HHS Public Access

Author manuscript

Trends Neurosci. Author manuscript; available in PMC 2022 May 10.

Published in final edited form as:

Trends Neurosci. 2021 November ; 44(11): 888–902. doi:10.1016/j.tins.2021.09.001.

50 Years of Persistent Activity: Quo Vadis?

Xiao-Jing Wang

Center for Neural Science, New York University, 4 Washington Place, New York, NY 20003, USA.

Abstract

Persistent spiking activity in the neocortex was discovered a half century ago as a neural substrate of working memory. Research on its brain mechanism has strived for understanding a core cognitive function across biological and computational levels. Here I review studies that cumulatively lend support to a synaptic theory of recurrent circuits for mnemonic persistent activity that depend on various cellular and network substrates, mathematically described by a multiple-attractor network model. An attractor state is consistent with temporal variations and heterogeneity across neurons in a subspace of population activity. Activity-silent state mechanisms are suitable for storing passive short-term memory traces, but not working memory characterized by executive control for filtering our distractors, limited capacity and internal manipulation of information.

Editorial note:

In view of past scientific affiliation between the author and the current editor of *Trends in Neurosciences*, editorial handling of this manuscript and management of peer-review were conducted by Dr. Lindsey Drayton, editor of *Trends in Cognitive Sciences*.

Keywords

working memory; persistent activity; multiple-attractor network model; NMDA receptor; diverse interneuron types; activity-silent state; short-term memory; subspace analysis; cognition; psychiatry

Mnemonic persistent activity as an atom of cognition

The year 2021 marks the fiftieth anniversary of the discovery that single-cell persistent activity is associated with working memory. The story of this discovery began in the 1960s when Joaquin M. Fuster happened to make the acquaintance of Larry Ott, an engineer at Hughes Aircraft who invented a cryogenic device used to cool the electronic components of space satellites. At that time, Fuster was impressed by the studies of C. F. Jacobsen and others showing that lesioning the prefrontal cortex (PFC) impaired macaque monkeys' performance in a delayed response task [1, 2]. In a typical delayed response task, a sensory stimulus (e.g., green visual object) and an appropriate response (go) are separated by a short time interval (delay period). Consequently, the probed behavior depends on working

Corresponding author: xjwang@nyu.edu.

memory, the brain's ability to hold and manipulate information when sensory stimulation is absent [3, 4]. Could Ott's new gadget help neuroscientists study the brain mechanisms supporting working memory? Fuster and his colleague Garrett Alexander adopted the cryogenic device to inactivate by cooling circumscribed brain regions of monkeys in a delayed response task [5]. They then proceeded to neurophysiological recordings, which revealed that a substantial number of prefrontal units showed persistent elevations of firing rate during the delay, the memory retention period of the task (for Fuster's recollection, see Box 1). The resulting publication in 1971 [6] and another independent publication that same year [7] ushered in additional single-neuron investigations of brain circuits underlying working memory.

This article takes stock of the last fifty years of research exploring persistent neural activity as it pertains to the foundation of working memory. This work has provided substantial support for the multiple-attractor network model of self-sustained mnemonic persistent activity. The central tenet of this theory is that a memory representation is not a transient signal that passively decays away in time, instead, it corresponds to a dynamically stable state of the brain. A working memory system is in turn conceptualized as a neural circuit endowed with multiple attractor states encoding different memory items that coexist with a baseline state. As an analogy, imagine a hilly golf course with many valleys, akin to a state space of neural population activity in a working memory system. The bottom (attractor) of a valley (basin of attraction) is "attractive" in the sense that a ball (the position of which corresponds to the state of the neural system) naturally rolls down towards it. This way, a sufficiently large transient input (hitting hard a ball to the air) can switch the system from rest (one valley) to a stimulus-selective mnemonic state (a different valley) which remains after stimulus withdrawal; such a state is robust against small perturbations (gentile taps of the ball with a club). A subsequent brief but potent signal can switch the system back to the resting state, thereby erasing a memory trace. Unlike a golf course, however, attractors in a neural system may be characterized by complex spatiotemporal patterns such as stochastic network oscillations or propagation waves (sequential activation of different neural groups) rather than steady states. Furthermore, the landscape of multiple attractors is readily modifiable by a sustained input, which is essential for executive control of working memory.

Here, I will first review studies that cumulatively lend support to the recurrent neural circuit mechanism of working memory representation, mathematically corresponding to the multiple-attractor network model of persistent activity. This theoretical framework predicts that (1) mnemonic activity is maintained over time when the delay period duration is varied considerably, (2) intracellular current injection cannot switch off persistent activity of a neuron engaged in working memory, and (3) after a brief optogenetic perturbation persistent activity reverts to the same pattern in the control condition. These predictions have recently received experimentally confirmations in behaving animals. In the sections that follow, I discuss developments that address some recent challenges to the theory and suggest areas for future work.

An attractor network model of persistent activity

Following the original discovery, studies of single-neuron recording in delay dependent tasks have documented persistent activity encoding discrete items (visual objects, categories, task rules) [8,9,10,11,12, 13, 14] and continuous space [15,16,17, 18, 19, 20, 21, 22]; parametric working memory with monotonical encoding of a behavioral attribute was discovered in a vibrotactile delayed discrimination (VDD) task [23]. These experiments identified the PFC (especially its superficial layers [24]), the posterior parietal cortex (PPC), and other brain regions engaged in working memory representation. Functional magnetic resonance imaging (fMRI) uncovered similar brain structures activated by working memory in humans [25], also differentially engaging the superficial layers [26]. In close interplay with experimentation, neural network models for stimulus-selective persistent activity were developed. Following pioneering work [28, 29], self-sustained memory states began to be conceptualized as attractor states [30, 31]. Mathematically, an attractor denotes a state of a nonlinear dynamical system that is stable such that after a small transient perturbation the system will revert to the original state [32].

In the late 90s and early 2000s, the attractor network paradigm was tested using spiking neural network models endowed with biologically constrained synaptic connections [37, 38, 39, 40, 41, 42]. These studies provided initial support for the attractor network model (see review in [31]). Has the attractor network model stood the test of time over the last twenty years? Biologically, a mnemonic attractor is sustained by reverberatory dynamics through feedback loops in a neural assembly [30,31]. One early theoretical prediction was that the posited reverberation must be slow and dependent on the NMDA receptors at local recurrent excitatory synapses in a working memory circuit [38]. This model prediction was confirmed in experiments where iontophoresis of an antagonist for NR2B-subunit containing NMDA receptors essentially abolished mnemonic persistent activity in PFC neurons recorded from monkeys performing an ODR task [43]. Subsequent studies showed that both the NMDA and AMPA receptors contributed to working memory function, with the fast AMPA receptors predominantly signaling sensory information [44, 45]. Another model prediction was a disinhibitory motif composed of three types of inhibitory neurons for gating access to working memory and filtering out distractors [46]. This theoretical prediction has been supported experimentally and shown to be a canonical feature of the neocortex (reviewed in [47, 48]).

The theoretical finding that NMDA receptors play a critical role in working memory offered an example of how a core cognitive function can be elucidated in neuroscience across levels, from receptors to recurrent neural circuit dynamics to function. It also explained why low dose ketamine, an NMDA receptor antagonist, could induce in healthy subjects working memory deficits [49] similar to those observed in schizophrenic subjects, who experience NMDA receptor hypofunction [50, 51, 52]. This insight helped prompt the emergence of the field of Computational Psychiatry [53, 54]. Slow reverberation is also suitable for temporal accumulation of evidence to inform decision-making [55,56, 57], suggesting a shared mechanism for working memory and decision-making in “cognitive-type” neural circuits [58, 59].

Rigorous experimental tests of the attractor network model of working memory became possible only recently thanks to advances of experimental tools such as cell-type specific optogenetic manipulation. In a mouse experiment, subjects learn to associate one of two sensory cues to left and right licking responses. The two sensory stimuli may be somatosensory (far and near objects that touch whiskers) or auditory (high and low tones; Figure 1a). Before the response is allowed to take place, there is a short delay period. Single neurons in the premotor area called anterior lateral motor (ALM) cortex display elevated firing activity during the delay period. A series of experiments, in close interplay with computational modeling, have led to a wealth of information about the underlying neural circuit mechanisms supporting short-term memory in this task. First, optogenetic inactivation systematically done across the cortex demonstrated that ALM is the crucial node for maintaining short-term memory [60]. Second, if persistent activity is a single-cell phenomenon rather than maintained by synaptic reverberation, current injection into a cell should be able to turn off ongoing persistent activity [61]. This was not found to be the case using intracellular recording in behaving animals during a delay period [62], in support of a network mechanism. Third, despite optogenetic perturbations that transiently alter the time course of the ALM neural firing, the trajectory of population activity converges to one of two fixed endpoints in the state space of recorded neural population activity, in support of discrete attractor models (Figure 1b) [63, 61]. Fourth, optogenetic inactivation during the delay period revealed that thalamo-cortical connections are important for the maintenance of delay period activity in the ALM [64].

