HHS Public Access

Author manuscript

Nat Rev Neurosci. Author manuscript; available in PMC 2022 January 30.

Published in final edited form as:

Nat Rev Neurosci. 2019 August ; 20(8): 466–481. doi:10.1038/s41583-019-0176-7.

The what, where and how of delay activity

Kartik K. Sreenivasan[1,*], **Mark D'Esposito**[2,*]

[1]Division of Science and Mathematics, New York University Abu Dhabi, UAE.

[2]Helen Wills Neuroscience Institute and Department of Psychology, University of California, Berkeley, CA, USA.

Abstract

Working memory is characterized by neural activity that persists during the retention interval of delay tasks. Despite the ubiquity of this 'delay activity' across tasks, species and experimental techniques, our understanding of this phenomenon remains incomplete. Although initially there was a narrow focus on sustained activation in a small number of brain regions, methodological and analytical advances have allowed researchers to uncover previously unobserved forms of delay activity across the entire brain. In light of these new findings, this Review reconsiders what delay activity is, where in the brain it is found, what roles it serves and how it may be generated.

Introduction

To follow a conversation, you must mentally represent the overall topic, what was said in the last sentence and what you intend to say next. Critically, these representations need to be connected despite occurring seconds apart. This ability to link multiple events over brief intervals is essential for cognition and is the core feature of working memory (WM) — the set of operations that support the temporary retention of behaviourally relevant information. One of the enduring aims of neuroscience is to understand the neurobiology underlying WM[1].

In 1971, Fuster and Alexander[2] and Kubota and Niki[3] described the activity of individual neurons in the lateral prefrontal cortex (lPFC) of macaque monkeys performing a task that required storing information over a delay of several seconds and using this memory to guide a response. Strikingly, lPFC neurons remained active during the memory delay when no stimulus was present, bridging the temporal gap between perception and the contingent motor response. This phenomenon was later termed 'delay activity'[4], and is thought to reflect the sustained representation of WM content or WM-related goals[5]. Delay activity is seen more generally in contexts that require an organism to link a sensory stimulus to a delayed behaviour, such as during tasks of sustained attention[6] or decision making[7];

* kartik.sreenivasan@nyu.edu, despo@berkeley.edu.

Author contributions

Both authors researched data for article and made substantial contributions to the discussion of content. K.K.S. wrote the article and K.K.S. and M.D. reviewed or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

here, however, we examine delay activity in tasks explicitly testing WM, where it has been characterized most extensively.

In recent years, a wealth of theoretical and experimental work has prompted an expanded consideration of delay activity, its function and the mechanisms that generate it. This Review integrates findings across subfields of neuroscience to critically evaluate the what, where and how of delay activity. There are two key areas of emphasis. First, in contrast to discussions of delay activity that typically focus on increases in spike rate or functional MRI (fMRI) signal that persist throughout a WM delay, we consider delay activity to be any task-related change in neural activity that spans the interval between a stimulus and the behavioural response in a **WM delay task**. This expanded definition allows for a more inclusive and comprehensive evaluation of delay activity, but at the same time calls attention to the challenge of comparing data across different experimental methods, brain regions and species. We address this issue by highlighting studies that record delay activity across multiple brain regions in the same experiment. Second, we emphasize points of connection between empirical studies and neural models of delay activity.

What is delay activity?

Delay activity can be observed in many forms at various scales — from individual neurons, to cell assemblies comprised of tens of neurons, meso-scale interregional circuits or macro-level networks that recruit regions throughout the brain. Recording delay activity at these different scales requires multiple techniques, each of which is sensitive to different aspects of the underlying neuronal signals (see Box 1 for a discussion of non-neuronal contributions to delay activity) and has a unique set of limitations. By considering these many forms of delay activity (Fig. 1), we aim to provide a more complete picture of its purpose and origin.

Neuronal spike rate

Extracellular recordings have documented increases in spike rate (relative to a pre-trial baseline) that persist throughout WM delays; for brevity, this finding is referred to as 'delay spiking' (Fig. 1a). Delay spiking is found in rodent PFC[8,9]; in various regions in the brains of non-human primates (NHPs), including prefrontal, parietal and sensory cortices[2,3,10–14]; and in human medial temporal lobe (MTL)[15,16]. Notably, delay spiking persists beyond the tens of milliseconds over which a neuron integrates its synaptic inputs[17], indicating that it may involve complex network dynamics[18,19] or intrinsic cellular properties[20] that can sustain spiking in the absence of external stimulation (see below). Studying spike rates enables specific hypotheses about the cellular basis of delay activity to be tested (Box 2) and can provide insights into local network properties that support WM.

There are two caveats to this approach. First, although spiking studies can elucidate how individual neurons encode WM information, analyses of networks of neurons are increasingly considered to be more informative for understanding brain function[21,22]. Important information is also encoded in large-scale network activity[23]. Unfortunately, concurrent recording of spiking of meso- and macro-scale populations remains relatively rare (but see below). Second, the extensive task and stimulus training that animals undergo in such studies increases delay spiking in NHPs[24,25] and reduces the correlation between

delay spiking and behaviour in mice[26]. Thus, despite some evidence for sustained delay spiking in the absence of training or familiarity in the human MTL[15,16], the potential influence of task and stimulus familiarity should be considered when interpreting studies of delay spiking.

Extracellular field potentials

Extracellular field potentials (EFPs) represent neural activity summed over many neurons and can be measured intracranially with microelectrodes as local field potentials (LFPs); with electrodes on the surface of the brain using electrocorticography (ECoG); or with electrodes or sensors on the scalp using electroencephalography (EEG) or magnetoencephalography (MEG). In the delay period of WM tasks, sustained increases in EFP amplitude are found in human EEG[27,28] and NHP LFP[29] recordings. Moreover, enhancements in the oscillatory power of the EFP in the gamma band (40–100 Hz)[30–32], beta band (15–30 Hz)[33] and theta band (4–8 Hz)[34] are also sustained during WM delays. Interactions between oscillatory power in the beta and gamma frequency bands have been proposed to underlie top-down control of WM content[35]. Human studies have also reported delay-period increases[36,37] and decreases[38,39] in oscillatory power in the alpha band (8–12 Hz) over posterior electrodes. Alpha attenuation may be related to the maintenance of task-relevant stimuli, whereas alpha enhancement may reflect the inhibition of the processing of task-irrelevant stimuli[40]. In addition to sustained modulation, discrete bursts of gamma oscillations in NHP lPFC are thought facilitate the storage of WM information[41,42].

Despite controversy over the relative contributions of spiking and synaptic activity to the LFP signal[43] and the difficulty of localizing EEG or MEG signal[44], there are several features of EFPs that make them a valuable measure of delay activity. First, EFPs are straightforward to measure and compare across species. For example, one study noted that delay period EFP signals from NHP lPFC and human parietal EEG electrodes have very similar features[29]. Second, LFP power often predicts behaviour in NHPs as well as or better than single-neuron spike rates[45,46]. Third, EFPs measure micro-, meso- and macro-level phenomena, and thus may help to bridge our understanding of detailed stimulus-coding properties of neurons and the activity of large-scale brain networks. At the micro-level, LFPs coordinate local spiking activity[47–50], and different oscillatory frequencies are linked to unique cellular and local circuit mechanisms[51–53]. At the meso- and macro-levels, EFPs may coordinate activity over local and brain-wide networks[54] through cross-frequency synchrony[55,56].

BOLD signal amplitude

The blood-oxygenation-level-dependent (BOLD) signal measured by fMRI remains elevated above pre-trial baseline throughout the delay period of WM tasks in several brain regions[57–60] (Fig. 1a). The BOLD signal reflects influences from spiking and LFPs[61], but is most closely correlated with LFP power[62–64] (however, see Ref.[65]). Indeed, one study found a positive relationship between the amplitude of the BOLD signal and ECoG gamma-band power throughout the human cortex during WM delays[66]. Uncertainty surrounding the source of the BOLD signal limits its usefulness for examining local circuit mechanisms, and its sluggish temporal resolution (on the order of seconds) imposes constraints on experimental design (Box 3). However, the meso-level spatial scale and broad coverage

of BOLD fMRI can be leveraged to uncover otherwise unobservable features of brain representations[67]. For example, one study found that attention shifted, rather than enhanced, the population tuning of voxels in human visual cortex[68] — an outcome that could not have been anticipated from single-unit data. Further, fMRI allows for comparisons of activity across brain regions and enables the investigation of large-scale brain networks[23] that are crucial for our understanding of WM[69,70].

Population coding

In contrast to early studies that hypothesized that memories are sustained via the persistent activation of small populations of neurons that are highly tuned to the maintained stimulus[10], recent results suggest that WM relies on **population coding** (Fig. 1c). Population coding distributes information across neurons or neural populations with diverse tuning preferences, potentially rendering representations more resistant to noise in individual neurons by taking into account correlations between neurons. Population activity in the lPFC of NHPs during a delay represents task variables such as stimulus–stimulus associations[71] in addition to WM content[72]. In humans, population analyses of fMRI data have revealed sustained representations of WM content in the pattern of activity of **voxels** distributed within parietal and sensory cortices, even when the mean BOLD signal does not differ from baseline across the delay[73–77]. Similarly, location information maintained in WM can be read out from the pattern of EEG-recorded alpha-band power[78]. Importantly, many of these findings revealed by population analyses were undetectable using traditional univariate analysis techniques.

Dynamic delay activity

Delay activity was traditionally depicted as stable over a delay, but closer analyses reveal that its magnitude often varies over time. These temporal dynamics were largely overlooked owing to analyses that averaged delay activity over the delay or over trials[35,79] (but see Refs[8,80–82]). A thorough understanding of these temporal patterns may provide valuable insight into how delay activity encodes WM information[35,83,84].

Variations in spike rate.—Single-trial spike data from NHPs indicates that some neurons exhibit monotonic increases or decreases in spike rate over the delay (Fig. 1b). These patterns have been hypothesized to represent the transformation from retrospective WM stimulus representations to prospective action plans[85,86], the anticipation of an upcoming response[86,87], or a computation of elapsed time since encoding[88]. Other neurons have spike rates that vary widely without a clear pattern over the delay[81,82]. Indeed, temporal fluctuations are the rule, rather than the exception; one study of lPFC estimated that only 3% of delay-active neurons spike at a stable rate throughout the delay[81]. It is therefore important that neural models of WM can capture this temporal variation.

Sequential delay spiking.—In contrast to individual cells showing temporal variations in delay spiking, other cells show sequential patterns of activation, in which each cell in a population spikes briefly at a specific point during the delay[89,90]. If the population has diverse temporal tuning preferences, then the population activity can 'tile' the entire delay period. Such sequential delay period spiking is robustly observed in area CA1 of the

hippocampus[91,92], but a similar phenomenon — sequential, delay-spiking 'relay-race' cells — has also been documented in the PFC[93].

Dynamic population codes.—WM information, including stimulus identity and task variables, can also be stored in a dynamic population code (reviewed elsewhere[94]). In a dynamic code, the same piece of information can be encoded by different patterns of population activity at different times. These dynamics can be revealed by a cross-temporal analysis in which a classifier is trained on patterns of activity at each time point during the delay and tested separately on every other time point of the delay, resulting in a training time-by-testing time matrix of classification accuracy (Fig. 1d). Successful classification along the diagonal, where training and testing time points are the same, combined with chance performance off the diagonal, indicates that distinct activity patterns encode memory content at different points during the delay.

What is the purpose of a time-varying population code? One intuitive account is that the temporal dynamics facilitate the computation of elapsed time, which may be useful in anticipating an upcoming response[95]. The strongest evidence supporting this idea is the finding that neural populations in NHP lPFC that dynamically encoded the category of the maintained stimulus simultaneously encoded elapsed maintenance time[96]. Another intriguing possibility is that dynamic WM codes could be less prone to interference from external input than static codes, rendering WM more resistant to distraction[97].

Despite their potential advantages, dynamic codes pose an obstacle to the faithful readout of WM information by downstream brain regions because they preclude a unique mapping from a pattern of activity to the information stored in that pattern. This challenge may be overcome in several ways. First, downstream regions may dynamically modulate their readout weights to match the dynamic changes in coding in the upstream region. Although this is theoretically plausible, empirical support for this argument is scant. A second possibility is that neurons in downstream regions maintain static readout weights and integrate evidence over the entire WM delay[94]. In this case, the mapping from activity pattern to information would only be valid for a single point in time, meaning that the WM information would only be available to a subpopulation of downstream neurons at one point during the delay. The activity at other points during the delay would therefore be noise. A third alternative was suggested by a recent NHP study that found that dynamic population codes in lPFC contained a low-dimensional subspace that encoded information stably over time, enabling downstream neurons to assign a single set of readout weights to read out WM content throughout the delay[98] (see also Refs[84,99,100]). It is important to note that arguments for a stable subspace typically rely on analyses of WM tasks in which a single item is to be maintained, whereas dynamic population codes are more frequently noted in complex tasks with multiple memory items. How task complexity or other factors, such as cognitive state or the connectivity of neurons in a network, influence whether information is encoded in a dynamic or static population code is unclear.

LFP bursts.—Single-trial analyses of NHP LFP data have revealed irregular patterns of LFPs that contribute to WM. Lundqvist et al. isolated narrowband LFP bursts on individual trials: gamma bursts were associated with increased information or more items stored in

WM, whereas beta bursts seemed to inhibit gamma bursts and reduced the information content of spiking activity[41] (Fig. 1b). Moreover, the frequency of gamma bursts also increased when the monkey accessed WM information to compare a memory item with a visually presented probe[42]. Only after averaging activity over trials did the classic patterns of sustained gamma and beta enhancement emerge. A possible advantage of LFP bursts is that intermittent bursting activity may be metabolically more efficient and less prone to disruption than persistent spiking[35]. Some have argued against this interpretation of LFP burst data, suggesting that gamma bursts may actually reflect spiking[101] (see Ref.[102] for a response). Nevertheless, the idea that information may be sustained by discrete bursting activity seriously challenges the long-held notion that WM information is encoded through persistent neural firing.

Where and why?

Delay activity is found throughout the brain, and probably represents many functions in service of WM[103] (as well as other cognitive processes not discussed here). We therefore consider location and function together in this section, using a representative set of brain regions to illustrate the anatomical breadth and functional diversity of delay activity (Fig. 2; see Ref.[104] for a comprehensive list of regions displaying delay activity).

Multimodal association cortex

Lateral prefrontal cortex.—A longstanding debate is whether lPFC delay activity preferentially represents WM content, or goal-related information such as response rules and WM operations (for more targeted reviews on PFC delay activity, see Refs[104–108]). Extensive evidence for both points of view exists, primarily from NHP single-unit and human fMRI studies.

