely a property of the silenced area’s physiology.

Concluding remarks

We found a few unexpected results which we have ignored in order to answer more immediate questions regarding the maintenance of working memory, and they should be investigated further. Namely, the dissociated function of the visual cortex during working memory guided action, evidence for a distinct dynamical regime for working memory coding representations, and the differences in robustness of areas AM and M2. As our experiments were not targeted to explore these phenomena, further investigations will need to be carried out in order to get a closer mechanistic understanding of working memory.

Works cited

1. Helmholtz, H. von. *Handbuch der physiologischen Optik.* (Leipzig : Leopold Voss, 1867).
2. Gold, J. I. & Shadlen, M. N. The Neural Basis of Decision Making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
3. Baddeley, A. Working memory. *Curr. Biol. CB* **20**, R136-140 (2010).
4. Lundqvist, M., Herman, P. & Miller, E. K. Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *J. Neurosci. Off. J. Soc. Neurosci.* **38**, 7013–7019 (2018).
5. Constantinidis, C. *et al.* Persistent Spiking Activity Underlies Working Memory. *J. Neurosci.* **38**, 7020–7028 (2018).
6. Zylberberg, J. & Strowbridge, B. W. Mechanisms of Persistent Activity in Cortical Circuits: Possible Neural Substrates for Working Memory. *Annu. Rev. Neurosci.* **40**, 603–627 (2017).
7. Druckmann, S. & Chklovskii, D. B. Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol. CB* **22**, 2095–2103 (2012).
8. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic Theory of Working Memory. *Science* **319**, 1543–1546 (2008).
9. Dotson, N. M. Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. *Neuron* **25** (2018).
10. Smith, E. E. & Jonides, J. Working memory: a view from neuroimaging. *Cognit. Psychol.* **33**, 5–42 (1997).
11. Li, N., Chen, T.-W., Guo, Z. V., Gerfen, C. R. & Svoboda, K. A motor cortex circuit for motor planning and movement. *Nature* **519**, 51–56 (2015).
12. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
13. Park, J. C., Bae, J. W., Kim, J. & Jung, M. W. Dynamically changing neuronal activity supporting working memory for predictable and unpredictable durations. *Sci. Rep.* **9**, 15512 (2019).
14. Lundqvist, M. *et al.* Gamma and Beta Bursts Underlie Working Memory. *Neuron* **90**, 152–164 (2016).
15. Shafi, M. *et al.* Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* **146**, 1082–1108 (2007).
16. Mumford, D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* **66**, 241–251 (1992).
17. Baddeley, A., Logie, R., Bressi, S., Sala, S. D. & Spinnler, H. Dementia and working memory. *Q. J. Exp. Psychol. Sect. A* **38**, 603–618 (1986).
18. Goldman-Rakic, P. S. Working memory dysfunction in schizophrenia. *J. Neuropsychiatry Clin. Neurosci.* **6**, 348–357 (1994).
19. Chai, W. J., Abd Hamid, A. I. & Abdullah, J. M. Working Memory From the Psychological and Neurosciences Perspectives: A Review. *Front. Psychol.* **9**, (2018).
20. Baddeley, A. Working memory: looking back and looking forward. Nat Rev Neurosci 4, 829–839 (2003).