In this task, because the sensori-motor transformation presumably occurs during external stimulation, persistent firing in the premotor area ALM encodes preparation for the impending movement rather than sensory working memory. This differs from other tasks, like the DMS, which require that delay activity represents the sample stimulus because the correct motor response is unknown (and thus cannot be prepared during the delay period). Using delay dependent tasks where remembering sensory information is essential, other rodent experiments found that frontal and parietal areas are engaged in working memory dependent behavior [65]. Results from neural data analyses and experimental manipulations combined with modeling lend further support to the attractor network paradigm [66, 67, 68]. Moreover, parametric working memory can be modeled as line attractors [69, 70], akin to a flat part of a golf course where the ball can stay at a continuum of positions. Finally, attractor models have also been extended to account for multiple-item working memory [71].

Dynamical coding and heterogenous delay activity

Although the attractor model has received theoretical and empirical support, it has been challenged on the ground that mnemonic neural activity often varies substantially over a delay period. In a working memory task, neurons in a cortical area tend to display temporal variations during the delay period [72, 73]. A relatively small number (5–10%) of recorded neurons show strictly tonic persistent activity. Others display time-varying patterns: some ramp-up while others ramp down their firing rates in time during the delay [74]. The percentage of sampled neurons showing delay period activity can be 30% or higher depending on the precise recording location, [11, 75]. Note that a brain region is engaged

in many tasks, thus the number of cells activated in a single task could be a small but significant fraction of the entire population. In delayed response tasks, persistent activity was reduced or absent in error trials [11, 21], in support of its importance at the behavioral level.

Critically, temporal changes of delay period activity, *per se*, are compatible with attractor network models. The misconception that an attractor must be in a steady state may result from the mere fact that mathematical models are easiest to describe and analyse if attractors are steady states [37, 38, 41, 76]. But attractor states do not have to be stationary, as illustrated by stimulus-selective attractors characterized by stochastic oscillations [39, 76] which have been observed in behaving monkeys during working memory tasks [18]. Chaotic attractors [32] may also support persistent activity [77, 78]. In principle, an attractor of a dynamical system may display complex spatiotemporal patterns, exemplified by fluid turbulence with vortices over many scales in space and time.

A more puzzling finding is that stimulus selectivity of a recorded neuron may be detectable only in a brief portion of the delay period, and each cell shows statistically significant selectivity at different times ([70] but see [80]). A method to quantify whether a working memory representation is stationary or time-varying is to train a linear classifier at time t to decode information from recorded neurons, which is then used to decode the stimulus at time t^0, thus the quality of decoding is shown in a two dimensional "cross-temporal classification matrix" [81]. Figure 2a shows such a matrix computed using 600 PFC neurons in a monkey delay dependent experiment [82]. During the cue presentation, reliable decoding (red to orange color) is confined near the diagonal line, which means that the classifier trained at a particular time cannot decode the trial type at a different time. On the other hand, during the delay period following the initial cue, good decoding fills a square, demonstrating that working memory representation is quite stable over time.

Studies using cross-temporal classification analysis yielded various cross-temporal classification matrices [85, 86]. In general, working memory representations are stable over time in tasks that mostly involve memory maintenance, but time-varying when information processing and manipulation are required during the delay period; sometimes a code is stable in a time window, then evolves into time-varying in another time window, yielding a mixture of stable code and dynamical code [87].

Can temporal variations of neural activity be compatible with a stable working memory representation during a delay period? To address this question, a principal component analysis (PCA) was applied to PFC neural trajectories using data from ODR [11] and VDD [23] monkey experiments [88]. This analysis revealed that single neurons display various temporal patterns in their delay period activities (Figure 3a). However, population coding of a stimulus stored in working memory is stable within a subspace where working memory coding is stationary, despite considerable temporal changes in the orthogonal subspace (Figure 3b). This observation was reproduced in attractor network models [89, 88]. In conclusion, temporal variations of delay period neural activities can be reconciled with a stable working memory representation over time in a low-dimensional subspace or manifold of neural population activity.

Firing activity in a delay period may move among different neural groups. In rodents, several studies found temporal "tiling" of a delay period by transiently active neurons [83, 90, 91]. In one mouse experiment, delay period activity of neurons (monitored by calcium imaging) in the posterior parietal cortex was transient rather than tonic: each firing cell briefly peaked at a different time of the delay period (Figure 2b) [83], demonstrating sequential activation of neural groups [92, 93]. Such a delay period activity pattern is incompatible with a stationary code. Transient activities were also found in the mice anterior agranular insular cortex in another delay dependent experiment [91]. On the other hand, in the aforementioned delayed response task (Figure 1), analysis of peak times of spiking activity of recorded neurons did not support sequential activation underlying delay period information coding [63], and so far no evidence has been reported for delay period sequential activity in monkey experiments.

If working memory is indeed represented by a sequence of transiently active neurons, the stored information must be read out from different neural groups at different times. In that case, would downstream neurons need to constantly change their input weights over time for decoding? A simple solution is for readout neurons to receive converging inputs from all mnemonic cells. However, in that case, a downstream neuron would display stationary persistent activity [92], and so the computational benefit of such a scheme in comparison with a stationary code in the first place remains unclear.

Certain types of temporal variations of delay activity are suitable to serve specific functions. For instance, ramping activity could reflect anticipated timing of the memory-guided [94,95,96]. Corroborative evidence was also reported in a mouse delayed response task [63] in which ALM neurons showed ramping activity when the delay duration was fixed but tonic persistent activity in trials where the delay duration varied probabilistically and therefore was not predictable [61]. Other temporal changes require different explanations, some of which may be related to uncontrolled factors in an experiment, such as micro-behavior not required to perform the task [97].

Separated from temporal variations of a cell's firing, delay period activity also varies considerably from cell to cell (e.g. Figure 3a). Whereas early models strove for simplicity to optimize analysis and interpretations of network behavior, more recent elaborated attractor models display considerable cell-to-cell heterogeneities [98, 99, 78, 100, 101, 102]. In the brain, heterogeneity could arise from variations of biological properties across individual cells in a well-defined population, or/and because a recording is done among several subtypes of neurons [54, 103]. Heterogeneity across neurons may also be understood in terms of desirable functions such as mixed-selectivity essential for flexible cognitive behaviors [104, 105].

Activity-silent states

The key assumption of the attractor model is that a biological working memory circuit has distinct stimulus-selective mnemonic attractor states that coexist with a stable resting state. Alternatively, and inconsistent with the attractor model, a network may have only a single attractor (the resting state) and delay period activity may be genuinely transient: a to-

be-remembered stimulus perturbs the system to another internal state, from which it returns to the resting state after the input offset during a delay period. The return trajectory may be slow, but eventually, elevated activity should disappear if the delay period is sufficiently long.

The "activity-silent state" model posits that a memorandum can be encoded by "hidden" variables unobservable at the level of neuronal spiking [106], in which case there would be no need for persistent activity in the form of an attractor state. A plausible biological substrate for such activity silent working memory is synaptic short-term facilitation (STF), which in rodent cortex is more prominent between excitatory neurons in frontal cortex than primary visual cortex [107, 108]. Importantly substantial STF does not automatically imply an activity-silent state; instead, it could be required for the maintenance of persistent activity [107]. Moreover, persistent activity that depends on STF could be repetition of brief population bursts (Figure 3 of [109]), which should still be considered an attractor rather than activity-silent state. Thus, "hidden" synaptic variables and spiking are not decoupled, and STF can contribute to the maintenance of persistent activity as part of synaptic machinery [110]. Interestingly, STF and other slow synaptic or cellular processes could induce history dependence across trials [110], which has been observed in monkey and human studies [111, 112].

On the other hand, short-term synaptic plasticity (STF) could maintain a short-term memory trace even when self-sustained neural spiking dies out [109]. In other words, the activity-silent state model assumes that a dynamical variable of STF, not observable by spiking activity, could mediate short-term memory. Results from neurophysiological tests of this idea are not clear cut, partly because interpretations are not straightforward for different kinds of measurements, ranging from single-neuron physiology, EEG/MEG to fMRI BOLD signal. For instance, in a monkey experiment, LFP displays brief episodes of synchrony at γ frequency band (~ 40 Hz), which was interpreted as inconsistent with the sustained activity model [113, 114]. However, persistent activity of single cells often coexists with intermittent and weak LFP rhythms [115, 116, 117, 118]. Furthermore, brief bursts as the neural substrate of working memory representation predict that variability of spike trains would be much higher during the delay period than in the resting state. This prediction is contradicted by single-cell data from three monkey experiments [119]. A unifying explanation of all these data is the theory of sparsely synchronous oscillations, where episodic bursts of network coherence coexist with sustained firing of single cells [116, 117], and temporally enhance information conveyed by spikes [113, 118].