In support of the claim that lPFC stores WM content, lPFC delay activity is selectively tuned to specific low-level features of WM memoranda[109], such as spatial location[10,110,111] or direction of motion[112,113]. Another form of content-selective lPFC activity was revealed by a study showing that NHP lPFC neurons monotonically vary their spike rate in proportion to the frequency of a vibrotactile stimulus held in WM[81,114]. lPFC delay activity can also be selective for abstract information about WM content, such as associations between stimuli[71,115,116] or stimulus category information[116–118].

By contrast, the view that lPFC encodes goal information is supported by evidence that lPFC delay activity is sensitive to response rules[119,120], reward expectation[121,122] and elapsed maintenance time[96], and varies in magnitude with demands on cognitive operations, such as manipulation of WM information[123]. The findings that lPFC neurons can change their tuning mid-delay[71,84] and switch from encoding WM content to encoding reward expectation[122] are also more compatible with this latter proposal.

Attempts to reconcile these perspectives have given rise to a more nuanced view: that lPFC delay activity simultaneously represents multiple dimensions of WM content and goals[124,125]. The property of **nonlinear mixed selectivity** enables individual neurons to respond to combinations of task and stimulus features such that feature combinations cannot

be read out from the responses of the individual neuron alone[124] and instead are encoded in the combined activity of several neurons[124]. The lPFC uses population coding by nonlinear mixed-selectivity neurons to store high-dimensional representations that can be decoded as low-dimensional representations of goals or content by hierarchically lower brain regions. Nonlinear mixed selectivity may even allow lPFC neurons to recode information in response to intervening input, potentially increasing the robustness of WM representations[126].

A noteworthy caveat to this debate over the role of lPFC in WM is that it largely ignores the possibility that frontal cortex may be adapted to perform different WM functions in humans and non-human animals. For example, recent work has challenged the long-assumed homology between the frontal eye fields in NHPs and the superior precentral sulcus in humans, complicating direct comparison between these two regions[127,128]. Such findings underline the need for systematic comparisons of delay activity across species.

Posterior parietal cortex.—Posterior parietal cortex (PPC) delay activity has been observed in single-unit[11,12,129] and LFP[45] recordings in NHPs, and using fMRI[76,130–132] and ECoG[133] in humans. Human fMRI studies find that PPC delay activity increases with **WM load**[130], and NHP studies show that PPC delay spiking is selective for the specific content of WM[12]. Both sets of findings are consistent with the idea that PPC delay activity represents low-level stimulus information. Alternatively, evidence that PPC delay activity preferentially encodes visual category information (for example, cat versus dog) over visual item information (for example, which specific cat or dog) when category information is task-relevant suggests that PPC delay activity can represent abstract features of WM content[134]. A third possibility — that PPC delay activity represents attention directed internally to individual WM representations — has been suggested by human fMRI and lesion studies of verbal WM[135,136]. This function is reminiscent of the well-documented role of the PPC in shifts of external attention[137]. Interestingly, all three of these putative roles of PPC echo those attributed to lPFC, which is consistent with the high degree of reciprocal connectivity between the two regions. However, observed differences in morphology[138] and effective connectivity[139] between the lPFC and PPC suggest that these regions have distinct roles (reviewed in[140]). Direct attempts to disentangle the functions of delay activity in PFC and PPC have yielded mixed results[141–143], highlighting the need for experiments that disrupt processing in one of these regions while recording delay activity in the other.

Medial temporal lobe.—Although the MTL is closely associated with long-term memory, it also exhibits delay activity during WM[144], particularly under conditions that can be described as 'high demand'[145]. Indeed, delay-related BOLD signal in the MTL is most robust at the upper limits of individuals' **WM capacity**[146]. In addition, patterns of delay spiking in the human and NHP MTL encode information about complex images[15,16,147], which are more difficult to maintain than simple features[148,149]. A related possibility is that, rather than encoding individual memory items, MTL delay activity might encode associations between features or items during WM for complex items[150]. This possible function is analogous to the well-known role of the MTL in relational binding in long-term memory[151].

The MTL may also support WM for challenging novel information[152]; MTL delay activity is modulated by the novelty of WM content, exhibiting a greater BOLD signal during WM for novel versus familiar items, even when novelty is task-irrelevant[153]. Understanding the role of the MTL delay activity in the storage of simple, non-novel features will be crucial for determining whether the MTL supports WM maintenance in general or has an auxiliary role only when demand is high. Another alternative, motivated by the hippocampus's purported role in computing sequential change[154], is that hippocampal populations within the MTL compute elapsed time during WM delays. Consistent with this hypothesis, rodent CA1 cells exhibit stereotyped patterns of sequential firing during delay tasks[91,92].

Unimodal cortices

Motor association cortex.—Motor association areas in frontal cortex are increasingly thought to contribute to cognition[155]. Indeed, these regions exhibit delay activity that may help maintain the rules required to convert sensory information to a behavioural response. In line with this notion, delay activity in NHP dorsal premotor cortex is selective for prospective motor plans[156–158] or, when the NHP cannot anticipate a specific motor plan, an abstract response rule[159]. Moreover, a human fMRI study revealed that delay activity in pre-supplementary motor area (pre-SMA) is largely independent of the sensory content of WM information[160]. Premotor and pre-SMA delay activity may also support WM for verbal information by organizing subvocal rehearsal processes[161].

Other fMRI and ECoG studies in humans have found that dorsal premotor regions are instead sensitive to features of WM content[123] and are preferentially activated during the maintenance of simple sensory information rather than complex motor sequences[162,163]. For example, pre-SMA delay activity scales with the frequency of an oscillating visual, auditory or tactile stimulus[164]. However, the fact that this activity is agnostic to the modality of the frequency information indicates abstract supramodal coding for WM memoranda. It will be important to determine whether the anatomical locus of rule- and content-based WM representations in these regions is consistent with the purported hierarchical organization of frontal cortex[165].

Sensory cortices.—Given that primary sensory cortex and sensory association cortex show largely similar delay-activity properties, we discuss them together. Higher-order sensory regions exhibit sustained spiking[13,166,167], BOLD signal modulation[168,169], and modulation of EFP alpha power[170] during WM delays. In NHPs, delay spiking has also been reported in primary auditory[171] and somatosensory cortex[172], but not visual cortex[173]. However, human fMRI studies find clear evidence for sustained population coding of information in both early and higher-order visual areas[73,74,77,174,175].

Importantly, the delay activity in sensory cortices is highly specific for individual WM items or features[73–75,176]. These findings support the 'sensory recruitment model' of WM[177,178], which proposes that content-selective WM representations recruit the same sensory regions that initially encode the information. However, this view has been challenged[104,108,179] on the grounds that, unlike in PFC[126,180] or PPC[132,142,181], content-selective delay activity in sensory cortices is often interrupted by intervening input[132,180,181]. Distractor resistance is

thought to be an essential property of delay activity that allows WM to function despite interference. Crucially, however, the sensory recruitment model does not claim that sensory cortex is sufficient for storing WM content; rather, it posits that sensory delay activity, together with influence from other regions, has a pivotal role in representing detailed information in WM[178,182]. Viewed through this lens, distractor-resistant delay activity is not a critical test of the sensory recruitment model. The fact that content-selective activity in sensory cortex is reinstated following the presentation of intervening stimuli indicates that these representations can still be used to guide behaviour[176,183].

Subcortical nuclei

Basal ganglia.—The basal ganglia (BG) transiently gate information during WM encoding[184] and influence behavioural responses based on currently relevant WM contents[185]; however, the BG also exhibit sustained activity during WM delays. NHP studies have found delay spiking in the putamen[186,187], caudate[188,189] and globus pallidus[188] that takes several trials to stabilize into persistent spiking[187] and ramps up over the delay[189] and that does not necessarily scale with the accuracy of the behavioural response[187]. Thus, the BG may encode the anticipation of an upcoming movement rather than the motor plan itself. Consistent with these results, human fMRI studies have revealed that delay activity in the caudate during WM for spatial location was enhanced[190,191] when an upcoming response could be prospectively encoded[191]. However, fMRI observations of delay activity in the caudate and putamen during tasks in which specific behavioural responses could not be anticipated, such as during verbal WM[192] or during WM for temporal duration[193], indicate that, at least in humans, BG delay activity is not limited to response anticipation.

Thalamus.—Elevated delay spiking in the thalamus was originally identified concurrently with lPFC delay activity in 1971[2], although the thalamic result received less attention (see also Refs[194,195]). Subsequent work characterized delay activity in the mediodorsal nucleus of the thalamus[196,197] (reviewed elsewhere[198,199]), which has dense reciprocal connections with the dorsolateral PFC in humans[200] and NHPs[201]. Similar to the lPFC, delay spiking in neurons of the mediodorsal nucleus was spatially specific, representing either the location of the visual cue or the target of the upcoming response[196].

New evidence about the role of thalamic delay activity comes from studies that optogenetically inhibited thalamocortical communication in mice. Photoinhibition of thalamic delay activity disrupted delay activity in frontal regions (including anterior lateral motor cortex and medial PFC, the putative rodent homologue of dorsolateral PFC and the recipient of projections from mediodorsal thalamus) and behaviour on simple delayed-discrimination tasks[202–204]. Interestingly, thalamic delay activity did not simply share the same properties as frontal delay activity; thalamic delay activity was less content-selective than frontal delay activity and causally modulated the selectivity[202] and magnitude[203] of frontal delay activity.

There are three key takeaway points from these studies. First, there is a tight causal relationship between thalamic delay activity, cortical delay spiking and behaviour (see also Ref.[205]). Second, local circuit mechanisms, which have long been implicated in delay

activity, may be insufficient to sustain delay activity in the absence of long-range input — in this case, from the thalamus. Third, thalamic delay activity does not merely relay information to and from cortex, but instead may directly shape sustained representations and the coordination of information storage required for WM[206,207]. This work is part of a growing appreciation of the role of the thalamus in cognition[208,209].

Comparing across brain regions

Differences in tasks, methodology and species in the above studies pose a significant obstacle to uncovering the location and function of delay activity. Here we consider studies that directly compare delay activity across regions. Our discussion centres on the storage of content-selective WM representations (rather than representations of WM operations) because nearly every delay-active region has been implicated in WM storage, making content-selective delay activity particularly amenable to this type of comparative analysis.

With its broad spatial coverage, fMRI is optimally suited to compare delay activity across brain regions. Human studies have started to capitalize on the power of **encoding models**[210] to directly contrast stimulus-encoding properties across the brain during WM for simple features. These studies have documented content-selective delay representations throughout the dorsal visual hierarchy, from early visual cortex to PPC and lPFC[211,212]. Subsequent work demonstrated that WM representations throughout the cortex are flexible, shifting from retrospectively representing stimulus position to prospectively representing the motor response, and that the amplitude of delay activity in lPFC and PPC correlates with the precision of encoding in visual regions, indicating that lPFC and PPC provide a top-down signal that tunes delay activity in sensory cortices[175].

Owing to limited spatial coverage, comparisons of delay spiking across regions are relatively rare. Nevertheless, the few NHP studies that have sufficient coverage to investigate multiple regions describe results that are largely compatible with the fMRI studies outlined above. One study recorded from lPFC, the motion-sensitive middle temporal area (MT), and the multimodal medial superior temporal area (MST) during a motion WM task[113]. Delay spiking sensitive to the maintained direction of motion was identified in lPFC and MST, but not MT; however, MT exhibited persistent, stimulus-selective increases in oscillatory power in multiple frequency bands. The results were interpreted as reflecting the maintenance of motion information in higher-order regions that biased synaptic activity in visual areas. Another study recorded from an impressive 42 cortical areas in NHPs and identified multiple regions showing content-selective delay spiking, including intraparietal and visual areas, with the most robust selectivity observed in ventrolateral PFC[213].

These results raise a key question: what is the benefit of storing WM information in multiple regions at once? A potential answer comes from a study that used calcium imaging to measure neural activity across mouse cortex during a delayed discrimination task[214]. Sustained delay activity was found in medial PFC and posterior perceptual regions, but the magnitude of delay activity in these areas was modulated by the mouse's strategy: a prospective motor-based strategy increased PFC delay activity, whereas a retrospective, sensory-based strategy increased delay activity in posterior regions. Importantly, experimental perturbation of delay activity in either region prompted strategy

switches. Thus, although content-selective delay activity can be observed across the brain, it may reflect different representational formats that are preferentially recruited depending on strategy or task demands.

What are the underlying mechanisms?

Models based on biophysical properties of neurons and neural circuits have provided valuable insights into how delay activity is generated. These models can be divided into three main classes (see Refs[20,215–218] for detailed reviews). Below, we describe these models and evaluate their ability to reproduce the experimental findings discussed above. As the relationship between the biophysical properties of neurons or micro-scale networks and the EFP or BOLD signals is incompletely understood, biophysical models almost exclusively attempt to describe spiking activity; accordingly, we focus our discussion here on delay spiking. The story that emerges is that no single model can account for all features of delay activity and that many results can be explained by multiple classes of models[20] (Table 1).

Intrinsic cellular properties

Under certain conditions, individual neurons can respond to a brief stimulation with persistently above-baseline firing[219]. This ability to settle into a stable, above-baseline state is known as bistability (or multistability, in the case of neurons able to achieve multiple above-baseline firing states) and can be induced in vitro in rodent MTL slices by activating muscarinic cholinergic receptors[220,221]. Importantly, models featuring intrinsic cellular bistability[222,223] can recapitulate persistent delay spiking, typically by incorporating features that mimic a variety of cell-intrinsic mechanisms observed in vitro (such as changes in calcium currents; reviewed in Ref[20]).

A critical feature of cellular bistability is that it does not require complex network connectivity (see below) to generate content-selective delay activity. Cellular models may therefore be well-suited to explain delay activity in situations where the rich connectivity and tuning for memoranda required for network-based persistent spiking is probably absent, such as in regions that lack complex networks[224] or during WM for novel items[216]. Models involving multistable neurons can also mimic the behaviour of lPFC cells that vary their firing rates according to the rate of vibrotactile stimulation being stored in WM[81].

There are some aspects of delay activity that models based on cellular bistability struggle to capture. For example, these models cannot reproduce complex temporal dynamics associated with delay spiking[218]. In addition, although neurons in these models can clearly switch between discrete states and therefore represent discrete information (for example, an integer from 1–12 on a clock face), how they might represent continuous quantities (for instance, an exact position in degrees on a clock face) in the absence of complex networks is unclear.