21. D'Esposito, M. & Postle, B. R. The Cognitive Neuroscience of Working Memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).
22. Fuster, J. M. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *J. Neurophysiol.* **36**, 61–78 (1973).
23. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
24. Postle, B. R. Working memory as an emergent property of the mind and brain. *Neuroscience* **139**, 23–38 (2006).
25. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
26. Tsotsos, J. Analyzing Vision at the Complexity Level. *Behav. Brain Sci.* **13**, (1990).
27. Carruthers, P. Evolution of working memory. *Proc. Natl. Acad. Sci.* **110**, 10371–10378 (2013).
28. Hahn, L. A. & Rose, J. Working Memory as an Indicator for Comparative Cognition – Detecting Qualitative and Quantitative Differences. *Front. Psychol.* **11**, (2020).
29. Baddeley, A. Working Memory: Theories, Models, and Controversies. *Annu. Rev. Psychol.* **63**, 1–29 (2012).
30. Kang, M.-S., Hong, S. W., Blake, R. & Woodman, G. F. Visual working memory contaminates perception. *Psychon. Bull. Rev.* **18**, 860–869 (2011).
31. Woodman, G. F. & Luck, S. J. Interactions between perception and working memory during visual search. *J. Vis.* **2**, 732–732 (2002).
32. Teng, C. & Kravitz, D. J. Visual working memory directly alters perception. *Nat. Hum. Behav.* **3**, 827–836 (2019).
33. Potter, M. C. Meaning in visual search. *Science* **187**, 965–966 (1975).
34. Rakic, P. Evolution of the neocortex: Perspective from developmental biology. *Nat. Rev. Neurosci.* **10**, 724–735 (2009).
35. Meynert. (1867).

103. Campbell, A. (1905). Histological studies on the localisation of cerebral function. Cambridge, University Press.

37. Brodmann.Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues (1909).
38. Wernicke. (1878).

103. Holmes, G. (1918). Disturbances of Vision by Cerebral Lesions, 156(3), 452–454.

40. Adrian, E. D. & Zotterman, Y. The impulses produced by sensory nerve-endings. *J. Physiol.* **61**, 151–171 (1926).
41. Bartley, S. H. & Bishop, Geo. H. The cortical response to stimulation of the optic nerve in the rabbit. *Am. J. Physiol.-Leg. Content* **103**, 159–172 (1932).
42. Bard, P. Studies on the Cortical Representation of Somatic Sensibility. *Bull. N. Y. Acad. Med.* **14**, 585–607 (1938).
43. Penfield, W. & Jasper, H. *Epilepsy and the functional anatomy of the human brain*. xv, 896 (Little, Brown & Co., 1954).
44. S.A. Talbot, W.H. Marshall. Binocular interaction and excitability cycles in cat and monkey. Proc. Amer. Physiol. Soc. (1941), pp. 279-280
45. Grether, W. F. (1941). Comparative visual acuity thresholds in terms of retinal image widths. Journal of Comparative Psychology, 31(1), 23–33.
46. Daniel, P. M. & Whitteridge, D. The representation of the visual field on the cerebral cortex in monkeys. *J. Physiol.* **159**, 203–221 (1961).

47. Mountcastle, V. B. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.* **20**, 408–434 (1957).

48. Rockel, A. J., Hiorns, R. W. & Powell, T. P. The basic uniformity in structure of the neocortex. *Brain J. Neurol.* **103**, 221–244 (1980).

49. Rockland, K. S. & Pandya, D. N. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res.* **179**, 3–20 (1979).

50. Maunsell, J. H. & van Essen, D. C. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci. Off. J. Soc. Neurosci.* **3**, 2563–2586 (1983).

51. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex N. Y. N 1991* **1**, 1–47 (1991).

52. Gămănuţ, R. *et al.* The Mouse Cortical Connectome Characterized by an Ultra Dense Cortical Graph Maintains Specificity by Distinct Connectivity Profiles. *Neuron* **97**, 698-715.e10 (2018).

53. Shipp, S. Structure and function of the cerebral cortex. *Curr. Biol. CB* **17**, R443-449 (2007).

54. Friston, K. Learning and inference in the brain. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **16**, 1325–1352 (2003).

55. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

56. Koffka, K. *Principles of Gestalt psychology*. 720 (Harcourt, Brace, 1935).

57. Ullman, S. Visual routines. *Cognition* **18**, 97–159 (1984).

58. Olshausen, B. A., Anderson, C. H. & Essen, D. V. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719 (1993).

59. Borst, G., Ganis, G., Thompson, W. L. & Kosslyn, S. M. Representations in mental imagery and working memory: Evidence from different types of visual masks. *Mem. Cognit.* **40**, 204–217 (2012).

60. Buschman, T. J. & Miller, E. K. Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science* **315**, 1860–1862 (2007).