Nevertheless, the activity-silent scenario has a specific prediction. If a brief stimulus activates one of neural assemblies in a network therefore induce STF at their interconnections, a later non-selective global signal (a "pinging" of the entire network) would "reawaken" selectively that particular neural assembly because its hidden state is differentially primed by STF [109]. This prediction has been tested in human experiments. In one study, a subject was shown two sample stimuli (a face and a word), followed by a delay period when a post-cue instructed which of the two would be probed (e.g. word). Then a test (the same or a different word) was shown and the participant responded match or nonmatch. The trial continued with a second delay when another post-cue instructed

which of the two would be probed next (which might be face or word), a final test stimulus was shown, and the subject responded match or nonmatch [84]. Multivoxel patterns from the BOLD signal were used to decode each of the items in the initial sample set. It was found that category (face or word) decoding by BOLD signal decayed to baseline. However, each post-cue "reawakened" significant decoding of the corresponding stimulus category (Figure 2c), supporting the idea that information remained in some hidden state not detectable by BOLD signal, with the caveat that fMRI measurements are not directly related to spiking neural activity. Moreover, transcranial magnetic stimulation (TMS) reactivated representation of the latest cued category, consistent with the model prediction about pinging a short-term memory system [109]. Similar findings were reported in another experiment with two to-be-remembered items, using decoding from EEG and pinging with nonspecific visual stimulation [120].

The two experiments [84, 120] were designed on the idea that a stored item can be "in" (if cued) or "out of" (when uncued) the focus of internal attention [121, 122]. These observations suggest that an item at the center of attention is represented by persistent activity, whereas information about another item encoded in a hidden variable can be reactivated when it becomes a priority. However, these studies did not distinguish behaviorally relevant stimuli from distractors. This is critical because a requirement for normal working memory function is the brain's ability to filter out irrelevant distracting sensory flow [39, 41, 123, 124, 22, 125]. Modeling work showed that a synaptic memory trace is strongest for the latest shown stimulus because signals of earlier stimuli decay[126]. Therefore, it cannot realize working memory in the face of distractors that are presented after behaviorally relevant stimulation in the absence of some additional control mechanism.

One argument for the activity-silent state model is that spikes are costly [127], therefore realizing a memory trace without spike firing would save energy [106]. If so, in the monkey ODR and VDD experiments, PFC neuronal spike firing rates during the delay period should be greater than during the baseline state of fixation ("foreperiod") at the start of a trial. This is not true; surprisingly, the distribution of firing rates across the recorded PFC neurons is roughly log-normal and the same across behavioral epochs for both experiments (Figure 3c). This is also the case for delay period activities in the mouse experiment of [63] (Figure 3d). Presumably, in a given trial neurons selective for an encoded stimulus have elevated spiking activity while others reduce their firing, in such a way that the total population activity remains similar to the baseline state. Therefore, the attractor network model for persistent activity cannot be discounted, and the activity-silent state model is not favored, on the ground of metabolic energy consumption in the brain.

Persistent activity is required for manipulation of information in working memory

A functional perspective distinguishes short-term memory (STM), possibly involving the hippocampus [128], from working memory for which information is not only maintained but also manipulated without direct sensory stimulation [3, 129, 4, 130]. Even simple delay dependent tasks may require information manipulation, by transforming a sensory cue into a

prospective plan for the future [131, 132]. How can one test computationally the hypothesis that maintenance and manipulation of information during a delay period have different demands and differentially engage persistent activity? In recent years, tools from machine learning have been used to train recurrent neural networks (RNNs) to perform tasks [133]. An RNN is initially a "blank slate", where connection weights are not specific and the network is incapable of any function. If a to-be-learnt task involves a mnemonic delay, this approach does not make an *à prior* assumption as to whether an RNN will solve the problem by virtue of a persistent activity pattern or an activity-silent state. Therefore, it offers an opportunity to investigate which of the two scenarios emerges from training [101].

In the model depicted in Figure 4a, an input layer signals spatial location and direction of motion stimuli, and an output layer generates a delayed response. The recurrent network between the input and output layers is wired with connections endowed with STF. Some are dominated by short-term depression (Figure 4b, left) while others by short-term facilitation (Figure 4b, right). The synaptic efficacy is the product of the depression factor and facilitation factor. In a motion-direction delayed DMS task, the sample is decoded either by recurrent neural population activity or by activity-silent synaptic efficacy. When the delay period is short, STF can keep a memory trace of the sample, in which case activity is not necessary. Indeed, a trained RNN found the solution with chance-level performance of decoding by activity of recurrent units, but decoding by activity increases with gradually prolonged delay duration (Figure 4c). This is because when the delay period is long compared to the biological time constants of STF, the network can no longer find a solution by an activity-silent state scenario, and persistent activity sustained by an attractor state emerges from training through experience.

What happens if an RNN is trained to perform a working memory task where information must be manipulated during the delay period? In a delayed match-to-rotated-sample (DMRS) task, subjects must decide whether the test direction is the same as the sample direction rotated by 90 degrees. In this case, even with a short delay, persistent activity naturally emerged from training, demonstrating that the amount of persistent activity (hence the accuracy of its sample decoding) depends on the behavioral demand for information manipulation during the delay period. This conclusion was further confirmed by training different RNNs to perform one of nine tasks for which the degree of required information manipulation was quantified. Generally, decoding accuracy from recurrent population activity increases with the task demand of information manipulation (Figure 4d). These findings highlight the importance of distinguishing passive short-term memory traces from active working memory: short-term memory traces do not always require persistent activity. On the other hand, internal computation is carried out and communicated by spikes; because information manipulation is an integral part of working memory at the cognitive level, persistent activity is essential for working memory.

Concluding remarks

I have reviewed experimental and theoretical research on selective self-sustained persistent activity as a neural substrate for working memory representation. Substantial progress has been made in our understanding of the neural circuit mechanisms of persistent

activity, through close interactions between experimentation using delay dependent tasks and biologically-based computational models. An important concept running through this research is attractors, stable states of a dynamical system that may be steady states (corresponding to tonic persistent activity) or complex spatiotemporal patterns. The workhorse for working memory maintenance is positive feedback, which depends on the recurrent synaptic excitation, but single neuronal and synaptic dynamical properties also play a role [31, 134]. Feedbacks include both local and long-distance connections such as the phonological loop in the case of human speech [135]. The attractor network model makes several testable predictions (Box 2). It is a synaptic theory because it mainly relies on network reverberations; short-term synaptic plasticity, which depends on neural firing and in turn can enhance spiking activity, represents one contributing factor and naturally fits into the attractor network model [107, 110, 112]. Alternatively, if a memory trace is encoded solely by a hidden state such as synaptic efficacy endowed with short-term plasticity, physiological experiments should be able to detect the trace [136, 112]. The energy-saving argument in favor of the activity-silent state scenario [106, 114] is inconsistent with the conserved totality of neural population spiking activity across different behavioral epochs. A hidden-variable mechanism is likely to be sufficient for passive short-term memory but not active working memory, because it works only when the delay period is short compared to the time constant of the underlying biological process, it does not filter out distractors, and it is not suited to subserve information manipulation internally in the brain [101].

In summary, the persistent firing mechanism has withstood challenges as the neural substrate of working memory coding. At the same time, recent work also highlights the need to better understand the complex spatiotemporal mnemonic processes in a working memory circuit and the benefit of distinguishing working memory from passive short-term memory. Efforts devoted to understanding the neural circuit mechanism of persistent activity have played a major role in revealing the mystery of the prefrontal cortex [59]. Among the most important challenges for future research (see Outstanding Questions) is the need to elucidate how the PFC works with the rest of brain in distributed working memory and related cognitive processes to advance the nascent neuroscience of large-scale brain systems [137, 138, 139, 126, 140].

Acknowledgements:

I thank Bijan Pesaran and Albert Compte for a critical reading of the manuscript. This work was supported by the NIH grant R01MH062349, ONR grant N00014-17-1-2041, NeuroNex grant NSF 2015276, and James Simons foundation grant 543057SPI.

References

[1]. Jacobsen CF Studies of cerebral function in primates: I. the functions of the frontal association areas in monkeys. Comp. Psychol. Monogr 13, 1–68 (1936).