Spiking in local cell assemblies

Feedforward models.—Cell assembly models are based on connectivity between multiple neurons. One group of such models of WM focuses on feedforward connections

between cells that share similar selectivity[225–227]. This connectivity enables the network to store information through brief periods of above-baseline spiking that are passed between cells[228] without requiring sustained firing in any one cell, mimicking the sequential activation of neurons observed in the hippocampus[91,92] and PFC[93]. Interestingly, feedforward networks can also reproduce stable delay spiking; a neuron that integrates activity across the feedforward chain of activity can sustain its spiking over several seconds[227]. However, the precise timing required to propagate activity in these networks may render them more susceptible to noise than networks that rely on population coding[216,225].

Attractor networks.—Another group of cell assembly models features recurrent excitatory connections between neurons in the network. When the connectivity between neurons in the network is properly structured, and when the balance between excitation and inhibition is well-tuned, spiking dynamics settle into a stable **attractor state** in the absence of any input[229]. Attractor networks can be discrete — with basins of attraction that encode discrete states — or continuous — encoding analogue values using a single continuous basin of attraction, a continuum of population firing rates along which network activity can stably traverse.

Attractor models have several appealing properties. First, and most critically, they replicate various empirical phenomena, including content-selective delay spiking and elevated gamma power[19]. Second, attractor networks inherently rely on a population code, which accords with several empirical observations described above. Third, continuous attractor models generate clear predictions about how WM representations are affected by internal neural noise; noise in model networks results in random drift along the basin of attraction and thus imprecise readout of the information encoded by the network[19]. One study tested this prediction in NHP lPFC and found that, in line with the model, drifts in population delay spiking predicted the degree of error in behavioural responses[230].

Attractor network models also have some notable drawbacks. The primary limitation of continuous attractor models is their fragility, which manifests itself in a few ways. Activity in continuous attractor networks is susceptible to drift and interference from external stimuli, meaning that encoded information is vulnerable to noise[231]. In addition, continuous attractor networks struggle to maintain multiple memory items simultaneously[102]. These issues can be overcome by approximating the behaviour of continuous attractors using discrete attractors with several basins of attraction[232]. Discrete attractor models have received recent empirical support; one study measured the response of delay-active neurons in mouse anterior lateral motor cortex and found that optogenetic perturbation resulted in neural and behavioral responses that were consistent with the predictions of a discrete – but not continuous – attractor model[233]. However, even these discrete approximations of continuous attractors require a very specific architecture, with precisely tuned and highly symmetric connectivity between individual neurons, as well as a specific balance of excitation and inhibition[18]. In practice, this architecture is probably present in networks that have been shaped by extensive learning. As a result, attractors are valuable for understanding WM for well-learned stimuli, such as spatial position[19], but are less useful for explaining WM for

novel stimuli or novel combinations of features. An additional challenge for attractor models is that they have trouble reproducing complex spiking dynamics[234].

Synaptic weights in local assemblies

A third class of models is based on the provocative claim that sustained activation is not necessary for WM. Instead, these 'activity silent' models propose that information can be maintained through rapid shifts in synaptic weights within local neural networks that encode WM information[235–237]. **Short-term plasticity (STP)**, which has been observed in PFC[238], has been proposed to underlie these changes in synaptic weights[235,239,240]. According to activity silent models, shifts in synaptic weights during stimulus encoding cause the network to act as a **matched filter**, responding to noisy input with activity that reveals the underlying state of the network (that is, the memory)[236]. As an analogy, a submarine uses sonar (that is, noisy input) to 'ping' the sea floor, resulting in an image of the otherwise unobservable state of the ocean floor (that is, the memory encoded within the network)[183]. In these models, brief changes in synaptic weights can store a memory for up to one second without spiking[235]; further maintenance requires the synaptic changes to be periodically 'refreshed' by the spiking of memory-encoding neurons[241]. Interestingly, activity-silent mechanisms raise the possibility that content-selective delay spiking, EFPs or BOLD responses may simply result from a non-specific input (that is, noise) that is delivered to a neuronal population that silently maintains a memory[183,242], rather than the mechanism of memory maintenance itself[236].

On one hand, models based on synaptic mechanisms are appealing for several reasons. First, unlike in many models (including spiking-based models) that struggle to explain how a network can simultaneously represent WM information and incoming sensory stimuli, in synaptic models spiking input does not interfere with the memory in the same way. Second, activity-silent models may explain how neurons with short **time constants**, such as those in sensory cortices, maintain information over longer intervals[243]; if the memory is stored in synaptic weights, sparse intermittent spiking may be sufficient to store it over tens of seconds. Third, activity-silent models may be more energy-efficient than spiking models[97,235], although a countervailing view from an analysis of NHP PFC neurons found that increased metabolic demand due to delay spiking was offset by decreases in the activity in other neurons, resulting in a network-wide metabolic cost that was virtually unchanged from baseline to delay[98].

On the other hand, most models of synaptic mechanisms are unable to recreate WM for novel items. A caveat when evaluating arguments for activity-silent mechanisms is that empirical support for these models is often based on the absence of delay activity[244]. One cannot rule out the possibility that the apparent absence of delay activity instead reflects very low levels of delay activity or an alternative form of delay activity to which the method is insensitive. Attempts to clarify how activity-silent and spiking mechanisms coexist are a logical next step towards an understanding of the synaptic mechanisms of WM.

Hybrid models

Elements of the models described above have been combined to overcome some of their individual limitations. For example, several models have been formulated to increase the stability of continuous attractor networks in the face of drift and external input, using either STP[231] or cellular bistability[232,245]. These models are able to stabilize the information encoded by delay firing over physiologically realistic intervals, thus improving memory readout. Hybrid models that blend elements of STP with traditional attractor network models are also able to reproduce irregular patterns of delay firing[234,241] that are consistent with experimental outcomes[81]. In addition, recurrent network models that incorporate STP can account for regular network-level oscillations in the gamma and beta frequencies, as well as intermittent bursting[246]. Two predictions of this latter model — that the gamma burst rate should increase with the anticipation of a behavioural response, and that the frequency of gamma bursts should increase with WM load — were confirmed empirically[41,42].

Hybrid models are also able to address another limitation of most traditional models — the inability to store novel information. One model embedded bistable cells in a network architecture that allowed the network to encode and maintain novel items through spiking[247], whereas a more recent study used Hebbian STP in an attractor model to enable the encoding of novel items[241]. Further refinement of hybrid models, and their use in generating network-level predictions in NHPs and humans, will be essential to fully elucidate the mechanisms that give rise to delay activity.

Concluding remarks

The work reviewed above suggests three key points about delay activity. First, although delay activity has classically been described as sustained activation of highly tuned neurons or neural populations, the abundance of evidence for time-varying forms of delay activity and population coding indicates that a broader view of delay activity is necessary. In particular, analytical advances have revealed WM information in LFP bursts[41] and dynamic coding[97] — information to which standard analyses were insensitive. An important avenue for future research will be to reconcile dynamic and stable modes of coding for WM information.

Second, the evidence suggests that widespread brain regions are able to exhibit delay activity in the service of WM[178]. Distinguishing between regions that generate delay activity and those that inherit delay activity as a result of connections with delay-active regions will be crucial in disentangling the many functions of delay activity. Even if a region does not generate delay activity itself, delay activity induced by another region may nevertheless have an important functional role. The ubiquity of delay activity across a diverse set of brain regions further calls for a significant emphasis on how information is integrated across the brain during WM maintenance. Accordingly, the study of meso- and macro-level networks using EFP and fMRI will be essential in providing data that complements single-neuron studies. The broad recruitment of brain regions during WM also highlights the need for methods that disrupt one region while recording delay activity from another, including the use of concurrent transcranial magnetic stimulation (TMS) and fMRI, and/or optogenetics with spike recordings.

Third, despite many attempts to draw links between empirical findings and the mechanisms hypothesized by biophysical models, the data are often consistent with several models. One strategy to overcome this limitation is to directly assess how different features of these models reproduce realistic data[248]. A second strategy is to identify differential predictions made by some of these models[233]. For example, synaptic mechanisms and continuous attractors seem to make opposite predictions about the effects of noise; noise in attractor networks results in poorer readout by downstream neurons, whereas noise delivered to a population of neurons that maintains information through shifts in synaptic weights could transiently enhance readout. Although some elements of this prediction have been explored empirically[183,249], these two models have not been directly pitted against one another. Connection between modelling and experimental work will also be necessary to explore whether distinct representational states within WM[250] correspond to different cellular and network mechanisms that produce delay activity. Further crosstalk between neuron- and circuit-level modelling work and systems-level empirical work will be crucial to advance our understanding of delay activity.

Acknowledgements

We thank Anastasia Kiyonaga, Elizabeth Lorenc, and Dan Bliss for their helpful comments on previous versions of this manuscript. This work was supported by NIH Grant MH63901 to Mark D'Esposito.

Glossary

WM delay task
Tasks that temporally segregate working memory encoding, maintenance and response by introducing an unfilled memory delay between a memory stimulus and the contingent behavioural response

Population coding
A coding scheme wherein information is encoded in the combined activity of a population of neurons (or electrodes, or voxels) as opposed to the activity of individual neurons (or electrodes, or voxels)

Voxel
The volumetric unit of functional MRI (fMRI) measurement. A 3D fMRI brain image contains ~100,000 voxels, each of which represents the activity of tens of thousands of neurons

Nonlinear mixed selectivity
A property that allows neurons to respond to combinations of stimulus or task features with nonlinear changes in firing rates

WM load
The amount of information that is held in working memory. Working memory load can be manipulated by varying the number or complexity of memory items

WM capacity

The upper bound on the amount of information that an individual can store at once in working memory

Encoding model
A model that forms a prediction of brain activity for given set of experimental features (for example, specific memory items during a working memory delay task)

Attractor state
A stable state of the activity of a network of (usually, recurrently connected) neurons that persists in the absence of input

Short-term plasticity
(STP). Synaptic plasticity in response to brief (~1 s) stimulation. Hebbian forms, (involving presynaptic and postsynaptic changes) and non-Hebbian forms (involving only presynaptic changes) of STP have been proposed to underlie working memory

Matched filter
A linear filter that can help detect the presence of a known stimulus in a noisy observed signal by correlating the known stimulus with the observed signal

Time constant
A value that describes the time required for a neuron to return to a baseline state following an input

Haemodynamic response
The temporal pattern of blood-oxygen-level-dependent signal observed by functional MRI in response to a brief impulse of neural activity. It takes ~20 s to return to baseline

General linear model
(GLM). A model that describes the output of a system as a linear combination of predictors. GLMs are used to estimate BOLD responses to features of an experimental task

Impulse response function
The output of a dynamic system in response to a brief input

References

1. D'Esposito M & Postle BR The Cognitive Neuroscience of Working Memory. Annual Review of Psychology 66, 115–142 (2015).
2. Fuster JM & Alexander GE Neuron activity related to short-term memory. Science 173, 652–654 (1971). [PubMed: 4998337]
3. Kubota K & Niki H Prefrontal cortical unit activity and delayed alternation performance in monkeys. J. Neurophysiol 34, 337–347 (1971). [PubMed: 4997822]
4. Rosenkilde CE, Bauer RH & Fuster JM Single cell activity in ventral prefrontal cortex of behaving monkeys. Brain Res. 209, 375–394 (1981). [PubMed: 7225799]
5. Goldman-Rakic PS Cellular basis of working memory. Neuron 14, 477–485 (1995). [PubMed: 7695894]
6. Chelazzi L, Miller EK, Duncan J & Desimone R A neural basis for visual search in inferior temporal cortex. Nature 363, 345–347 (1993). [PubMed: 8497317]

7. Curtis CE & Lee D Beyond working memory: the role of persistent activity in decision making. Trends in Cognitive Sciences 14, 216–222 (2010). [PubMed: 20381406]
8. Baeg EH et al. Dynamics of population code for working memory in the prefrontal cortex. Neuron 40, 177–188 (2003). [PubMed: 14527442]
9. Yang S-T, Shi Y, Wang Q, Peng J-Y & Li B-M Neuronal representation of working memory in the medial prefrontal cortex of rats. Molecular Brain 7, (2014).
10. Funahashi S, Bruce CJ & Goldman-Rakic PS Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol 61, 331–349 (1989). [PubMed: 2918358]
11. Gnadt JW & Andersen RA Memory related motor planning activity in posterior parietal cortex of macaque. Exp. Brain Res 70, 216–220 (1988). [PubMed: 3402565]
12. Chafee MV & Goldman-Rakic PS Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. J. Neurophysiol 79, 2919–2940 (1998). [PubMed: 9636098]
13. Fuster JM & Jervey JP Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. J. Neurosci 2, 361–375 (1982). [PubMed: 7062115]
14. Nakamura K & Kubota K Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. J. Neurophysiol 74, 162–178 (1995). [PubMed: 7472321]
15. Kamiński J et al. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. Nat. Neurosci 20, 590–601 (2017). [PubMed: 28218914]
16. Kornblith S, Quian Quiroga R, Koch C, Fried I & Mormann F Persistent Single-Neuron Activity during Working Memory in the Human Medial Temporal Lobe. Curr. Biol 27, 1026–1032 (2017). [PubMed: 28318972]
17. Wang XJ Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 24, 455–463 (2001). [PubMed: 11476885]
18. Amit DJ & Brunel N Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex 7, 237–252 (1997). [PubMed: 9143444]
19. Compte A, Brunel N, Goldman-Rakic PS & Wang XJ Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex 10, 910–923 (2000). [PubMed: 10982751]
20. Zylberberg J & Strowbridge BW Mechanisms of Persistent Activity in Cortical Circuits: Possible Neural Substrates for Working Memory. Annu. Rev. Neurosci 40, 603–627 (2017). [PubMed: 28772102]
21. Yuste R From the neuron doctrine to neural networks. Nat. Rev. Neurosci 16, 487–497 (2015). [PubMed: 26152865]
22. Averbeck BB, Latham PE & Pouget A Neural correlations, population coding and computation. Nat. Rev. Neurosci 7, 358–366 (2006). [PubMed: 16760916]
23. Sporns O Structure and function of complex brain networks. Dialogues Clin. Neurosci 15, 247–262 (2013). [PubMed: 24174898]
24. Qi X-L, Meyer T, Stanford TR & Constantinidis C Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. Cereb. Cortex 21, 2722–2732 (2011). [PubMed: 21527786]
25. Meyer T, Qi X-L, Stanford TR & Constantinidis C Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. J. Neurosci 31, 6266–6276 (2011). [PubMed: 21525266]
26. Liu D et al. Medial prefrontal activity during delay period contributes to learning of a working memory task. Science 346, 458–463 (2014). [PubMed: 25342800]
27. Vogel EK & Machizawa MG Neural activity predicts individual differences in visual working memory capacity. Nature 428, 748–751 (2004). [PubMed: 15085132]
28. Voytek B & Knight RT Prefrontal cortex and basal ganglia contributions to visual working memory. Proc. Natl. Acad. Sci. U. S. A 107, 18167–18172 (2010). [PubMed: 20921401]
29. Reinhart RMG et al. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. J. Neurosci 32, 7711–7722 (2012). [PubMed: 22649249]