61. Scocchia, L., Valsecchi, M. & Triesch, J. Top-down influences on ambiguous perception: the role of stable and transient states of the observer. *Front. Hum. Neurosci.* **8**, (2014).

62. Ashourian, P. & Loewenstein, Y. Bayesian Inference Underlies the Contraction Bias in Delayed Comparison Tasks. *PLoS* (2011).

63. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The Helmholtz machine. *Neural Comput.* **7**, 889–904 (1995).

64. Hedayati, S., O'Donnell, R. & Wyble, B. The Memory for Latent Representations: An Account of Working Memory that Builds on Visual Knowledge for Efficient and Detailed Visual Representations. *bioRxiv* 2021.02.07.430171 (2021) doi:10.1101/2021.02.07.430171.

65. Brody, C., Zainos, A. & Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex N. Y. N 1991* **13**, 1196–207 (2003).

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

67. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
68. Dotson, N. M., Hoffman, S. J., Goodell, B. & Gray, C. M. Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. *Neuron* **99**, 215-226.e4 (2018).
69. Zoltowski, D. M., Latimer, K. W., Yates, J. L., Huk, A. C. & Pillow, J. W. Discrete Stepping and Nonlinear Ramping Dynamics Underlie Spiking Responses of LIP Neurons during Decision-Making. *Neuron* **102**, 1249-1258.e10 (2019).
70. Stokes, M. & Spaak, E. The Importance of Single-Trial Analyses in Cognitive Neuroscience. *Trends Cogn. Sci.* **20**, 483–486 (2016).
71. Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *J. Neurophysiol.* **102**, 614–635 (2009).
72. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation Through Neural Population Dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
73. Sauerbrei, B. A. *et al.* Cortical pattern generation during dexterous movement is input-driven. *Nature* **577**, 386–391 (2020).
74. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).
75. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
76. Cueva, C. J. *et al.* Low-dimensional dynamics for working memory and time encoding. *Proc. Natl. Acad. Sci.* (2020) doi:10.1073/pnas.1915984117.
77. Tajima, S. *et al.* Task-dependent recurrent dynamics in visual cortex. *eLife* **6**, e26868 (2017).
78. Panichello, M. F. & Buschman, T. J. Selective control of working memory in prefrontal, parietal, and visual cortex. *bioRxiv* 2020.04.07.030718 (2020) doi:10.1101/2020.04.07.030718.
79. Teutsch, J. & Kätzel, D. Operant Assessment of DMTP Spatial Working Memory in Mice. *Front. Behav. Neurosci.* **13**, (2019).
80. Bolkan, S. S. *et al.* Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat. Neurosci.* **20**, 987–996 (2017).
81. Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810-824.e9 (2019).
82. Krumin, M., Lee, J. J., Harris, K. D. & Carandini, M. Decision and navigation in mouse parietal cortex. *eLife* **7**, e42583 (2018).
83. Willeke, K. F. *et al.* Memory-guided microsaccades. *Nat. Commun.* **10**, 3710 (2019).
84. Wu, Z. *et al.* Context-Dependent Decision Making in a Premotor Circuit. *Neuron* **106**, 316-328.e6 (2020).
85. Condylis, C. *et al.* Context-Dependent Sensory Processing across Primary and Secondary Somatosensory Cortex. *Neuron* **106**, 515-525.e5 (2020).
86. Liu, D. *et al.* Medial prefrontal activity during delay period contributes to learning of a working memory task. *Science* **346**, 458–463 (2014).
87. Pagan, M. & Rust, N. C. Dynamic Target Match Signals in Perirhinal Cortex Can Be Explained by Instantaneous Computations That Act on Dynamic Input from Inferotemporal Cortex. *J. Neurosci.* **34**, 11067–11084 (2014).

88. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).
89. Zhuang, J. *et al.* An extended retinotopic map of mouse cortex. *eLife* **6**, e18372 (2017).
90. Guo, Z. V. *et al.* Flow of cortical activity underlying a tactile decision in mice. *Neuron* **81**, 179–194 (2014).
91. Zariwala, H. A., Kepecs, A., Uchida, N., Hirokawa, J. & Mainen, Z. F. The Limits of Deliberation in a Perceptual Decision Task. *Neuron* **78**, 339–351 (2013).
92. Pertzov, Y., Manohar, S. & Husain, M. Rapid forgetting results from competition over time between items in visual working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 528–536 (2017).
93. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986-999.e16 (2017).
94. Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates, Compact - 5th Edition. https://www.elsevier.com/books/paxinos-and-franklins-the-mouse-brain-in-stereotaxic-coordinates-compact/franklin/978-0-12-816159-3.
95. Laubach, M., Amarante, L. M., Swanson, K. & White, S. R. What, If Anything, Is Rodent Prefrontal Cortex? *eNeuro* **5**, (2018).
96. Hayden, B. Y. & Gallant, J. L. Working memory and decision processes in visual area v4. *Front. Neurosci.* **7**, 18 (2013).
97. Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife* **5**, e14472 (2016).
98. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: Flexible software for operating laser scanning microscopes. *Biomed. Eng. OnLine* **2**, 13 (2003).
99. Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *eLife* **8**, e38173 (2019).
100. Vogelstein, J. T. *et al.* Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104**, 3691–3704 (2010).
101. Ganea, D. A. *et al.* Pupillary Dilations of Mice Performing a Vibrotactile Discrimination Task Reflect Task Engagement and Response Confidence. *Front. Behav. Neurosci.* **14**, 159 (2020).
102. Churchland, M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).
103. Harris, K.D., Nonsense correlations in neuroscience. bioRxiv (2020). https://www.biorxiv.org/content/10.1101/2020.11.29.402719v1.
104. Moreno-Bote, R. *et al.* Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
105. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**, 1434–1448 (2003).
106. Womelsdorf, T. & Fries, P. The role of neuronal synchronization in selective attention. *Curr. Opin. Neurobiol.* **17**, 154–160 (2007).
107. Liang, H. *et al.* Interactions between feedback and lateral connections in the primary visual cortex. *Proc. Natl. Acad. Sci.* (2017) doi:10.1073/pnas.1706183114.
108. Leinweber, M., Ward, D. R., Sobczak, J. M., Attinger, A. & Keller, G. B. A Sensorimotor Circuit in Mouse Cortex for Visual Flow Predictions. *Neuron* **95**, 1420-1432.e5 (2017).

109. Larkum, M. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* **36**, 141–151 (2013).

110. Roelfsema, P. R. & Holtmaat, A. Control of synaptic plasticity in deep cortical networks. *Nat. Rev. Neurosci.* **19**, 166–180 (2018).

111. Pachitariu M, Stringer C, Dipoppa M, et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. bioRxiv; 2016. DOI: 10.1101/061507.

112. Yang, W., Carrasquillo, Y., Hooks, B. M., Nerbonne, J. M. & Burkhalter, A. Distinct Balance of Excitation and Inhibition in an Interareal Feedforward and Feedback Circuit of Mouse Visual Cortex. *J. Neurosci.* **33**, 17373–17384 (2013).

113. Musall, S., Kaufman, M., Gluf, S. & Churchland, A. *Movement-related activity dominates cortex during sensory-guided decision making*. (2018). doi:10.1101/308288.

114. Larsen, R. S. & Waters, J. Neuromodulatory Correlates of Pupil Dilation. *Front. Neural Circuits* **12**, (2018).

115. Vogels, R. & Orban, G. A. Decision processes in visual discrimination of line orientation. *J. Exp. Psychol. Hum. Percept. Perform.* **12**, 115–132 (1986).

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

117. Gilad, A., Gallero-Salas, Y., Groos, D. & Helmchen, F. Behavioral Strategy Determines Frontal or Posterior Location of Short-Term Memory in Neocortex. *Neuron* **99**, 814-828.e7 (2018).

118. Lee, J. J., Krumin, M., Harris, K. D. & Carandini, M. Task specificity in mouse parietal cortex. *bioRxiv* 2020.12.18.423543 (2020) doi:10.1101/2020.12.18.423543.