[2]. Pribram KH, Mishkin M, Rosvold HE & Kaplan SJ Effects on delayed-response performance of lesions of dorsolateral and ventromedial frontal cortex of baboons. Journal of Comparative and Physiological Psychology 45, 565 (1952). [PubMed: 13000029]

[3]. Baddeley A & Hitch GJ Working memory. In Bower GA (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, 47–89 (New York: Academic Press, 1974).

[4]. Baddeley A Working memory: theories, models, and controversies. Annu Rev Psychol 63, 1–29 (2012). [PubMed: 21961947]

[5]. Fuster JM & Alexander GE Delayed response deficit by cryogenic depression of frontal cortex. Brain Research 20, 85–90 (1970). [PubMed: 4986430]

[6]. Fuster JM & Alexander G Neuron activity related to short-term memory. Science 173, 652–654 (1971). [PubMed: 4998337]

[7]. Kubota K & Niki H Prefrontal cortical unit activity and delayed alternation performance in monkeys. J. Neurophysiol 34, 337–347 (1971). [PubMed: 4997822]

[8]. Miyashita Y Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335, 817–820 (1988). [PubMed: 3185711]

[9]. Miyashita Y & Chang HS Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature 331, 68–70 (1988). [PubMed: 3340148]

[10]. 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]

[11]. Freedman DJ, Riesenhuber M, Poggio T & Miller EK Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291, 312–316 (2001). [PubMed: 11209083]

[12]. Wallis J, Anderson K & Miller E Single neurons in prefrontal cortex encode abstract rules. Nature 411, 953–956 (2001). [PubMed: 11418860]

[13]. Freedman DJ & Assad JA Experience-dependent representation of visual categories in parietal cortex. Nature 443, 85–88 (2006). [PubMed: 16936716]

[14]. 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]

[15]. Gnadt JW & Andersen RA Memory related motor planning activity in posterior parietal cortex of macaque. Exp. Brain Res 70, 216–220 (1988). [PubMed: 3402565]

[16]. 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]

[17]. Rao SC, Rainer G & Miller EK Integration of what and where in the primate prefrontal cortex. Science 276, 821–824 (1997). [PubMed: 9115211]

[18]. Pesaran B, Pezaris JS, Sahani M, Mitra PP & Andersen RA Temporal structure in neuronal activity during working memory in marcaque pariental cortex. Nat. Neurosci 5, 805–811 (2002). [PubMed: 12134152]

[19]. Pasternak T & Greenlee M Working memory in primate sensory systems. Nat. Rev. Neurosci 6, 97–107 (2005). [PubMed: 15654324]

[20]. Vijayraghavan S, Wang M, Birnbaum SG, Williams GV & Arnsten AF Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. Nat. Neurosci 10, 376–384 (2007). [PubMed: 17277774]

[21]. Wang M et al. Neuronal basis of age-related working memory decline. Nature 476, 210–213 (2011). [PubMed: 21796118]

[22]. Suzuki M & Gottlieb J Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. Nat. Neurosci 16, 98–104 (2013). [PubMed: 23242309]

[23]. Romo R, Brody CD, Hernández A & Lemus L Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399, 470–474 (1999). [PubMed: 10365959]

[24]. Bastos AM, Loonis R, Kornblith S, Lundqvist M & Miller EK Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. Proceedings of the National Academy of Sciences 115, 1117–1122 (2018).

[25]. Courtney SM, Petit L, Maisog JM, Ungerleider LG & Haxby JV An area specialized for spatial working memory in human frontal cortex. Science 279, 1347–1351 (1998). [PubMed: 9478894]

[26]. Finn ES, Huber L, Jangraw DC, Molfese PJ, and Bandettini PA. Layer-dependent activity in human prefrontal cortex during working memory. Nature Neuroscience 22: 1687–1695 (2019). [PubMed: 31551596]

[27]. Miller EK & Cohen JD An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24, 167–202 (2001). [PubMed: 11283309]

[28]. Amari S Dynamics of pattern formation in lateral-inhibition type neural fields. Biol. Cybern 27, 77–87 (1977). [PubMed: 911931]

[29]. Hopfield JJ Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. (USA) 79, 2554–2558 (1982). [PubMed: 6953413]

[30]. Amit DJ The Hebbian paradigm reintegrated: local reverberations as internal representations. Behav. Brain Sci 18, 617–626 (1995).

[31]. Wang X-J Synaptic reverberation underlying mnemonic persistent activity. Trends in Neurosci. 24, 455–463 (2001).

[32]. Strogatz SH Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering (Oxford, Britain: Taylor & Francis Group, 2016), second edition edn.

[33]. Goldman-Rakic PS Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In Plum F & Mountcastle V (eds.) Handbook of Physiology – The nervous system V, chap. 9, 373–417 (Bethesda, Maryland: American Physiological Society, 1987).

[34]. Goldman-Rakic PS Working memory and the mind. Sci. Am 267, 110–117 (1992).

[35]. Goldman-Rakic PS Cellular basis of working memory. Neuron 14, 477–485 (1995). [PubMed: 7695894]

[36]. Arnsten AF, Paspalas CD, Gamo NJ, Yang Y & Wang M Dynamic network connectivity: a new form of neuroplasticity. Trends Cogn. Sci 14, 365–375 (2010). [PubMed: 20554470]

[37]. Amit DJ & Brunel N Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cerebral Cortex 7, 237–252 (1997). [PubMed: 9143444]

[38]. Wang X-J Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. J. Neurosci 19, 9587–9603 (1999). [PubMed: 10531461]

[39]. Compte A, Brunel N, Goldman-Rakic PS & Wang X-J Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex 10, 910–923 (2000). [PubMed: 10982751]

[40]. 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]

[41]. Brunel N & Wang X-J Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11, 63–85 (2001). [PubMed: 11524578]

[42]. Tegnér J, Compte A & Wang X-J The dynamical stability of reverberatory neural circuits. Biol. Cybern 87, 471–481 (2002). [PubMed: 12461636]

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

[44]. van Vugt B, van Kerkoerle T, Vartak D & Roelfsema PR The contribution of ampa and nmda receptors to persistent firing in the dorsolateral prefrontal cortex in working memory. Journal of Neuroscience 40, 2458–2470 (2020). [PubMed: 32051326]

[45]. Yang S, Seo H, Wang M & Arnsten AF NMDAR neurotransmission needed for persistent neuronal firing: Potential roles in mental disorders. Frontiers in Psychiatry 12, 337 (2021).

[46]. Wang X-J, Tegnér J, Constantinidis C & Goldman-Rakic PS Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. Proc Natl Acad Sci U S A 101, 1368–1373 (2004). [PubMed: 14742867]

[47]. Kepecs A & Fishell G Interneuron cell types are fit to function. Nature 505, 318–326 (2014). [PubMed: 24429630]

[48]. Tremblay R, Lee S & Rudy B GABAergic interneurons in the neocortex: From cellular properties to circuits. Neuron 91, 260–292 (2016). [PubMed: 27477017]

[49]. 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]

[50]. Coyle JT, Tsai G & Goff D Converging evidence of NMDA receptor hypofunction in the pathophysiology of schizophrenia. Ann N Y Acad Sci 1003, 318–327 (2003). [PubMed: 14684455]

[51]. Wang X-J Toward a prefrontal microcircuit model for cognitive deficits in schizophrenia. Pharmacopsychiatry 39 Suppl 1, 80–87 (2006). [PubMed: 16555171]

[52]. Stein H et al. Reduced serial dependence suggests deficits in synaptic potentiation in anti-nmdar encephalitis and schizophrenia. Nature Communications 11, 10.1038/s41467-020-18033-3 (2020).

[53]. Montague PR, Dolan RJ, Friston KJ & Dayan P Computational psychiatry. Trends Cogn. Sci 16, 72–80 (2012). [PubMed: 22177032]

[54]. Wang X-J & Krystal JH Computational psychiatry. Neuron 84, 638–654 (2014). [PubMed: 25442941]

[55]. Wang X-J Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36, 955–968 (2002). [PubMed: 12467598]

[56]. Roitman JD & Shadlen MN Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci 22, 9475–9489 (2002). [PubMed: 12417672]

[57]. Gold JI & Shadlen MN The neural basis of decision making. Annu. Rev. Neurosci 30, 535–574 (2007). [PubMed: 17600525]

[58]. Wang X-J Decision making in recurrent neuronal circuits. Neuron 60, 215–234 (2008). [PubMed: 18957215]

[59]. Wang X-J The prefrontal cortex as a quintessential 'cognitive-type' neural circuit: Working memory and decision making. In Stuss DT & Knight RT (eds.) Principles of Frontal Lobe Function, 226–248 (New York: Cambridge University Press, 2013), second edn.