30. Pipa G et al. Performance- and stimulus-dependent oscillations in monkey prefrontal cortex during short-term memory. Front. Integr. Neurosci 3, 25 (2009). [PubMed: 19862343]
31. Haller M et al. Persistent neuronal activity in human prefrontal cortex links perception and action. Nat Hum Behav 2, 80–91 (2018). [PubMed: 29963646]
32. Honkanen R, Rouhinen S, Wang SH, Palva JM & Palva S Gamma Oscillations Underlie the Maintenance of Feature-Specific Information and the Contents of Visual Working Memory. Cereb. Cortex 25, 3788–3801 (2015). [PubMed: 25405942]
33. Tallon-Baudry C, Bertrand O & Fischer C Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance. J. Neurosci 21, RC177 (2001). [PubMed: 11588207]
34. Raghavachari S et al. Gating of human theta oscillations by a working memory task. J. Neurosci 21, 3175–3183 (2001). [PubMed: 11312302]
35. Miller EK, Lundqvist M & Bastos AM Working Memory 2.0. Neuron 100, 463–475 (2018). [PubMed: 30359609]
36. Klimesch W, Doppelmayr M, Schwaiger J, Auinger P & Winkler T 'Paradoxical' alpha synchronization in a memory task. Brain Res. Cogn. Brain Res 7, 493–501 (1999). [PubMed: 10076094]
37. Jensen O, Gelfand J, Kounios J & Lisman JE Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. Cereb. Cortex 12, 877–882 (2002). [PubMed: 12122036]
38. Jokisch D & Jensen O Modulation of gamma and alpha activity during a working memory task engaging the dorsal or ventral stream. J. Neurosci 27, 3244–3251 (2007). [PubMed: 17376984]
39. van Ede F, Jensen O & Maris E Supramodal Theta, Gamma, and Sustained Fields Predict Modality-specific Modulations of Alpha and Beta Oscillations during Visual and Tactile Working Memory. J. Cogn. Neurosci 29, 1455–1472 (2017). [PubMed: 28358658]
40. van Ede F Mnemonic and attentional roles for states of attenuated alpha oscillations in perceptual working memory: a review. Eur. J. Neurosci 48, 2509–2515 (2018). [PubMed: 29068095]
41. Lundqvist M et al. Gamma and Beta Bursts Underlie Working Memory. Neuron 90, 152–164 (2016). [PubMed: 26996084] By analyzing single trial LFP data, this paper demonstrates that WM delay activity is characterized by transient bursts of activity in the gamma and beta frequency ranges. Importantly, (gamma) LFP bursts were associated with spiking activity that encoded information about WM memoranda, while sustained LFP activity did not exhibit a relationship with information encoding.
42. Lundqvist M, Herman P, Warden MR, Brincat SL & Miller EK Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nat. Commun 9, 394 (2018). [PubMed: 29374153]
43. Mitzdorf U Evoked potentials and current source densities in the cat visual cortex. Electroencephalography and Clinical Neurophysiology 61, S179 (1985).
44. Baillet S, Mosher JC & Leahy RM Electromagnetic brain mapping. IEEE Signal Processing Magazine 18, 14–30 (2001).
45. Pesaran B, Pezaris JS, Sahani M, Mitra PP & Andersen RA Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat. Neurosci 5, 805–811 (2002). [PubMed: 12134152]
46. Backen T, Treue S & Martinez-Trujillo JC Encoding of Spatial Attention by Primate Prefrontal Cortex Neuronal Ensembles. eNeuro 5, (2018).
47. O'Keefe J & Recce ML Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus 3, 317–330 (1993). [PubMed: 8353611]
48. Jacobs J, Kahana MJ, Ekstrom AD & Fried I Brain oscillations control timing of single-neuron activity in humans. J. Neurosci 27, 3839–3844 (2007). [PubMed: 17409248]
49. Rasch MJ, Gretton A, Murayama Y, Maass W & Logothetis NK Inferring spike trains from local field potentials. J. Neurophysiol 99, 1461–1476 (2008). [PubMed: 18160425]
50. Siegel M, Warden MR & Miller EK Phase-dependent neuronal coding of objects in short-term memory. Proc. Natl. Acad. Sci. U. S. A 106, 21341–21346 (2009). [PubMed: 19926847]

51. Buzsáki G, Anastassiou CA & Koch C The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci 13, 407–420 (2012). [PubMed: 22595786]
52. Roux F & Uhlhaas PJ Working memory and neural oscillations: α-γ versus θ-γ codes for distinct WM information? Trends Cogn. Sci 18, 16–25 (2014). [PubMed: 24268290]
53. Lisman JE & Jensen O The θ-γ neural code. Neuron 77, 1002–1016 (2013). [PubMed: 23522038]
54. Fries P A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn. Sci 9, 474–480 (2005). [PubMed: 16150631]
55. Palva JM, Palva S & Kaila K Phase synchrony among neuronal oscillations in the human cortex. J. Neurosci 25, 3962–3972 (2005). [PubMed: 15829648]
56. Womelsdorf T et al. Modulation of neuronal interactions through neuronal synchronization. Science 316, 1609–1612 (2007). [PubMed: 17569862]
57. Courtney SM, Ungerleider LG, Keil K & Haxby JV Transient and sustained activity in a distributed neural system for human working memory. Nature 386, 608–611 (1997). [PubMed: 9121584]
58. Zarahn E, Aguirre G & D'Esposito M A trial-based experimental design for fMRI. Neuroimage 6, 122–138 (1997). [PubMed: 9299386]
59. Jha AP & McCarthy G The influence of memory load upon delay-interval activity in a working-memory task: an event-related functional MRI study. J. Cogn. Neurosci 12 Suppl 2, 90–105 (2000). [PubMed: 11506650]
60. Leung H-C, Gore JC & Goldman-Rakic PS Sustained mnemonic response in the human middle frontal gyrus during on-line storage of spatial memoranda. J. Cogn. Neurosci 14, 659–671 (2002). [PubMed: 12126506]
61. Logothetis NK, Pauls J, Augath M, Trinath T & Oeltermann A Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157 (2001). [PubMed: 11449264]
62. Goense JBM & Logothetis NK Neurophysiology of the BOLD fMRI signal in awake monkeys. Curr. Biol 18, 631–640 (2008). [PubMed: 18439825]
63. Murayama Y et al. Relationship between neural and hemodynamic signals during spontaneous activity studied with temporal kernel CCA. Magn. Reson. Imaging 28, 1095–1103 (2010). [PubMed: 20096530]
64. Winawer J et al. Asynchronous broadband signals are the principal source of the BOLD response in human visual cortex. Curr. Biol 23, 1145–1153 (2013). [PubMed: 23770184]
65. Takata N et al. Optogenetic astrocyte activation evokes BOLD fMRI response with oxygen consumption without neuronal activity modulation. Glia 66, 2013–2023 (2018). [PubMed: 29845643]
66. Khursheed F et al. Frequency-specific electrocorticographic correlates of working memory delay period fMRI activity. Neuroimage 56, 1773–1782 (2011). [PubMed: 21356314]
67. Serences JT & Saproo S Computational advances towards linking BOLD and behavior. Neuropsychologia 50, 435–446 (2012). [PubMed: 21840553]
68. Vo VA, Sprague TC & Serences JT Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. J. Neurosci 37, 3386–3401 (2017). [PubMed: 28242794]
69. Constantinidis C & Procyk E The primate working memory networks. Cogn. Affect. Behav. Neurosci 4, 444–465 (2004). [PubMed: 15849890]
70. Gazzaley A, Rissman J & D'Esposito M Functional connectivity during working memory maintenance. Cogn. Affect. Behav. Neurosci 4, 580–599 (2004). [PubMed: 15849899]
71. Stokes MG et al. Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron 78, 364–375 (2013). [PubMed: 23562541]
72. Barak O, Tsodyks M & Romo R Neuronal population coding of parametric working memory. J. Neurosci 30, 9424–9430 (2010). [PubMed: 20631171]
73. Serences JT, Ester EF, Vogel EK & Awh E Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci 20, 207–214 (2009). [PubMed: 19170936]
74. Harrison SA & Tong F Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635 (2009). [PubMed: 19225460]

75. Riggall AC & Postle BR The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci 32, 12990–12998 (2012). [PubMed: 22993416]
76. Christophel TB, Hebart MN & Haynes J-D Decoding the contents of visual short-term memory from human visual and parietal cortex. J. Neurosci 32, 12983–12989 (2012). [PubMed: 22993415]
77. Sreenivasan KK, Vytlacil J & D'Esposito M Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. J. Cogn. Neurosci 26, 1141–1153 (2014). [PubMed: 24392897]
78. Foster JJ, Sutterer DW, Serences JT, Vogel EK & Awh E The topography of alpha-band activity tracks the content of spatial working memory. J. Neurophysiol 115, 168–177 (2016). [PubMed: 26467522]
79. Stokes M & Spaak E The Importance of Single-Trial Analyses in Cognitive Neuroscience. Trends Cogn. Sci 20, 483–486 (2016). [PubMed: 27237797]
80. Compte A et al. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J. Neurophysiol 90, 3441–3454 (2003). [PubMed: 12773500]
81. Brody CD, Hernández A, Zainos A & Romo R Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex 13, 1196–1207 (2003). [PubMed: 14576211]
82. Shafi M et al. Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience 146, 1082–1108 (2007). [PubMed: 17418956]
83. Durstewitz D & Seamans JK Beyond bistability: biophysics and temporal dynamics of working memory. Neuroscience 139, 119–133 (2006). [PubMed: 16326020]
84. Spaak E, Watanabe K, Funahashi S & Stokes MG Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. J. Neurosci 37, 6503–6516 (2017). [PubMed: 28559375]
85. Watanabe K & Funahashi S Prefrontal delay-period activity reflects the decision process of a saccade direction during a free-choice ODR task. Cereb. Cortex 17 Suppl 1, i88–100 (2007). [PubMed: 17726006]
86. Quintana J & Fuster JM Mnemonic and predictive functions of cortical neurons in a memory task. Neuroreport 3, 721–724 (1992). [PubMed: 1520863]
87. Quintana J & Fuster JM From perception to action: temporal integrative functions of prefrontal and parietal neurons. Cereb. Cortex 9, 213–221 (1999). [PubMed: 10355901]
88. Kojima S & Goldman-Rakic PS Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. 248, 43–49 (1982). [PubMed: 7127141]
89. Goldman-Rakic PS Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. Comprehensive Physiology (2011). doi:10.1002/cphy.cp010509
90. Howard MW Memory as Perception of the Past: Compressed Time inMind and Brain. Trends Cogn. Sci 22, 124–136 (2018). [PubMed: 29389352]
91. Pastalkova E, Itskov V, Amarasingham A & Buzsáki G Internally generated cell assembly sequences in the rat hippocampus. Science 321, 1322–1327 (2008). [PubMed: 18772431]
92. MacDonald CJ, Lepage KQ, Eden UT & Eichenbaum H Hippocampal 'time cells' bridge the gap in memory for discontiguous events. Neuron 71, 737–749 (2011). [PubMed: 21867888]
93. Batuev AS, Pirogov AA, Orlov AA & Sheafer VI Cortical mechanisms of goal-directed motor acts in the rhesus monkey. Acta Neurobiol. Exp 40, 27–49 (1980).
94. Meyers EM Dynamic population coding and its relationship to working memory. J. Neurophysiol 120, 2260–2268 (2018). [PubMed: 30207866] This review provides an in-depth discussion of the potential benefits and costs of dynamic population coding of WM information, highlighting the ways in which information from dynamics codes may be interpreted by downstream brain regions.
95. Wang J, Narain D, Hosseini EA & Jazayeri M Flexible timing by temporal scaling of cortical responses. Nat. Neurosci 21, 102–110 (2018). [PubMed: 29203897]
96. Tiganj Z, Cromer JA, Roy JE, Miller EK & Howard MW Compressed Timeline of Recent Experience in Monkey Lateral Prefrontal Cortex. J. Cogn. Neurosci 30, 935–950 (2018). [PubMed: 29698121]

97. Stokes MG 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci 19, 394–405 (2015). [PubMed: 26051384]

98. Murray JD et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. U. S. A 114, 394–399 (2017). [PubMed: 28028221] This study addresses the question of how WM information is stably encoded by dynamic population codes. The authors applied principal components analysis to the complex temporal dynamics exhibited by NHP lPFC neurons and identified a low-dimensional population code that was stable across the delay.

99. Druckmann S & Chklovskii DB Neuronal circuits underlying persistent representations despite time varying activity. Curr. Biol 22, 2095–2103 (2012). [PubMed: 23084992]

100. Myers NE et al. Testing sensory evidence against mnemonic templates. eLife 4, (2015).

101. Constantinidis C et al. Persistent Spiking Activity Underlies Working Memory. J. Neurosci 38, 7020–7028 (2018). [PubMed: 30089641]

102. Lundqvist M, Herman P & Miller EK Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. J. Neurosci 38, 7013–7019 (2018). [PubMed: 30089640]

103. Lee S-H & Baker CI Multi-Voxel Decoding and the Topography of Maintained Information During Visual Working Memory. Frontiers in Systems Neuroscience 10, (2016).

104. Leavitt ML, Mendoza-Halliday D & Martinez-Trujillo JC Sustained Activity Encoding Working Memories: Not Fully Distributed. Trends Neurosci. 40, 328–346 (2017). [PubMed: 28515011]

105. Curtis CE & D'Esposito M Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences 7, 415–423 (2003). [PubMed: 12963473]

106. Sreenivasan KK, Curtis CE & D'Esposito M Revisiting the role of persistent neural activity during working memory. Trends in Cognitive Sciences 18, 82–89 (2014). [PubMed: 24439529]

107. Lara AH & Wallis JD The Role of Prefrontal Cortex in Working Memory: A Mini Review. Frontiers in Systems Neuroscience 9, (2015).