[60]. Guo ZV et al. Flow of cortical activity underlying a tactile decision in mice. Neuron 81, 179–194 (2014). [PubMed: 24361077]

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

[62]. 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]

[63]. Li N, Daie K, Svoboda K & Druckmann S Robust neuronal dynamics in premotor cortex during motor planning. Nature 532, 459–464 (2016). [PubMed: 27074502]

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

[65]. Kopec CD, Erlich JC, Brunton BW, Deisseroth K & Brody CD Cortical and subcortical contributions to short-term memory for orienting movements. Neuron 88, 367–377 (2015). [PubMed: 26439529]

[66]. 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]

[67]. Piet AT, Erlich JC, Kopec CD & Brody CD Rat prefrontal cortex inactivations during decision making are explained by bistable attractor dynamics. Neural Comput 29, 2861–2886 (2017). [PubMed: 28777728]

[68]. Finkelstein A et al. Attractor dynamics gate cortical information flow during decision-making. Nature Neuroscience 10.1038/s41593-021-00840-6 (2021).

[69]. Seung HS How the brain keeps the eyes still. Proc. Natl. Acad. Sci. (USA) 93, 13339–13344 (1996). [PubMed: 8917592]

[70]. Lim S, & Goldman MS Balanced cortical microcircuitry for maintaining information in working memory. Nature Neuroscience 16, 1306–1314 (2013) [PubMed: 23955560]

[71]. Wei Z, Wang XJ & Wang DH From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. J. Neurosci 32, 11228–11240 (2012). [PubMed: 22895707]

[72]. Batuev A, Pirogov A & Orlov A Unit activity of the prefrontal cortex during delayed alternation performance in monkey. Acta physiologica Academiae Scientiarum Hungaricae 53, 345–353 (1979). [PubMed: 120674]

[73]. Baeg EH et al. Dynamics of population code for working memory in the prefrontal cortex. Neuron 40, 177–188 (2003). [PubMed: 14527442]

[74]. Fuster JM The Prefrontal Cortex (Academic Press: New York, 2008), Fourth edn.

[75]. Constantinidis C et al. Persistent spiking activity underlies working memory. Journal of Neuroscience 38, 7020–7028 (2018). [PubMed: 30089641]

[76]. Renart A, Brunel N & Wang X-J Mean-field theory of recurrent cortical networks: Working memory circuits with irregularly spiking neurons. In Feng J (ed.) Computational Neuroscience: A Comprehensive Approach, 432–490 (Boca Raton: CRC Press, 2003).

[77]. Barbieri F & Brunel N Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? Frontiers in Neurosci. 2, 114–122 (2008).

[78]. Barak O, Sussillo D, Romo R, Tsodyks M & Abbott LF From fixed points to chaos: three models of delayed discrimination. Prog. Neurobiol 103, 214–222 (2013). [PubMed: 23438479]

[79]. 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]

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

[81]. 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]

[82]. Stokes MG et al. Dynamic coding for cognitive control in prefrontal cortex. Neuron 78, 364–375 (2013). [PubMed: 23562541]

[83]. Harvey CD, Coen P & Tank DW Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484, 62–68 (2012). [PubMed: 22419153]

[84]. Rose NS et al. Reactivation of latent working memories with transcranial magnetic stimulation. Science 354, 1136–1139 (2016). [PubMed: 27934762]

[85]. Meyers EM Dynamic population coding and its relationship to working memory. J. Neurophysiol 120, 2260–2268 (2018). [PubMed: 30207866]

[86]. Kaminski J & Rutishauser U Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. Annals of New York Academy of Sciences 1459, doi: 10.1111/nyas.14213 (2019).

[87]. 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]

[88]. 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]

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

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

[91]. Zhu J et al. Transient delay-period activity of agranular insular cortex controls working memory maintenance in learning novel tasks. Neuron 105, 934–946 (2020). [PubMed: 32135091]

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

[93]. Rajan K, Harvey CD & Tank DW Recurrent network models of sequence generation and memory. Neuron 90, 128–142 (2016). [PubMed: 26971945]

[94]. Machens CK, Romo R & Brody CD Functional, but not anatomical, separation of ”what” and ”when” in prefrontal cortex. J. Neurosci 30, 350–360 (2010). [PubMed: 20053916]

[95]. Markowitz DA, Curtis CE & Pesaran B Multiple component networks support working memory in prefrontal cortex. Proceedings of the National Academy of Sciences 112, 11084–11089 (2015).

[96]. Brody C, 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]

[97]. Musall S, Kaufman MT, Juavinett AL, Gluf S & Churchland AK Single-trial neural dynamics are dominated by richly varied movements. Nat. Neurosci 22, 1677–1686 (2019). [PubMed: 31551604]

[98]. Renart A, Song P & Wang X-J Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron 38, 473–485 (2003). [PubMed: 12741993]

[99]. Hansel D & Mato G Short-term plasticity explains irregular persistent activity in working memory tasks. J. Neurosci 33, 133–149 (2013). [PubMed: 23283328]

[100]. Chaisangmongkon W, Swaminathan SK, Freedman DJ & Wang XJ Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions. Neuron 93, 1504–1517 (2017). [PubMed: 28334612]

[101]. Masse NY, Yang GR, Song HF, Wang X-J & Freedman DJ Circuit mechanisms for the maintenance and manipulation of information in working memory. Nat. Neurosci 22, 1159–1167 (2019). [PubMed: 31182866]

[102]. Yang GR, Joglekar MR, Song HF, Newsome WT & Wang X-J Task representations in neural networks trained to perform many cognitive tasks. Nat. Neurosci 22, 297–306 (2019). [PubMed: 30643294]

[103]. Hirokawa J, Vaughan A, Masset P, Ott T & Kepecs A Frontal cortex neuron types categorically encode single decision variables. Nature 576, 446–451 (2019). [PubMed: 31801999]

[104]. Rigotti M, Rubin DB, Wang X-J & Fusi S Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. Front Comput Neurosci 4, 24 (2010). [PubMed: 21048899]

[105]. Rigotti M et al. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590 (2013). [PubMed: 23685452]

[106]. Stokes MG ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. (Regul. Ed.) 19, 394–405 (2015).

[107]. Hempel CM, Hartman KH, Wang X-J, Turrigiano G & Nelson SB Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. J. Neurophysiol 83, 3031–3041 (2000). [PubMed: 10805698]

[108]. Wang Y et al. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat Neurosci 9, 534–542 (2006). [PubMed: 16547512]

[109]. Mongillo G, Barak O & Tsodyks M Synaptic theory of working memory. Science 319, 1543–1546 (2008). [PubMed: 18339943]

[110]. Pereira J & Wang XJ A tradeoff between accuracy and flexibility in a working memory circuit endowed with slow feedback mechanisms. Cereb. Cortex 25, 3586–3601 (2015). [PubMed: 25253801]

[111]. Bliss DP, Sun JJ & D’Esposito M Serial dependence is absent at the time of perception but increases in visual working memory. Sci Rep 7, 14739 (2017). [PubMed: 29116132]

[112]. Barbosa J et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature Neuroscience 23, 1016–1024 (2020). [PubMed: 32572236]

[113]. Lundqvist M et al. Gamma and Beta Bursts Underlie Working Memory. Neuron 90, 152–164 (2016). [PubMed: 26996084]

[114]. Lundqvist M, Herman P & Miller EK Working memory: delay activity, yes! persistent activity? maybe not. Journal of Neuroscience 38, 7013–7019 (2018). [PubMed: 30089640]

[115]. Brunel N & Hakim V Fast global oscillations in networks of integrate-andfire neurons with low firing rates. Neural Computation 11, 1621–1671 (1999). [PubMed: 10490941]

[116]. Brunel N & Wang X-J What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitationinhibition balance. J Neurophysiol 90, 415–430 (2003). [PubMed: 12611969]

[117]. Wang X-J Neurophysiological and computational principles of cortical rhythms in cognition. Physiol. Rev 90, 1195–1268 (2010). [PubMed: 20664082]

[118]. Palmigiano A, Geisel T, Wolf F & Battaglia D Flexible information routing by transient synchrony. Nat. Neurosci 20, 1014–1022 (2017). [PubMed: 28530664]

[119]. Li D, Constantinidis C & Murray JD Trial-to-trial variability of spiking delay activity in prefrontal cortex constrains burst-coding models of working memory. bioRxiv doi: 10.1101/2021.01.30.428962 (2021).

[120]. Wolff MJ, Jochim J, Akyürek EG & Stokes MG Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci 20, 864–871 (2017). [PubMed: 28414333]

[121]. Myers NE, Stokes MG & Nobre AC Prioritizing information during working memory: beyond sustained internal attention. Trends Cogn. Sci. (Regul. Ed.) 21, 449–461 (2017).

[122]. Christophel TB, Iamshchinina P, Yan C, Allefeld C & Haynes JD Cortical specialization for attended versus unattended working memory. Nature neuroscience 21, 494–496 (2018). [PubMed: 29507410]

[123]. Sakai K, Rowe JB & Passingham RE Active maintenance in prefrontal area 46 creates distractor-resistant memory. Nat Neurosci 5, 479–484 (2002). [PubMed: 11953754]

[124]. Gazzaley A & Nobre AC Top-down modulation: bridging selective attention and working memory. Trends in Cognitive Sciences 16, 129–135 (2012). [PubMed: 22209601]

[125]. Buschman TJ Balancing flexibility and interference in working memory. Annual Review of Vision Science 7, VS07CH0 (2021).

[126]. Froudist-Walsh S et al. A dopamine gradient controls access to distributed working memory in monkey cortex. BioRxiv doi:10.1101/2020.09.07.286500 (2020).

[127]. Attwell D & Laughlin SB An energy budget for signaling in the grey matter of the brain. Journal of Cerebral Blood Flow & Metabolism 21, 1133–1145 (2001). [PubMed: 11598490]

[128]. Beukers AO, Buschman TJ, Cohen JD & Norman KA Is activity silent working memory simply episodic memory? Trends in Cognitive Sciences 25, 284–293 (2021). [PubMed: 33551266]

[129]. Cowan N What are the differences between long-term, short-term, and working memory? Progress in brain research 169, 323–338 (2008). [PubMed: 18394484]

[130]. Trübutschek D, Marti S, Ueberschär H & Dehaene S Probing the limits of activity-silent non-conscious working memory. Proceedings of the National Academy of Sciences 116, 14358–14367 (2019).

[131]. Wu Z et al. Context-dependent decision making in a premotor circuit. Neuron 106, 316–328 (2020). [PubMed: 32105611]

[132]. Ehrlich DB & Murray JD Geometry of neural computation unifies working memory and planning. bioRxiv doi: 10.1101/2021.02.01.429156 (2021).

[133]. Yang GR & Wang X-J Artificial neural networks for neuroscientists: a primer. Neuron 107, 1048–1070 (2020). [PubMed: 32970997]

[134]. 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]

[135]. Cogan GB et al. Sensory–motor transformations for speech occur bilaterally. Nature 507, 94–98 (2014). [PubMed: 24429520]

[136]. Fujisawa S, Amarasingham A, Harrison MT & Buzsáki G Behavior dependent short-term assembly dynamics in the medial prefrontal cortex. Nature Neurosci. 11, 823 (2008). [PubMed: 18516033]

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

[138]. Christophel TB, Klink PC, Spitzer B, Roelfsema PR & Haynes JD The distributed nature of working memory. Trends Cogn. Sci 21, 111–124 (2017). [PubMed: 28063661]

[139]. Mejias JF & Wang X-J Mechanisms of distributed working memory in a large-scale model of the macaque neocortex. BioRxiv 760231 (2020).

[140]. Wang X-J Macroscopic gradients of synaptic excitation and inhibition in the neocortex. Nature Reviews Neuroscience 21, 169–178 (2020). [PubMed: 32029928]

Highlights

Working memory actively engages stimulus-selective persistent activity mathematically described as an attractor state of a reverberatory neural circuit.

The attractor network model is compatible with temporal variations of mnemonic neural firing in a subspace of population activity.

Sustained activity during working memory coexists with intermittent bursts of frequency-dependent network synchronization.

There is no increase in the total number of spikes in a neural population during a mnemonic delay period compared to a baseline state. Thus, persistent activity is not more energetically costly than an alternative memory mechanism using hidden variables.

Activity-silent state mechanisms like synaptic short-term facilitation are suitable for the storage of passive memory traces, but not working memory, which also involves manipulation of information online in the absence of external stimulation.

Box 1.

Fuster's reminiscence

In the late 1960's we found in my laboratory that by cryogenic inactivation of the lateral prefrontal cortex we could produce a reversible deficit in monkeys' performance of a delayed response task, a test of working memory. Thus we reestablished by reversible lesion what Jacobsen had established many years before by ablation. The beauty of our method was that it allowed us to use each animal repeatedly as its own control. From the results of that experiment it became clear to me that the lateral prefrontal cortex was critical for the temporary retention of a form of short-term memory that later Baddeley called working memory. It was therefore reasonable to expect that the nerve cells in that part of the cortex would be actively involved in that form of memory. Because at the same time we were becoming proficient at recording with microelectrodes single units from chronic animals, it occurred to me that those cells had to undergo recordable activity changes during delayed response, that is, during memory retention. With the help of my graduate student Gary Alexander, we trained monkeys to perform the delayed-response task and surgically prepared them for single-cell recording from the prefrontal cortex. My expectation was happily fulfilled: a substantial number of prefrontal units showed persistent elevations of firing rate during the delay, the memory retention period of the task. Never in my scientific life have I experienced a cleaner confirmation of a hypothesis (many have failed!), though later it turned out that the sustained delay activity reflects the influence of other factors in addition to memory.

Box 2.

Predictions of an attractor state in contrast to a decaying transient

An attractor as the substrate of an internal brain state is robust against brief and modest perturbations, which can be noise, sensory distractors, or intruding thoughts. This can be tested experimentally using optogenetic perturbations.

A working memory representation sustained by an attractor is insensitive to the duration of a mnemonic time period, which can be varied systematically in an experiment. Forgetting is not due to passive decay but interference by other mental processes.

Neurobiologically, an activity-silent memory trace can be instantiated by a purely feedforward process. By contrast, the attractor model predicts that memory relies on sufficiently strong reverberation through feedback loops on multiple levels in a subnetwork of the brain.

The coexistence of multiple attractors enables a working memory circuit to rapidly switch between a resting state and an information-specific mnemonic state, in contrast to slow transients that cannot be turned off by a brief input.

The attractor network model but not the activity-silent state model is capable of filtering out behaviorally irrelevant distractors in working memory, this can be verified experimentally using distracting stimuli shown after a behaviorally relevant one is stored in working memory.

The landscape of multiple attractors can be modified flexibly by executive control signals, which vary depending on cognitive load.

Outstanding questions

Under what behavioral circumstances is memory instantiated by sequential activation of different neural groups, each firing briefly? What is the mechanism for a downstream system to readout the stored information at different time points?

What is the precise dynamical nature of persistent activity? How can one distinguish an attractor of highly complex spatiotemporal neural activity from slowly decaying transients?

What biologically-realistic neural circuit model accounts for the preserved totality of neural population activity during rest and active working memory?

During the mnemonic period of a working memory task, is the internal representation retrospective about previously shown stimuli or prospective about upcoming events and actions? How does the transformation from retrospective to retrospective coding take place in a neural circuit?

What is the biological mechanism of history dependence of working memory behavior across trials? What would be its functional utility?

How can the limited working memory capacity be explained mechanistically? How is the content of working memory controlled and flexibly updated according to behavioral demands?

What is the large-scale brain circuit basis of distributed working memory? What would constitute an adequate mathematical model of such distributed representation?