108. Riley MR & Constantinidis C Role of Prefrontal Persistent Activity in Working Memory. Front. Syst. Neurosci 9, 181 (2015). [PubMed: 26778980]

109. Constantinidis C, Franowicz MN & Goldman-Rakic PS The sensory nature of mnemonic representation in the primate prefrontal cortex. Nat. Neurosci 4, 311–316 (2001). [PubMed: 11224549]

110. Sprague TC & Serences JT Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci 16, 1879–1887 (2013). [PubMed: 24212672]

111. Curtis CE, Rao VY & D'Esposito M Maintenance of spatial and motor codes during oculomotor delayed response tasks. J. Neurosci 24, 3944–3952 (2004). [PubMed: 15102910]

112. Zaksas D & Pasternak T Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. J. Neurosci 26, 11726–11742 (2006). [PubMed: 17093094]

113. Mendoza-Halliday D, Torres S & Martinez-Trujillo JC Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nat. Neurosci 17, 1255–1262 (2014). [PubMed: 25108910]

114. Romo R, Brody CD, Hernández A & Lemus L Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399, 470–473 (1999). [PubMed: 10365959]

115. Lewis-Peacock JA & Postle BR Temporary activation of long-term memory supports working memory. J. Neurosci 28, 8765–8771 (2008). [PubMed: 18753378]

116. Rainer G, Rao SC & Miller EK Prospective coding for objects in primate prefrontal cortex. J. Neurosci 19, 5493–5505 (1999). [PubMed: 10377358]

117. Meyers EM, Freedman DJ, Kreiman G, Miller EK & Poggio T Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol 100, 1407–1419 (2008). [PubMed: 18562555]

118. Wutz A, Loonis R, Roy JE, Donoghue JA & Miller EK Different Levels of Category Abstraction by Different Dynamics in Different Prefrontal Areas. Neuron 97, 716–726.e8 (2018). [PubMed: 29395915]

119. Warden MR & Miller EK Task-dependent changes in short-term memory in the prefrontal cortex. J. Neurosci 30, 15801–15810 (2010). [PubMed: 21106819]
120. Wallis JD, Anderson KC & Miller EK Single neurons in prefrontal cortex encode abstract rules. Nature 411, 953–956 (2001). [PubMed: 11418860]
121. Leon MI & Shadlen MN Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24, 415–425 (1999). [PubMed: 10571234]
122. Cavanagh SE, Towers JP, Wallis JD, Hunt LT & Kennerley SW Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nat. Commun 9, 3498 (2018). [PubMed: 30158519]
123. Mohr HM, Goebel R & Linden DEJ Content- and task-specific dissociations of frontal activity during maintenance and manipulation in visual working memory. J. Neurosci 26, 4465–4471 (2006). [PubMed: 16641225]
124. Fusi S, Miller EK & Rigotti M Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol 37, 66–74 (2016). [PubMed: 26851755]
125. Rigotti M et al. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590 (2013). [PubMed: 23685452] This paper presents evidence for nonlinear mixed selectivity in NHP lPFC neurons and shows that the high-dimensional representations that are enabled by nonlinear mixed selectivity are crucial for behavior.
126. Parthasarathy A et al. Mixed selectivity morphs population codes in prefrontal cortex. Nat. Neurosci 20, 1770–1779 (2017). [PubMed: 29184197]
127. Percheron G, François C & Pouget P What makes a frontal area of primate brain the frontal eye field? Front. Integr. Neurosci 9, 33 (2015). [PubMed: 26042006]
128. Schall JD et al. On the Evolution of the Frontal Eye Field: Comparisons of Monkeys, Apes, and Humans. Evolution of Nervous Systems 249–275 (2017). doi:10.1016/b978-0-12-804042-3.00130-5
129. Constantinidis C & Steinmetz MA Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. J. Neurophysiol 76, 1352–1355 (1996). [PubMed: 8871242]
130. Todd JJ & Marois R Capacity limit of visual short-term memory in human posterior parietal cortex. Nature 428, 751–754 (2004). [PubMed: 15085133]
131. Schluppeck D, Curtis CE, Glimcher PW & Heeger DJ Sustained activity in topographic areas of human posterior parietal cortex during memory-guided saccades. J. Neurosci 26, 5098–5108 (2006). [PubMed: 16687501]
132. Bettencourt KC & Xu Y Decoding the content of visual short-term memory under distraction in occipital and parietal areas. Nat. Neurosci 19, 150–157 (2016). [PubMed: 26595654]
133. Meltzer JA et al. Effects of working memory load on oscillatory power in human intracranial EEG. Cereb. Cortex 18, 1843–1855 (2008). [PubMed: 18056698]
134. Sarma A, Masse NY, Wang X-J & Freedman DJ Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. Nat. Neurosci 19, 143–149 (2016). [PubMed: 26595652]
135. Chein JM, Ravizza SM & Fiez JA Using neuroimaging to evaluate models of working memory and their implications for language processing. Journal of Neurolinguistics 16, 315–339 (2003).
136. Berryhill ME, Chein J & Olson IR At the intersection of attention and memory: the mechanistic role of the posterior parietal lobe in working memory. Neuropsychologia 49, 1306–1315 (2011). [PubMed: 21345344]
137. Corbetta M, Miezin FM, Shulman GL & Petersen SE A PET study of visuospatial attention. J. Neurosci 13, 1202–1226 (1993). [PubMed: 8441008]
138. Elston GN Pyramidal cells of the frontal lobe: all the more spinous to think with. J. Neurosci 20, RC95 (2000). [PubMed: 10974092]
139. Katsuki F et al. Differences in intrinsic functional organization between dorsolateral prefrontal and posterior parietal cortex. Cereb. Cortex 24, 2334–2349 (2014). [PubMed: 23547137]
140. Katsuki F & Constantinidis C Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions. Front. Integr. Neurosci 6, 17 (2012). [PubMed: 22563310]

141. Mackey WE & Curtis CE Distinct contributions by frontal and parietal cortices support working memory. Sci. Rep 7, 6188 (2017). [PubMed: 28733684]
142. Jacob SN & Nieder A Complementary roles for primate frontal and parietal cortex in guarding working memory from distractor stimuli. Neuron 83, 226–237 (2014). [PubMed: 24991963]
143. Masse NY, Hodnefield JM & Freedman DJ Mnemonic Encoding and Cortical Organization in Parietal and Prefrontal Cortices. J. Neurosci 37, 6098–6112 (2017). [PubMed: 28539423]
144. Ranganath C & Blumenfeld RS Doubts about double dissociations between short- and long-term memory. Trends Cogn. Sci 9, 374–380 (2005). [PubMed: 16002324]
145. Jeneson A & Squire LR Working memory, long-term memory, and medial temporal lobe function. Learn. Mem 19, 15–25 (2012). [PubMed: 22180053]
146. Rissman J, Gazzaley A & D'Esposito M Dynamic adjustments in prefrontal, hippocampal, and inferior temporal interactions with increasing visual working memory load. Cereb. Cortex 18, 1618–1629 (2008). [PubMed: 17999985]
147. Suzuki WA, Miller EK & Desimone R Object and place memory in the macaque entorhinal cortex. J. Neurophysiol 78, 1062–1081 (1997). [PubMed: 9307135]
148. Alvarez GA & Cavanagh P The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychol. Sci 15, 106–111 (2004). [PubMed: 14738517]
149. Brady TF & Alvarez GA No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. J. Exp. Psychol. Learn. Mem. Cogn 41, 921–929 (2015). [PubMed: 25419824]
150. Olson IR, Page K, Moore KS, Chatterjee A & Verfaellie M Working memory for conjunctions relies on the medial temporal lobe. J. Neurosci 26, 4596–4601 (2006). [PubMed: 16641239]
151. Davachi L Item, context and relational episodic encoding in humans. Curr. Opin. Neurobiol 16, 693–700 (2006). [PubMed: 17097284]
152. Hasselmo ME & Stern CE Mechanisms underlying working memory for novel information. Trends Cogn. Sci 10, 487–493 (2006). [PubMed: 17015030]
153. Ranganath C & D'Esposito M Medial Temporal Lobe Activity Associated with Active Maintenance of Novel Information. Neuron 31, 865–873 (2001). [PubMed: 11567623]
154. Buzsáki G & Tingley D Space and Time: The Hippocampus as a Sequence Generator. Trends Cogn. Sci 22, 853–869 (2018). [PubMed: 30266146]
155. Chung GH, Han YM & Kim CS Functional MRI of the Supplementary Motor Area: Comparison of Motor and Sensory Tasks. Journal of Computer Assisted Tomography 24, 521–525 (2000). [PubMed: 10966180]
156. Kaufman MT, Churchland MM, Ryu SI & Shenoy KV Cortical activity in the null space: permitting preparation without movement. Nat. Neurosci 17, 440–448 (2014). [PubMed: 24487233]
157. di Pellegrino G & Wise SP Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J. Neurosci 13, 1227–1243 (1993). [PubMed: 8441009]
158. Ohbayashi M, Ohki K & Miyashita Y Conversion of working memory to motor sequence in the monkey premotor cortex. Science 301, 233–236 (2003). [PubMed: 12855814]
159. Wallis JD & Miller EK From rule to response: neuronal processes in the premotor and prefrontal cortex. J. Neurophysiol 90, 1790–1806 (2003). [PubMed: 12736235]
160. Petit L, Courtney SM, Ungerleider LG & Haxby JV Sustained activity in the medial wall during working memory delays. J. Neurosci 18, 9429–9437 (1998). [PubMed: 9801381]
161. Buchsbaum BR & D'Esposito M A sensorimotor view of verbal working memory. Cortex 112, 134–148 (2019). [PubMed: 30639088]
162. Simon SR et al. Spatial attention and memory versus motor preparation: premotor cortex involvement as revealed by fMRI. J. Neurophysiol 88, 2047–2057 (2002). [PubMed: 12364527]
163. Brovelli A, Lachaux J-P, Kahane P & Boussaoud D High gamma frequency oscillatory activity dissociates attention from intention in the human premotor cortex. Neuroimage 28, 154–164 (2005). [PubMed: 16023374]

164. Vergara J, Rivera N, Rossi-Pool R & Romo R A Neural Parametric Code for Storing Information of More than One Sensory Modality in Working Memory. Neuron 89, 54–62 (2016). [PubMed: 26711117]
165. Badre D & D'Esposito M Is the rostro-caudal axis of the frontal lobe hierarchical? Nat. Rev. Neurosci 10, 659–669 (2009). [PubMed: 19672274]
166. Miyashita Y & Chang HS Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature 331, 68–70 (1988). [PubMed: 3340148]
167. Scott BH, Mishkin M & Yin P Neural Correlates of Auditory Short-Term Memory in Rostral Superior Temporal Cortex. Current Biology 24, 2767–2775 (2014). [PubMed: 25456448]
168. Ranganath C, DeGutis J & D'Esposito M Category-specific modulation of inferior temporal activity during working memory encoding and maintenance. Brain Res. Cogn. Brain Res 20, 37–45 (2004). [PubMed: 15130587]
169. Lepsien J & Nobre AC Attentional modulation of object representations in working memory. Cereb. Cortex 17, 2072–2083 (2007). [PubMed: 17099066]
170. Bonnefond M & Jensen O Alpha oscillations serve to protect working memory maintenance against anticipated distracters. Curr. Biol 22, 1969–1974 (2012). [PubMed: 23041197]
171. Huang Y, Matysiak A, Heil P, König R & Brosch M Persistent neural activity in auditory cortex is related to auditory working memory in humans and nonhuman primates. Elife 5, (2016).
172. Wang L et al. Persistent neuronal firing in primary somatosensory cortex in the absence of working memory of trial-specific features of the sample stimuli in a haptic working memory task. J. Cogn. Neurosci 24, 664–676 (2012). [PubMed: 22098263]
173. Super H A Neural Correlate of Working Memory in the Monkey Primary Visual Cortex. Science 293, 120–124 (2001). [PubMed: 11441187]
174. Ester EF, Anderson DE, Serences JT & Awh E A neural measure of precision in visual working memory. J. Cogn. Neurosci 25, 754–761 (2013). [PubMed: 23469889]
175. Rahmati M, Saber GT & Curtis CE Population Dynamics of Early Visual Cortex during Working Memory. J. Cogn. Neurosci 30, 219–233 (2018). [PubMed: 28984524] This study uses fMRI to examine the precision with which encoding models can reconstruct WM representations in human visual cortex. The authors use an innovative method to model and quantify fMRI delay activity in order to demonstrate a link between the precision of model reconstruction and activity in higher-order parietal and frontal regions.
176. Woloszyn L & Sheinberg DL Neural dynamics in inferior temporal cortex during a visual working memory task. J. Neurosci 29, 5494–5507 (2009). [PubMed: 19403817]
177. Pasternak T & Greenlee MW Working memory in primate sensory systems. Nat. Rev. Neurosci 6, 97–107 (2005). [PubMed: 15654324]
178. D'Esposito M From cognitive to neural models of working memory. Philosophical Transactions of the Royal Society B: Biological Sciences 362, 761–772 (2007).
179. Xu Y Reevaluating the Sensory Account of Visual Working Memory Storage. Trends Cogn. Sci 21, 794–815 (2017). [PubMed: 28774684]
180. Miller EK, Erickson CA & Desimone R Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci 16, 5154–5167 (1996). [PubMed: 8756444]
181. Lorenc ES, Sreenivasan KK, Nee DE, Vandenbroucke ARE & D'Esposito M Flexible Coding of Visual Working Memory Representations during Distraction. J. Neurosci 38, 5267–5276 (2018). [PubMed: 29739867]
182. Scimeca JM, Kiyonaga A & D'Esposito M Reaffirming the Sensory Recruitment Account of Working Memory. Trends in Cognitive Sciences 22, 190–192 (2018). [PubMed: 29475635]
183. Wolff MJ, Jochim J, Akyürek EG & Stokes MG Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci 20, 864–871 (2017). [PubMed: 28414333]
184. O'Reilly RC & Frank MJ Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. Neural Computation 18, 283–328 (2006). [PubMed: 16378516]
185. Chatham CH, Frank MJ & Badre D Corticostriatal output gating during selection from working memory. Neuron 81, 930–942 (2014). [PubMed: 24559680]

186. Alexander GE Selective neuronal discharge in monkey putamen reflects intended direction of planned limb movements. Exp. Brain Res 67, 623–634 (1987). [PubMed: 3653320]

187. Johnstone S & Rolls ET Delay, discriminatory, and modality specific neurons in striatum and pallidum during short-term memory tasks. Brain Res. 522, 147–151 (1990). [PubMed: 2224509]

188. Soltysik S, Hull CD, Buchwald NA & Fekete T Single unit activity in basal ganglia of monkeys during performance of a delayed response task. Electroencephalography and Clinical Neurophysiology 39, 65–78 (1975). [PubMed: 50201]

189. Hikosaka O & Sakamoto M Cell activity in monkey caudate nucleus preceding saccadic eye movements. Exp. Brain Res 63, 659–662 (1986). [PubMed: 3758274]

190. Postle BR & D'Esposito M Dissociation of human caudate nucleus activity in spatial and nonspatial working memory: an event-related fMRI study. Cognitive Brain Research 8, 107–115 (1999). [PubMed: 10407200]

191. Postle BR & D'esposito M Spatial working memory activity of the caudate nucleus is sensitive to frame of reference. Cognitive, Affective, & Behavioral Neuroscience 3, 133–144 (2003).