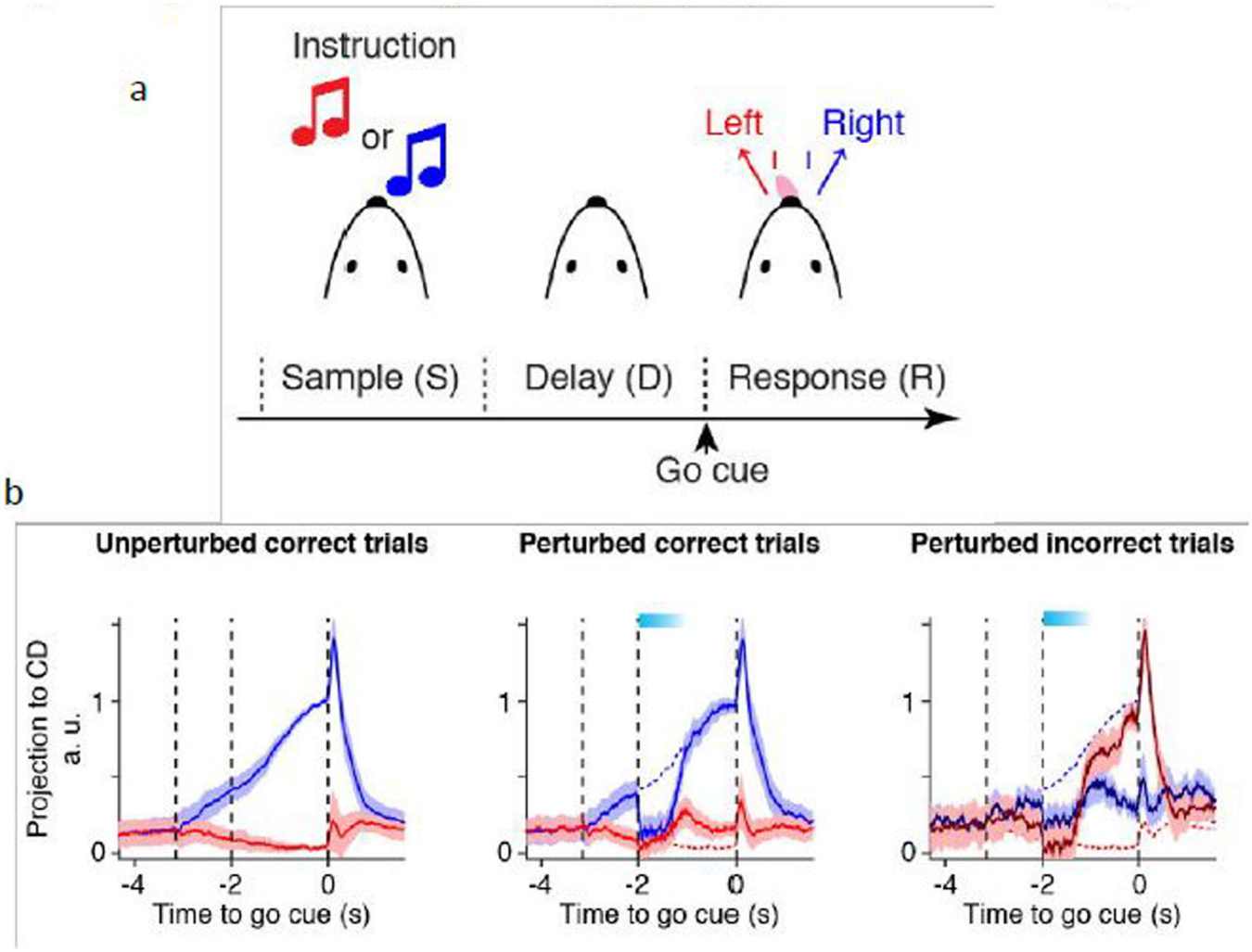

Figure 1: Mnemonic activity in mouse performing a delayed response task.

(a) one of two sensory cues is presented briefly, which can be a high or low tone in an auditory task or near or far location of an object on the whisker in a somatosensory task. The two stimuli are mapped to left and right licking responses, shown in red and blue, respectively. A correct motor response after a delay yields a reward. (b) Population activity from ~ 10 simultaneously recorded ALM neurons, projected in the one-dimensional subspace optimized for mnemonic representation. Left: control when the choice is correct (right, blue). Middle: optogenetic inactivation at the start of the delay period suppresses right-selective neural activity, which recovers and the ultimate choice is correct. Right: same as middle but this time optogenetic manipulation induces an incorrect response (red, left). The ALM decoding still predicts the erroneous movement direction, demonstrating its correlate with behavior performance. Reproduced from [61].

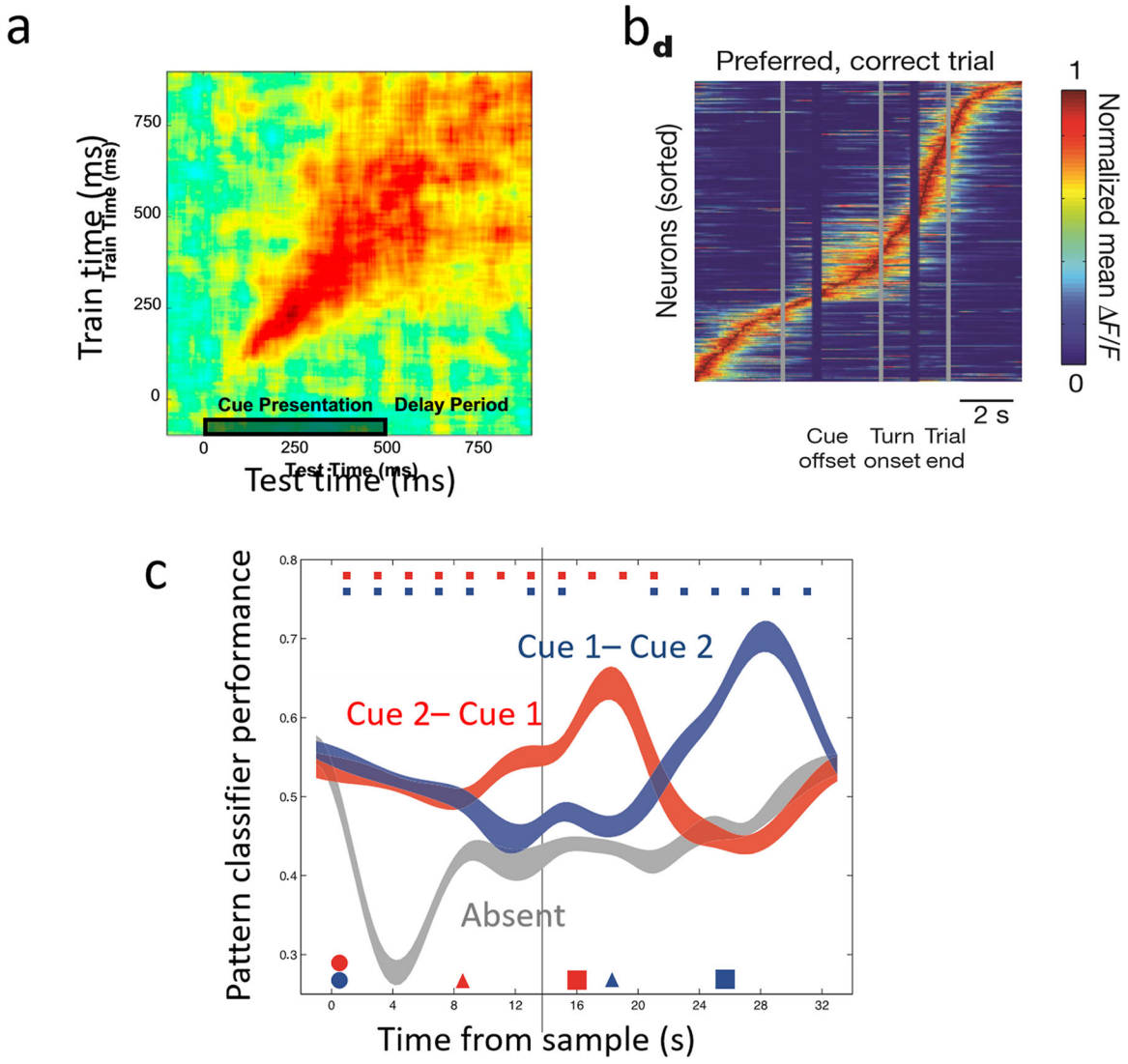

Figure 2:

Analysis of information coding by delay period activity, (a) Cross-time classification matrix of recorded neurons for a remembered item. Classifiers are trained to discriminate trial type at time *t* (y-axis) and tested at time *t* (x-axis). (b) In a delayed response task, calcium imaging of choice-specific cells (one cell per row) in the posterior parietal cortex of a behaving mouse. Traces were normalized to each cell's maximal activity on preferred trials and sorted by the peak time. (c) Decoding from human fMRI BOLD signals in a multi-step task in which two items were presented as memoranda for each trial. A cue indicated which item would be tested by the impending recognition memory probe, followed by the probe, then by a second cue, and then a second probe. Red and blue dots: stimulus presentation; red triangle: first cue; blue triangle: second cue. After the first cue, decoding by a classifier of the first cued item (red) increases whereas that of the uncued item (blue) decays to the baseline (grey). Upon the presentation of the second cue, decoded evidence for the two categories reversed for the remainder of the trial. Panel (a) is reproduced from [82], (b) from [83], (c) from [84].

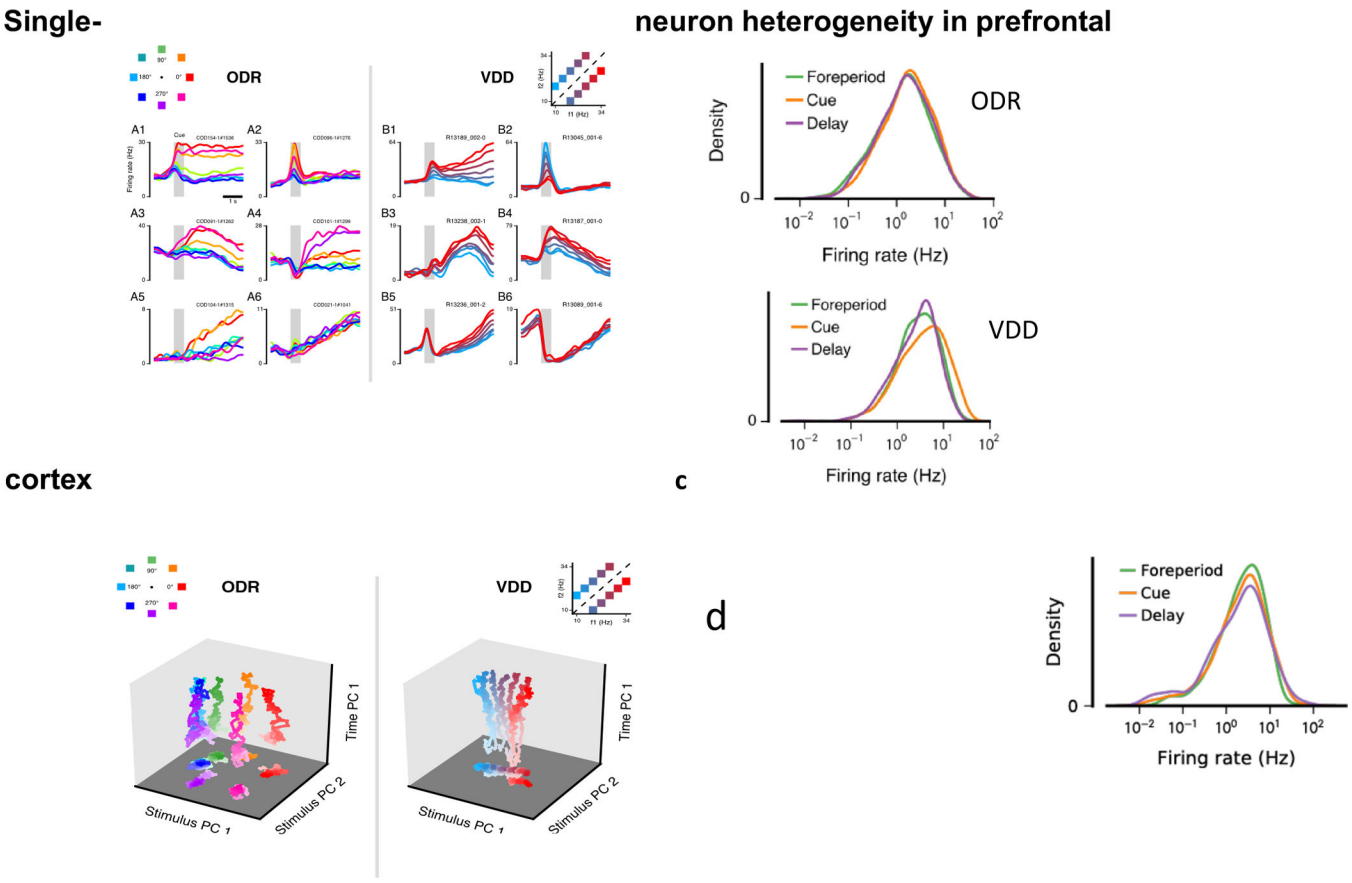

Figure 3:

Coexistence of stable working memory coding and temporal dynamics of delay period neural population activity. (a) Six individual neurons are shown for each of two monkey experiments using ODR (left) and VDD (right) tasks. Different colors correspond to different stimuli. (b) PFC analysis of population activity reveals that coding is stable (traces for different colors are distinct) in a subspace of the population activity state space (PC1 and PC2), whereas temporal changes are confined in the orthogonal subspace (PC 3). (c) Firing rate distributions of PFC neurons in behaving monkeys, plotted with logarithmic scale along the x-axis and linear scale along the y-axis for the ODR and VDD experiments, respectively. (d) Firing rate distributions of ALM neurons from mice performing a delay dependent task. Panels (a-c) are reproduced from [88], (d) from data provided by Nuo Li [63].

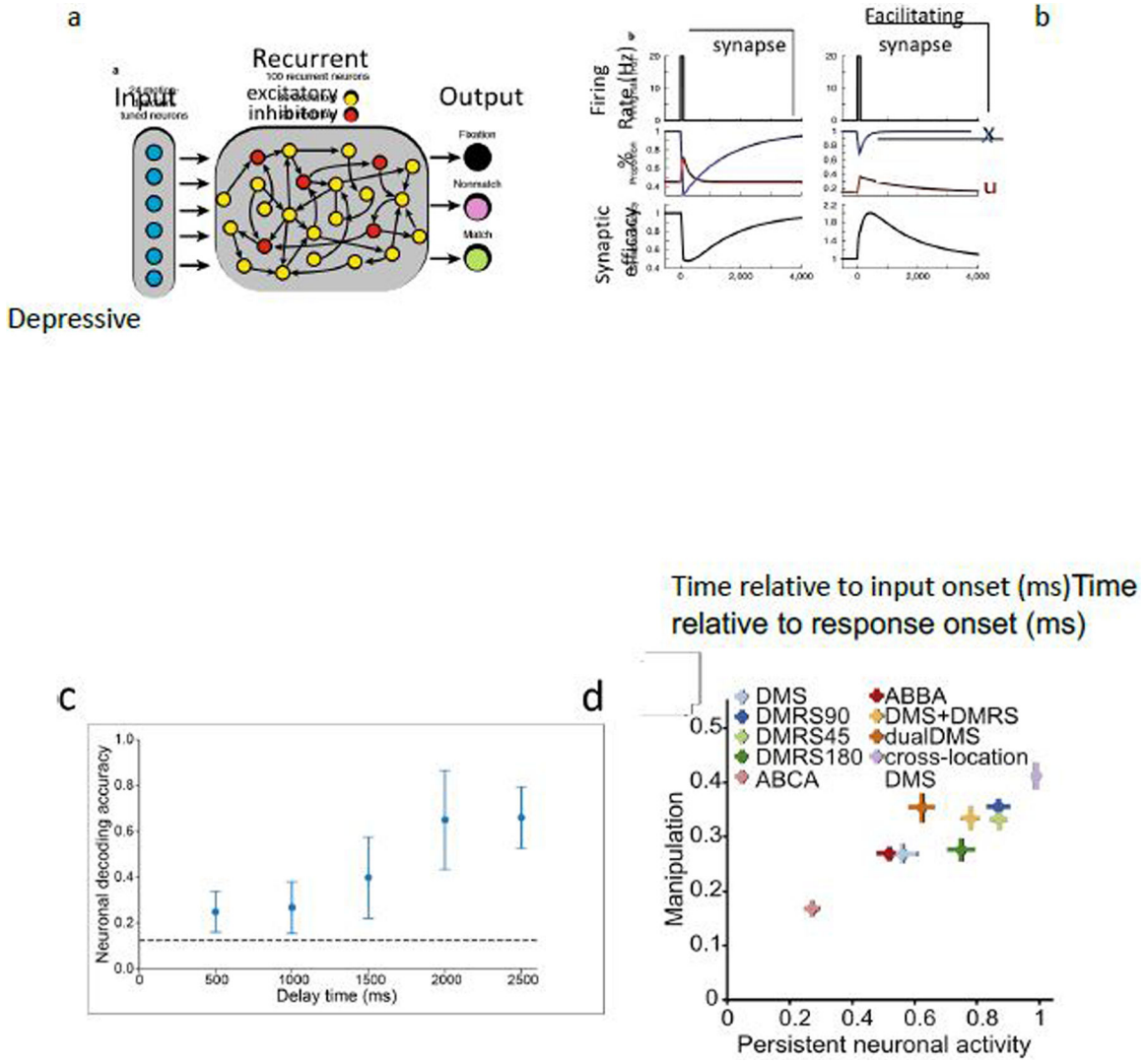

Figure 4:

A recurrent neural network trained by machine learning to perform working memory tasks. (a) Model scheme. (b) Short-term facilitation and short-term depression in response to a pulse input. Variable u (red): facilitation factor; x (blue): depression factor, both defined between 0 and 1. Synaptic efficacy is proportional to the product ux. (c) After the model is trained to perform a delayed match-to-sample (DMS) task, decoding accuracy from the recurrent population activity is poor with short delay duration, but gradually increases when delay becomes longer than the biological time constants of STF. (d) Scatterplot shows the level of persistent neuronal activity, measured as the neuronal decoding accuracy during the last 100 ms of the delay (x-axis), versus the level of manipulation (y-axis) across nine different tasks (indicated by colored crosses). Adapted from [101].