192. Chang C, Crottaz-Herbette S & Menon V Temporal dynamics of basal ganglia response and connectivity during verbal working memory. Neuroimage 34, 1253–1269 (2007). [PubMed: 17175179]

193. Harrington DL, Zimbelman JL, Hinton SC & Rao SM Neural modulation of temporal encoding, maintenance, and decision processes. Cereb. Cortex 20, 1274–1285 (2010). [PubMed: 19778958]

194. Fuster JM & Alexander GE Firing changes in cells of the nucleus medialis dorsalis associated with delayed response behavior. Brain Res. 61, 79–91 (1973). [PubMed: 4204130]

195. Kubota K, Niki H & Goto A Thalamic unit activity and delayed alternation performance in the monkey. Acta Neurobiol. Exp 32, 177–192 (1972).

196. Watanabe Y & Funahashi S Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. II. Activity encoding visual versus motor signal. J. Neurophysiol 92, 1756–1769 (2004). [PubMed: 15140912]

197. Watanabe Y, Takeda K & Funahashi S Population vector analysis of primate mediodorsal thalamic activity during oculomotor delayed-response performance. Cereb. Cortex 19, 1313–1321 (2009). [PubMed: 18832329]

198. Watanabe Y & Funahashi S Thalamic mediodorsal nucleus and working memory. Neurosci. Biobehav. Rev 36, 134–142 (2012). [PubMed: 21605592]

199. Funahashi S Thalamic mediodorsal nucleus and its participation in spatial working memory processes: comparison with the prefrontal cortex. Front. Syst. Neurosci 7, 36 (2013). [PubMed: 23914160]

200. Klein JC et al. Topography of connections between human prefrontal cortex and mediodorsal thalamus studied with diffusion tractography. Neuroimage 51, 555–564 (2010). [PubMed: 20206702]

201. McFarland NR & Haber SN Thalamic relay nuclei of the basal ganglia form both reciprocal and nonreciprocal cortical connections, linking multiple frontal cortical areas. J. Neurosci 22, 8117–8132 (2002). [PubMed: 12223566]

202. Schmitt LI et al. Thalamic amplification of cortical connectivity sustains attentional control. Nature 545, 219–223 (2017). [PubMed: 28467827] This study, along with Ref 203, demonstrates that the thalamus has a key role in sustaining PFC delay spiking. Optogenetic suppression of thalamic delay activity eliminated sustained WM representations in PFC and impaired behavior.

203. Guo ZV et al. Maintenance of persistent activity in a frontal thalamocortical loop. Nature 545, 181–186 (2017). [PubMed: 28467817]

204. Bolkan SS et al. Thalamic projections sustain prefrontal activity during working memory maintenance. Nat. Neurosci 20, 987–996 (2017). [PubMed: 28481349]

205. Peräkylä J et al. Causal Evidence from Humans for the Role of Mediodorsal Nucleus of the Thalamus in Working Memory. J. Cogn. Neurosci 29, 2090–2102 (2017). [PubMed: 28777058]

206. Nakajima M & Halassa MM Thalamic control of functional cortical connectivity. Curr. Opin. Neurobiol 44, 127–131 (2017). [PubMed: 28486176]

207. Pergola G et al. The Regulatory Role of the Human Mediodorsal Thalamus. Trends Cogn. Sci 22, 1011–1025 (2018). [PubMed: 30236489]

208. Halassa MM & Kastner S Thalamic functions in distributed cognitive control. Nat. Neurosci 20, 1669–1679 (2017). [PubMed: 29184210]

209. Rikhye RV, Gilra A & Halassa MM Thalamic regulation of switching between cortical representations enables cognitive flexibility. Nat. Neurosci 21, 1753–1763 (2018). [PubMed: 30455456]

210. Naselaris T, Kay KN, Nishimoto S & Gallant JL Encoding and decoding in fMRI. Neuroimage 56, 400–410 (2011). [PubMed: 20691790]

211. Ester EF, Sprague TC & Serences JT Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron 87, 893–905 (2015). [PubMed: 26257053] This fMRI study uses encoding models to demonstrate stimulus-selective delay activity throughout the human dorsal visual hierarchy – most notably in frontal and parietal regions.

212. Sprague TC, Ester EF & Serences JT Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91, 694–707 (2016). [PubMed: 27497224]

213. Dotson NM, Hoffman SJ, Goodell B & Gray CM Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. Neuron 99, 215–226.e4 (2018). [PubMed: 29909999]

214. 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). [PubMed: 30100254] This paper represents an exciting attempt to reconcile evidence for WM storage-related delay activity in frontal and sensory regions. The key finding is that delay activity in these two regions reflects different behavioral strategies: prospective and action-oriented for frontal regions, and retrospective and sensory-oriented for sensory regions.

215. Chaudhuri R & Fiete I Computational principles of memory. Nat. Neurosci 19, 394–403 (2016). [PubMed: 26906506]

216. Durstewitz D, Seamans JK & Sejnowski TJ Neurocomputational models of working memory. Nat. Neurosci 3 Suppl, 1184–1191 (2000). [PubMed: 11127836]

217. Barak O & Tsodyks M Working models of working memory. Curr. Opin. Neurobiol 25, 20–24 (2014). [PubMed: 24709596]

218. Major G & Tank D Persistent neural activity: prevalence and mechanisms. Curr. Opin. Neurobiol 14, 675–684 (2004). [PubMed: 15582368]

219. Traub RD & Jefferys JG Are there unifying principles underlying the generation of epileptic afterdischarges in vitro? Prog. Brain Res 102, 383–394 (1994). [PubMed: 7800828]

220. Egorov AV, Hamam BN, Fransén E, Hasselmo ME & Alonso AA Graded persistent activity in entorhinal cortex neurons. Nature 420, 173–178 (2002). [PubMed: 12432392]

221. Navaroli VL, Zhao Y, Boguszewski P & Brown TH Muscarinic receptor activation enables persistent firing in pyramidal neurons from superficial layers of dorsal perirhinal cortex. Hippocampus 22, 1392–1404 (2012). [PubMed: 21956787]

222. Guigon E, Dorizzi B, Burnod Y & Schultz W Neural correlates of learning in the prefrontal cortex of the monkey: a predictive model. Cereb. Cortex 5, 135–147 (1995). [PubMed: 7620290]

223. Fransén E, Tahvildari B, Egorov AV, Hasselmo ME & Alonso AA Mechanism of graded persistent cellular activity of entorhinal cortex layer v neurons. Neuron 49, 735–746 (2006). [PubMed: 16504948]

224. Russo RE & Hounsgaard J Dynamics of intrinsic electrophysiological properties in spinal cord neurones. Prog. Biophys. Mol. Biol 72, 329–365 (1999). [PubMed: 10605293]

225. Diesmann M, Gewaltig MO & Aertsen A Stable propagation of synchronous spiking in cortical neural networks. Nature 402, 529–533 (1999). [PubMed: 10591212]

226. Rajan K, Harvey CD & Tank DW Recurrent Network Models of Sequence Generation and Memory. Neuron 90, 128–142 (2016). [PubMed: 26971945]

227. Goldman MS Memory without feedback in a neural network. Neuron 61, 621–634 (2009). [PubMed: 19249281]

228. Abeles M Corticonics. (1991). doi:10.1017/cbo9780511574566

229. Brody CD, Romo R & Kepecs A Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. Curr. Opin. Neurobiol 13, 204–211 (2003). [PubMed: 12744975]
230. Wimmer K, Nykamp DQ, Constantinidis C & Compte A Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci 17, 431–439 (2014). [PubMed: 24487232] This single-unit study is noteworthy for testing and confirming a fundamental prediction – that drift in population delay spiking should predict the direction and magnitude of behavioral errors – that was generated by a continuous attractor model. This work demonstrates how biophysical models can be used to inform empirical studies.
231. Itskov V, Hansel D & Tsodyks M Short-Term Facilitation may Stabilize Parametric Working Memory Trace. Front. Comput. Neurosci 5, 40 (2011). [PubMed: 22028690]
232. Koulakov AA, Raghavachari S, Kepecs A & Lisman JE Model for a robust neural integrator. Nat. Neurosci 5, 775–782 (2002). [PubMed: 12134153]
233. Inagaki HK, Fontolan L, Romani S & Svoboda K Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature 566, 212–217 (2019). [PubMed: 30728503]
234. Hansel D & Mato G Short-term plasticity explains irregular persistent activity in working memory tasks. J. Neurosci 33, 133–149 (2013). [PubMed: 23283328]
235. Mongillo G, Barak O & Tsodyks M Synaptic theory of working memory. Science 319, 1543–1546 (2008). [PubMed: 18339943]
236. Sugase-Miyamoto Y, Liu Z, Wiener MC, Optican LM & Richmond BJ Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. PLoS Comput. Biol 4, e1000073 (2008). [PubMed: 18464917]
237. Mi Y, Katkov M & Tsodyks M Synaptic Correlates of Working Memory Capacity. Neuron 93, 323–330 (2017). [PubMed: 28041884]
238. Wang Y et al. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nature Neuroscience 9, 534–542 (2006). [PubMed: 16547512]
239. Sandberg A, Tegnér J & Lansner A A working memory model based on fast Hebbian learning. Network 14, 789–802 (2003). [PubMed: 14653503]
240. Erickson MA, Maramara LA & Lisman J A single brief burst induces GluR1-dependent associative short-term potentiation: a potential mechanism for short-term memory. J. Cogn. Neurosci 22, 2530–2540 (2010). [PubMed: 19925206]
241. Fiebig F & Lansner A A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. J. Neurosci 37, 83–96 (2017). [PubMed: 28053032] This paper incorporates Hebbian STP into an attractor model in order to explain how synaptic and spiking delay mechanisms can be used to encode WM for multiple novel items.
242. Sreenivasan KK, Katz J & Jha AP Temporal characteristics of top-down modulations during working memory maintenance: an event-related potential study of the N170 component. J. Cogn. Neurosci 19, 1836–1844 (2007). [PubMed: 17958486]
243. Murray JD et al. A hierarchy of intrinsic timescales across primate cortex. Nat. Neurosci 17, 1661–1663 (2014). [PubMed: 25383900]
244. Schneegans S & Bays PM Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. J. Cogn. Neurosci 29, 1977–1994 (2017). [PubMed: 28820674]
245. Camperi M & Wang XJ A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. J. Comput. Neurosci 5, 383–405 (1998). [PubMed: 9877021]
246. Lundqvist M, Herman P & Lansner A Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. J. Cogn. Neurosci 23, 3008–3020 (2011). [PubMed: 21452933] The model described in this paper combines elements of STP with an attractor network model to recapitulate empirical LFP findings, including the relationship between LFP amplitude and WM load. The key advance of this model is that its architecture results in WM storage-related LFP bursting – a prediction that was later confirmed empirically in Refs 42 and 43.
247. Lisman JE, Fellous JM & Wang XJ A role for NMDA-receptor channels in working memory. Nat. Neurosci 1, 273–275 (1998). [PubMed: 10195158]

248. Orhan AE & Ma WJ A diverse range of factors affect the nature of neural representations underlying short-term memory. Nat. Neurosci 22, 275–283 (2019). [PubMed: 30664767]
249. Rose NS et al. Reactivation of latent working memories with transcranial magnetic stimulation. Science 354, 1136–1139 (2016). [PubMed: 27934762]
250. Myers NE, Stokes MG & Nobre AC Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. Trends Cogn. Sci 21, 449–461 (2017). [PubMed: 28454719]
251. Azevedo FAC et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol 513, 532–541 (2009). [PubMed: 19226510]
252. Ransom BR Neuroglia. Oxford Medicine Online (2013). doi:10.1093/med/9780199794591.001.0001
253. Araque A, Parpura V, Sanzgiri RP & Haydon PG Tripartite synapses: glia, the unacknowledged partner. Trends Neurosci. 22, 208–215 (1999). [PubMed: 10322493]
254. Santello M, Toni N & Volterra A Astrocyte function from information processing to cognition and cognitive impairment. Nat. Neurosci 22, 154–166 (2019). [PubMed: 30664773]
255. Vardjan N, Parpura V & Zorec R Loose excitation-secretion coupling in astrocytes. Glia 64, 655–667 (2016). [PubMed: 26358496]
256. Halassa MM et al. Astrocytic modulation of sleep homeostasis and cognitive consequences of sleep loss. Neuron 61, 213–219 (2009). [PubMed: 19186164]
257. Haydon PG GLIA: listening and talking to the synapse. Nat. Rev. Neurosci 2, 185–193 (2001). [PubMed: 11256079]
258. Papouin T, Dunphy J, Tolman M, Foley JC & Haydon PG Astrocytic control of synaptic function. Philos. Trans. R. Soc. Lond. B Biol. Sci 372, (2017).
259. Wang X et al. Astrocytic Ca2+ signaling evoked by sensory stimulation in vivo. Nat. Neurosci 9, 816–823 (2006). [PubMed: 16699507]
260. Schummers J, Yu H & Sur M Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. Science 320, 1638–1643 (2008). [PubMed: 18566287]
261. Lee HS et al. Astrocytes contribute to gamma oscillations and recognition memory. Proc. Natl. Acad. Sci. U. S. A 111, E3343–52 (2014). [PubMed: 25071179]
262. Pittà MD, De Pittà M, Ben-Jacob E & Berry H Astrocytic theory of working memory. BMC Neuroscience 15, (2014).
263. Wang XJ Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. J. Neurosci 19, 9587–9603 (1999). [PubMed: 10531461]
264. Aura J & Riekkinen P Jr. Blockade of NMDA receptors located at the dorsomedial prefrontal cortex impairs spatial working memory in rats. Neuroreport 10, 243–248 (1999). [PubMed: 10203316]
265. Verma A & Moghaddam B NMDA receptor antagonists impair prefrontal cortex function as assessed via spatial delayed alternation performance in rats: modulation by dopamine. J. Neurosci 16, 373–379 (1996). [PubMed: 8613804]
266. Baron SP & Wenger GR Effects of drugs of abuse on response accuracy and bias under a delayed matching-to-sample procedure in squirrel monkeys. Behav. Pharmacol 12, 247–256 (2001). [PubMed: 11548110]
267. Krystal JH et al. Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. Arch. Gen. Psychiatry 51, 199–214 (1994). [PubMed: 8122957]
268. Ghoneim MM, Hinrichs JV, Mewaldt SP & Petersen RC Ketamine: behavioral effects of subanesthetic doses. J. Clin. Psychopharmacol 5, 70–77 (1985). [PubMed: 3988972]
269. Driesen NR et al. The impact of NMDA receptor blockade on human working memory-related prefrontal function and connectivity. Neuropsychopharmacology 38, 2613–2622 (2013). [PubMed: 23856634]
270. Dudkin KN, Kruchinin VK & Chueva IV Effect of NMDA on the activity of cortical glutaminergic structures in delayed visual differentiation in monkeys. Neurosci. Behav. Physiol 27, 153–158 (1997). [PubMed: 9168485]

271. Wang M et al. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. Neuron 77, 736–749 (2013). [PubMed: 23439125]

272. Wang H, Stradtman GG, -J. Wang X & -J. Gao W A specialized NMDA receptor function in layer 5 recurrent microcircuitry of the adult rat prefrontal cortex. Proceedings of the National Academy of Sciences 105, 16791–16796 (2008).

273. McQuail JA et al. NR2A-Containing NMDARs in the Prefrontal Cortex Are Required for Working Memory and Associated with Age-Related Cognitive Decline. J. Neurosci 36, 12537–12548 (2016). [PubMed: 27807032]

274. Buzsáki G & Wang X-J Mechanisms of gamma oscillations. Annu. Rev. Neurosci 35, 203–225 (2012). [PubMed: 22443509]

275. Wang J D1 Dopamine Receptors Potentiate NMDA-mediated Excitability Increase in Layer V Prefrontal Cortical Pyramidal Neurons. Cerebral Cortex 11, 452–462 (2001). [PubMed: 11313297]

276. Williams GV & Goldman-Rakic PS Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. Nature 376, 572–575 (1995). [PubMed: 7637804]

277. Durstewitz D, Seamans JK & Sejnowski TJ Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. J. Neurophysiol 83, 1733–1750 (2000). [PubMed: 10712493]

278. Arnsten AF, Cai JX, Murphy BL & Goldman-Rakic PS Dopamine D1 receptor mechanisms in the cognitive performance of young adult and aged monkeys. Psychopharmacology 116, 143–151 (1994). [PubMed: 7862943]

279. Wang M et al. Alpha2A-adrenoceptors strengthen working memory networks by inhibiting cAMP-HCN channel signaling in prefrontal cortex. Cell 129, 397–410 (2007). [PubMed: 17448997]

280. Thuault SJ et al. Prefrontal cortex HCN1 channels enable intrinsic persistent neural firing and executive memory function. J. Neurosci 33, 13583–13599 (2013). [PubMed: 23966682]

281. Zhang Z, Cordeiro Matos S, Jego S, Adamantidis A & Séguéla P Norepinephrine drives persistent activity in prefrontal cortex via synergistic α1 and α2 adrenoceptors. PLoS One 8, e66122 (2013). [PubMed: 23785477]

282. Neymotin SA et al. Calcium regulation of HCN channels supports persistent activity in a multiscale model of neocortex. Neuroscience 316, 344–366 (2016). [PubMed: 26746357]

283. Arnsten AFT Stress signalling pathways that impair prefrontal cortex structure and function. Nat. Rev. Neurosci 10, 410–422 (2009). [PubMed: 19455173]

284. Ollinger JM, Shulman GL & Corbetta M Separating processes within a trial in event-related functional MRI I. The Method. Neuroimage 13, 210–217 (2001). [PubMed: 11133323]

285. Ruge H, Goschke T & Braver TS Separating event-related BOLD components within trials: the partial-trial design revisited. Neuroimage 47, 501–513 (2009). [PubMed: 19422920]

Box 1 |

Do astrocytes contribute to delay activity?

Astrocytes, the most abundant glial cell in the brain, are comparable in number to neurons[251]. Classical views of astrocytic function proposed supportive roles, including nutritional support of neurons, the maintenance of ion concentrations in the extracellular space, support of the blood–brain barrier and the repair of injured brain tissue[252]. In the 1990s, astrocytic function was reconsidered, with the idea of the 'tripartite synapse' put forth based on empirical evidence that astrocytes can integrate neuronal inputs and modulate synaptic activity[253]. Particularly in light of recent findings linking blood-oxygen-level-dependent responses and astrocytic function[65], as well as increased appreciation for the role of astrocytes in cognition[254], one must consider how astrocytes might contribute to delay activity. Although there is currently no published empirical data linking astrocytes to delay activity, three features make astrocytes an intriguing candidate for participating in delay activity.

First, and most importantly, although astrocytes respond to fast neural dynamics, they process information over a much slower time scale than do neurons (for example, over tens of milliseconds to seconds versus sub-millisecond to milliseconds)[255], suggesting that they may promote prolonged brain states such as sleep, potentially by modulating neuronal activity via gliotransmission[256]. Indeed, slow astrocytic dynamics may overcome some of the challenges for sustaining information over seconds that are imposed by the short time constants of neurons.

Second, as astrocytes enwrap nerve terminals, they are perfectly positioned to sense and emit informative signals from and to synapses[257]. They can sense neural activity and are activated during synaptic transmission and can in turn modulate neuronal activity by releasing transmitters such as glutamate[258]. For example, cystolic calcium levels in astrocytes of the mouse barrel cortex increase following whisker stimulation, in line with the frequency of stimulation[259]. Likewise, in ferret visual cortex, astrocytes respond to visual stimuli, and show tuning to stimulus features such as orientation and spatial frequency[260]. This modulation of activity in astrocytes as a function of stimulus properties suggests that they may be able to encode information about memoranda.

Third, in addition to being involved in localized neuronal activity, astrocytes also regulate network-level activity. Inhibition of glutamate release by astrocytes significantly reduced the duration of gamma oscillations in hippocampal slices, as well as gamma power in awake behaving mice, which corresponded with behavioural performance on a recognition memory task[261]. Thus, gliotransmission may potentially similarly contribute to gamma oscillations in WM.

These new insights into astrocytic function have led to the proposal that astrocytes may have a role in WM. One study provided a computational model that simulated individual cortical neurons in a delay task. Bistable delay spiking emerged from short-term (on the order of several seconds) synaptic facilitation that was initiated by the cue stimulus and mediated by astrocytes[262]. This model also recapitulated the finding of irregular patterns of delay spiking by individual neurons (see main text). Thus, astrocytes may have an

essential role in WM by promoting cellular bistability via synaptic plasticity. Future models of WM should take astrocytic function into account.

Box 2 |

The cellular basis of delay activity

Computational models have long theorized that NMDA receptors (NMDARs) have a prominent role in delay spiking[216,247,263]. The relatively slow excitatory dynamics of NMDAR activation can provide a stabilizing influence on attractor networks, allowing these networks to sustain a representation through persistent spiking over several seconds[263]. These models accord well with behavioural evidence that systemic administration of NMDAR antagonists disrupts WM in rats[264,265], non-human primates (NHPs)[266] and humans[267–269], whereas perfusion of NMDA in monkey visual cortex improves WM performance[270]. This hypothesized role of NMDARs in delay spiking was recently confirmed empirically: blockade or antagonism of prefrontal NMDARs eliminated delay spiking in NHPs[271]. This effect was specific to the NR2B subunit, presumably owing to its slow kinetics[17,272] (although see Ref.[273] for a discussion of the importance of NR2A versus NR2B subunits).

NMDARs have also been implicated in fMRI and EFP measures of delay activity. Systemic administration of the NMDAR antagonist ketamine in humans reduced BOLD responses in lPFC during the early delay period in a WM task[269]. Moreover, NMDAR dynamics, in concert with the contributions of AMPA receptors to recurrent excitation, have been found to give rise to sustained gamma frequency oscillations during the delay period in an attractor network model[19], consistent with the role for NMDARs in generating gamma oscillations[274]. NMDARs may also help to mediate neuromodulatory effects on WM. Stimulation of dopamine D1 receptors facilitates the activity of NMDARs in prefrontal neurons in vitro[275], which may be the mechanism by which dopamine stabilizes or enhances delay spiking in lPFC[276,277], thus promoting more stable WM[278].

In addition to NMDARs, hyperpolarization-activated cyclic nucleotide-gated (HCN) channels — specifically, HCN1 channels — have been implicated in delay spiking, although their exact role is debated. One study in NHP lPFC demonstrated that α2A-adrenoreceptor antagonism indirectly potentiates HCN1 channels, leading to an attenuation of delay spiking, whereas HCN1 blockade enhances delay spiking[279]. By contrast, a study of mouse PFC neurons in vitro showed that HCN1 channels were crucial for delay activity; facilitation of the HCN1 channel-mediated current resulted in intrinsic (that is, connectivity-independent) persistent spiking[280,281] (see Ref.[282] for a model that examines the influence of HCN channels on network and cell-intrinsic mechanisms of persistent activity). HCN1 channel activity is influenced by dopamine, noradrenaline and acetylcholine[280,281–283], making these channels another important target for understanding neuromodulatory effects on delay activity.

Box 3 |

Measuring delay activity with functional MRI

The temporal resolution of functional MRI presents a challenge for studying delay activity in humans. The blood-oxygen-level-dependent (BOLD) signal represents the transformation of neural activity to a **haemodynamic response**, and therefore acts as a low-pass filter with a response that peaks several seconds after an isolated neural event. Consequently, in a typical working-memory delay task consisting of a to-be-remembered sample stimulus, an empty delay period and a probe and/or response, accurately estimating the magnitude of delay activity with fMRI is nontrivial. Although analytical methods for estimating delay-period activity from the BOLD signal were introduced more than 20 years ago[57,58], there does not seem to be a consensus among strategies used in contemporary empirical studies. Here, we review the three most common approaches.

Fixed delays modelled with a boxcar function

The most widely used approach uses a long delay period (typically 10–15 s) that is fixed in length across trials. To estimate the magnitude of delay activity, the BOLD response during the delay is modelled with a **general linear model (GLM)** as a single event that spans the entire delay period (that is, a boxcar function, usually convolved with the haemodynamic response function). In part **a** of the figure, the left column shows the three predictors — sample, delay and response. The right column shows the data (navy line) and the three predictors convolved with the haemodynamic response function. This method accurately estimates the magnitude of activity during the delay but is unable to independently estimate the contributions of encoding-, maintenance- and response-related activity. As a consequence, any estimate of delay activity is necessarily contaminated by activity during encoding and response, and caution is required when interpreting the results.

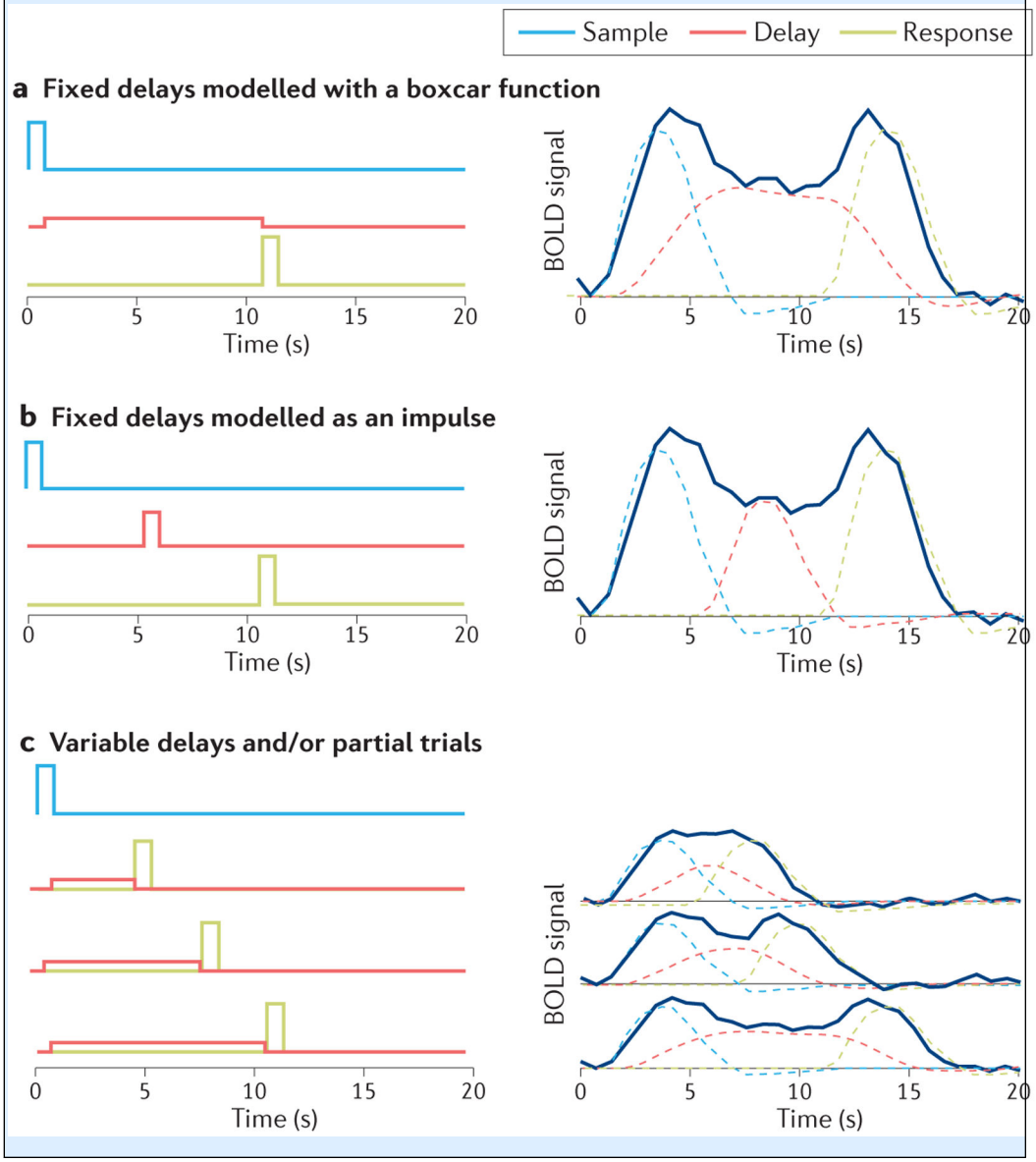
Fixed delays modelled as an impulse

Some GLMs model the BOLD signal during the delay period as a single **impulse response function** (or 'event') centred in the middle of the delay period and convolved with the haemodynamic response function (part **b** of the figure). Separating the delay event from the sample presentation and probe or response by 4 s or more improves the ability to estimate the independent contribution of each event to the BOLD signal[58]. Therefore, designing experiments with delay periods longer than 8 s and analysing the data with a GLM that places the delay-related impulse response function at least 4 s after the offset of the sample stimulus provides an estimation of the magnitude of delay activity that is reasonably independent of the activity associated with encoding or response. However, this method tends to sacrifice accuracy of the estimation.

Variable delays and/or partial trials

A more recent strategy is to use variable delay lengths (part **c**). Varying the length of the delay helps to reduce the statistical dependencies between the different trial components, and allows for an independent and accurate estimation of delay activity magnitude. The use of partial trials, which terminate after the sample or the delay period, has similar

benefits[284,285]. The cost of this method is that it complicates efforts to estimate the magnitude of delay period activity for single trials, as is typically required for population analyses (although see Ref.[175] for a notable exception).



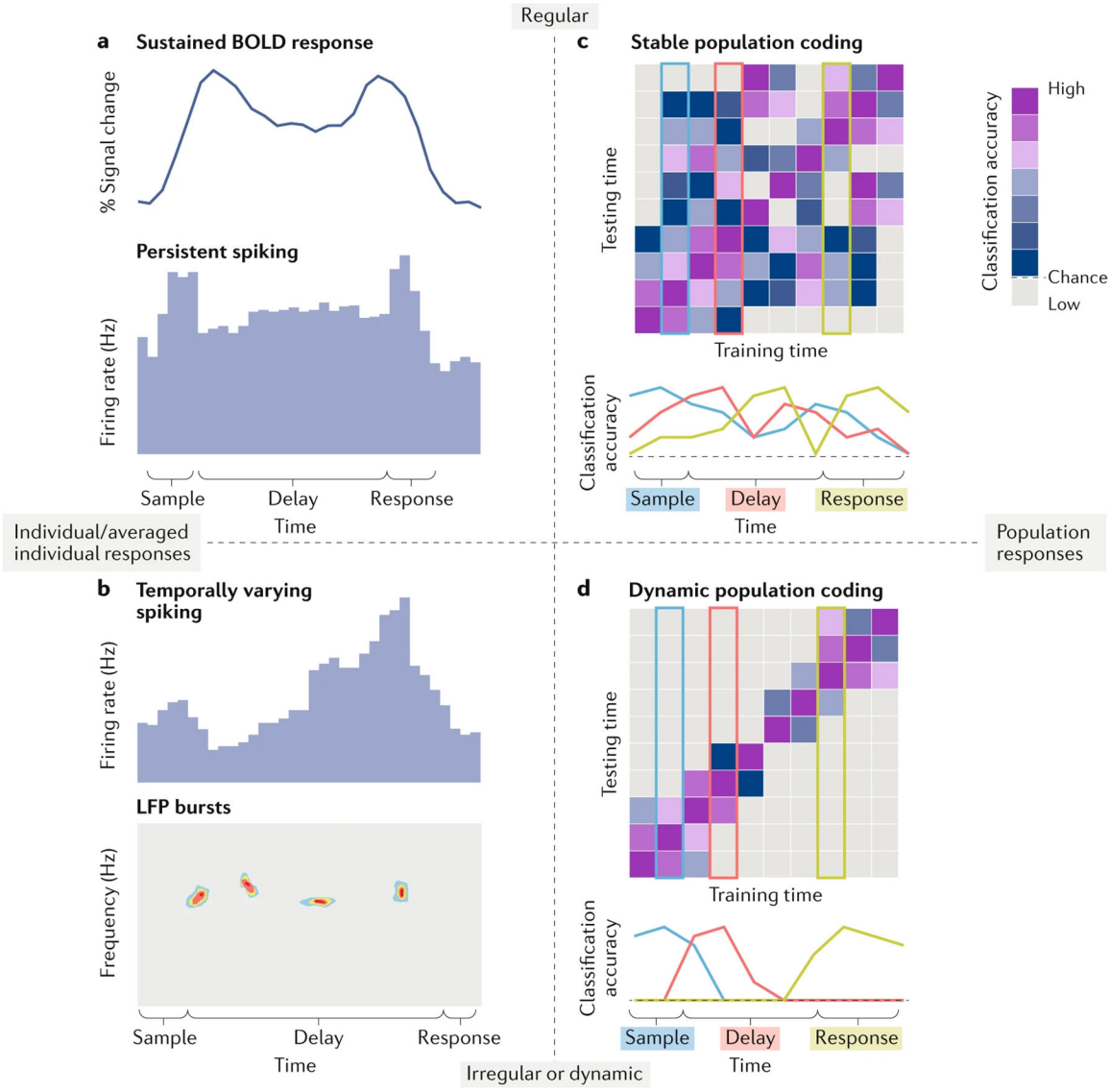


Fig. 1 |. Schematic examples of different types of delay activity.

Here we use schematics to highlight two properties of delay activity. First, delay activity can be stable in time (top row), or it can be temporally irregular or dynamic (bottom row). Second, delay activity can be measured within or averaged over individual neurons, voxels or electrodes (left column), or measured as the combined response across populations of neurons/voxels/electrodes (right column). **a** | Temporally stable delay activity in individual neurons or blood-oxygen-level-dependent (BOLD) response averaged over functional MRI voxels. Activity remains elevated above baseline throughout the delay period, signifying the sustained representation of information. **b** | Temporally irregular responses in individual neurons or electrodes. Individual neurons can display spiking activity that varies over the course of the delay (schematicized at the top of this panel). Recent results demonstrate intermittent bursts in the LFP signal throughout the delay (schematicized at the bottom)[41]. **c** | Stable population coding. WM information (for example, WM content, task rules or planned responses) can be decoded from the combined activity of populations of neurons

or voxels. A pattern classifier can be trained and tested on independent data sets recorded during a WM task. Above-chance classification accuracy indicates that a representation of the information being classified exists in those neurons or voxels. The classifier can be trained on data from a specific time (for example, during the early delay) in the trial, and tested (on independent data) from the same time point or different time points, allowing one to measure the stability of the representation over the course of WM maintenance. In stable population codes, the pattern of activity that encodes specific information at any given time point is the same as the pattern of activity that encodes that same information at any other time point; thus, a classifier trained at one time point will be more accurate than chance at other time points throughout the trial. This is indicated by above chance classification at each time point regardless of which time point or period (for example, those represented in light blue, red or green) the classifier is trained on. **d** | Dynamic population coding. In contrast to stable population coding, information encoded in the population activity at a certain time point is encoded in different forms of activity at other time points. A classifier trained on a time point will therefore perform successfully at that time point, but not at other time points in the trial. This time-dependent classification is indicated by above-chance classification accuracy along the diagonal of the training-by-testing matrix (top; compare to the matrix in **c**), but chance classification elsewhere in the matrix, and above-chance classification accuracy that is limited to brief periods of time for classifiers trained on the sample, delay and response periods (bottom; light blue, red and green, respectively).

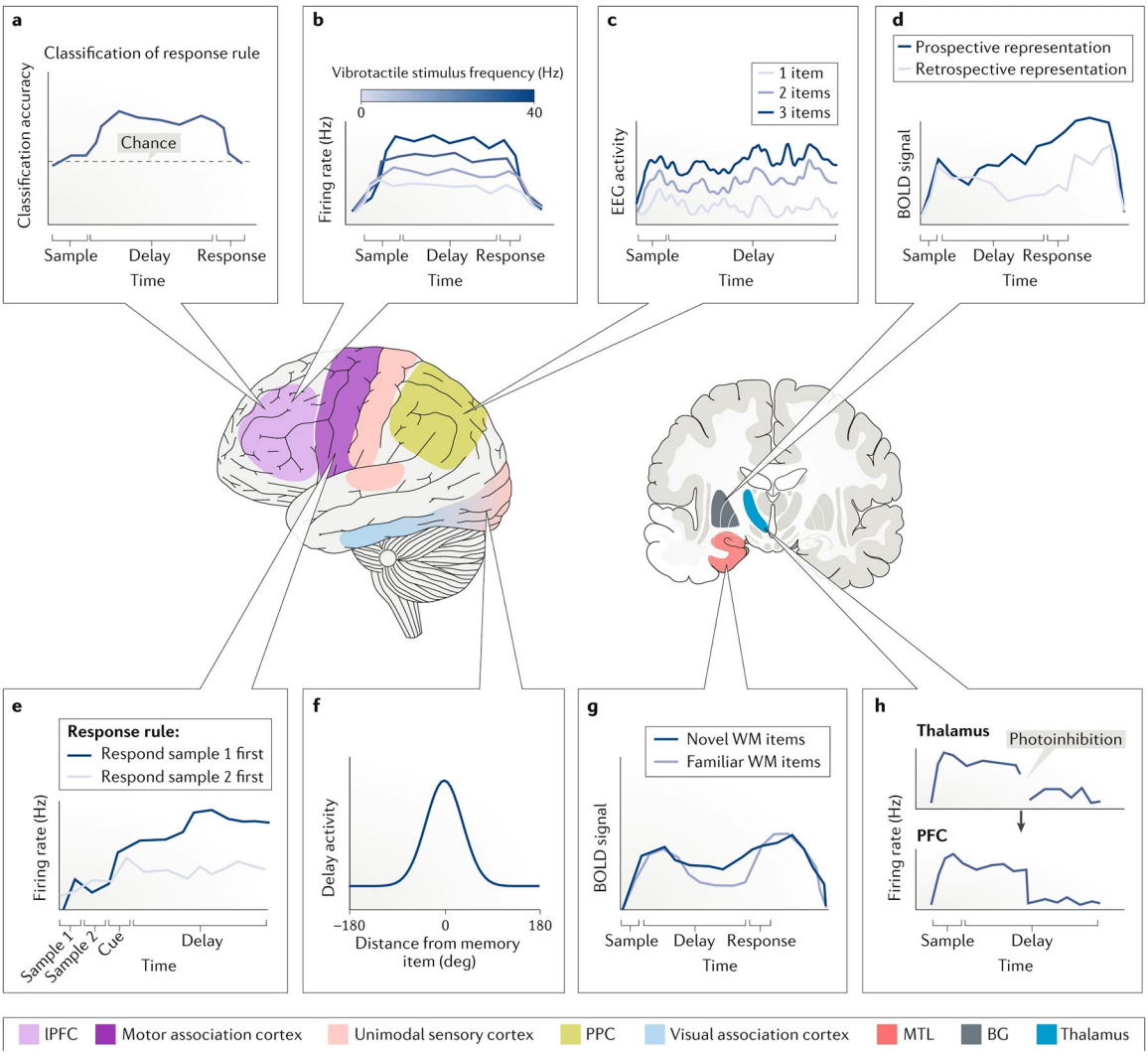


Fig. 2 |. Schematic depictions of the properties of delay activity in different brain regions.

In each panel, simplified schematic diagrams are provided to illustrate the type of information represented by delay activity in different parts of the brain. Note that not all of the findings were shown in humans, but the regions have been depicted on a human brain for illustrative purposes. **a** | Delay activity in the lateral prefrontal cortex (lPFC) represents task rules. Population analyses involving pattern classification approaches have demonstrated that classification of the task rule (for example, which feature of a memorandum is relevant for response[75]) is above chance during the memory delay, suggesting that lPFC delay activity represents aspects of task rules. **b** | lPFC delay activity represents working memory (WM) content. Neurons in the lPFC of non-human primates (NHPs) exhibit delay-period activity that varies in spike frequency with the properties of the memory stimulus, such as vibrotactile frequency[114]. **c** | Electroencephalogram (EEG) electrodes over PPC reveal delay activity that increases in magnitude and oscillatory power with WM load, and plateaus at an individual's WM capacity[27]. This is consistent with the notions that PPC delay activity encodes WM content and that PPC delay activity represents internal attention directed to items in WM. Challenges in localizing EEG activity make it unclear where in the

brain these signals originate. **d** | Basal ganglia (BG) delay activity is associated with the upcoming behavioural response. BG delay activity is greater in magnitude when participants can anticipate the upcoming response (prospective representation) than when they cannot (retrospective representation)[191]. **e** | Neurons in motor association cortex show preference for specific response rules. Tasks that require NHPs to maintain response rules (for example, which memory item to indicate first with a behavioural response[158]), suggest that this region encodes the anticipation of specific planned responses. **f** | Population activity in visual sensory regions is tuned to features of WM content. Functional MRI studies can identify meso-scale responses to features (such as orientation) of memory items, and have found that populations that represent features of WM content are preferentially active during the delay[174]. **g** | Delay activity in the medial temporal lobe (MTL) may be involved in storing complex information in WM. The magnitude of MTL delay activity is larger for novel items than for familiar items[153], and MTL delay activity is often observed during WM for complex items[15]. **h** | The thalamus exhibits delay activity that seems to drive delay responses in PFC. Experimental disruption of thalamic delay activity in mice resulted in reduced or abolished delay activity in PFC[202,203].

Table 1 |

Features and limitations of models of delay activity

Class of model	Notable features	Limitations	Refs
Spike-based network models			
Feedforward networks	• Do not require recurrent connections (although similar activity can also be achieved in recurrent networks[227,227]) • Exhibit both sequential and stable patterns of delay spiking	• Timing precision required for feedforward signals may render these models more susceptible to noise	225, 228
Attractor networks	• Can describe network delay activity tuned to features of WM memoranda • Can encode discrete or continuous quantities and mimic parametric WM[114] • Disruptions in attractor network activity may account for behavioural errors[230,233]	• Slight changes in parameters lead to instability • Precise network connectivity required for these models makes it difficult to account for WM for novel items	18,19
Other			
Synaptic mechanisms	• Do not require persistent neural activity to sustain WM representations • Can explain absence or recovery of delay activity	• Unable to explain WM for novel items without STP[241] • Difficult to confirm absence of delay activity experimentally	235–237
Cellular bistability	• Can achieve two or more levels of above-baseline delay firing that can encode features of WM stimuli • Do not require complex network connectivity to generate delay activity • Can recapitulate delay activity for novel items • Can potentially encode parametric WM[114]	• Limited ability to exhibit complex temporal dynamics • Can represent discrete but not continuous quantities	222, 223
Hybrid models			
Cellular bistability plus recurrent network architecture	• Network can be more stable to initial conditions and perturbation • Bistable cells can help network encode novel items	• Unable to display complex single-cell spiking dynamics	232, 245, 247
Attractor networks plus STP	• Can increase resistance of the network to perturbation and drift • Firing patterns of individual cells capture complex temporal dynamics, including network oscillations and LFP bursts • Allows for encoding of multiple items and novel items	• Many of these networks still require specific architecture that involves extensive learning	231, 234, 241, 246