CellPress

Trends in Neurosciences

Review

50 years of mnemonic persistent activity: quo vadis?

Xiao-Jing Wang [1,*]

Half a century ago persistent spiking activity in the neocortex was discovered to be a neural substrate of working memory. Since then scientists have sought to understand this core cognitive function across biological and computational levels. Studies are reviewed here that cumulatively lend support to a synaptic theory of recurrent circuits for mnemonic persistent activity that depends on various cellular and network substrates and is mathematically described by a multiple-attractor network model. Crucially, a mnemonic attractor state of the brain is consistent with temporal variations and heterogeneity across neurons in a subspace of population activity. Persistent activity should be broadly understood as a contrast to decaying transients. Mechanisms in the absence of neural firing ('activity-silent state') are suitable for passive short-term memory but not for working memory – which is characterized by executive control for filtering out distractors, limited capacity, and internal manipulation of information.

Highlights

Working memory actively engages stimulus-selective persistent activity, which is 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, suggesting that persistent activity is not more energetically costly than an alternative memory mechanism using hidden variables.

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

Mnemonic persistent activity as an atom of cognition

The year 2021 marks the 50th 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 that was 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 the performance of macaque monkeys 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 memory – the ability of the brain to hold and manipulate information when sensory stimulation is absent [3,4]. Could Ott's new gadget help neuroscientists to study the brain mechanisms supporting working memory? Fuster and his student 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 (Fuster's recollection is presented in Box 1). The resulting publication in 1971 [6] and another independent publication that same year [7] ushered in single-neuron investigations of brain circuits underlying working memory.

The present article takes stock of the past 50 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 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

[1]Center for Neural Science, New York University, 4 Washington Place, New York, NY 20003, USA

*Correspondence: xjwang@nyu.edu (X.-J. Wang).

Box 1. Fuster's reminiscence (personal communication)

In the late 1960s we found in my laboratory that cryogenic inactivation of the lateral prefrontal cortex could produce a reversible deficit in the performance of monkeys in a delayed response task, a test of working memory. Thus we re-established 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 PFC was crucial for the temporary retention of a form of short-term memory that Baddeley later termed '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 micro-electrodes single units from chronic animals, it occurred to me that those cells must 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!), although later it turned out that the sustained delay activity reflects the influence of other factors in addition to memory.

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 a ball hard into 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.

I 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 (i) mnemonic activity is maintained over time when the delay period duration is varied considerably, (ii) 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) [8–14] and continuous space [15–22]. Parametric working memory with monotonic 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 that are engaged in working memory representation. Functional magnetic resonance imaging (fMRI) uncovered similar brain structures that are activated by working memory in humans [25], and these also differentially engage the superficial layers [26]. Research using human and other animals led to the proposal of an integrative theory of the PFC for behavioral flexibility [27]. 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].

Patricia Goldman-Rakic had the vision to tackle the neural circuit mechanism across biological levels of working memory as a window to cognition and mind [33–35]. This vision gradually became realizable with a cross-disciplinary approach combining human brain imaging, single-cell recording and pharmacology, anatomy, *in vitro neurophysiology, and computational modeling [*36]. In the late 1990s and early 2000s, the attractor network paradigm was tested using spiking neural network models endowed with biologically constrained synaptic connections [37–42]. These studies provided initial support for the attractor network model (reviewed in [31]). Has the attractor network model stood the test of time over the past 20 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 depend on 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 oculomotor delayed response (ODR) task [43]. Subsequent studies showed that both the NMDA and AMPA receptors contribute to working memory function in which the fast AMPA receptors predominantly signal 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 was shown to be a canonical feature of the neocortex (reviewed in [47,48]).

The theoretical finding that NMDA receptors play a crucial 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 in healthy subjects low-dose ketamine, an NMDA receptor antagonist, could induce working memory deficits [49] similar to those observed in schizophrenic subjects who display NMDA receptor hypofunction [50–52]. This insight helped to 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 [55–57], 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 with 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 a premotor area, the anterior lateral motor cortex (ALM), 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 performed systematically 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 being 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 timecourse of ALM neural firing, the trajectory of population activity converges to one of two

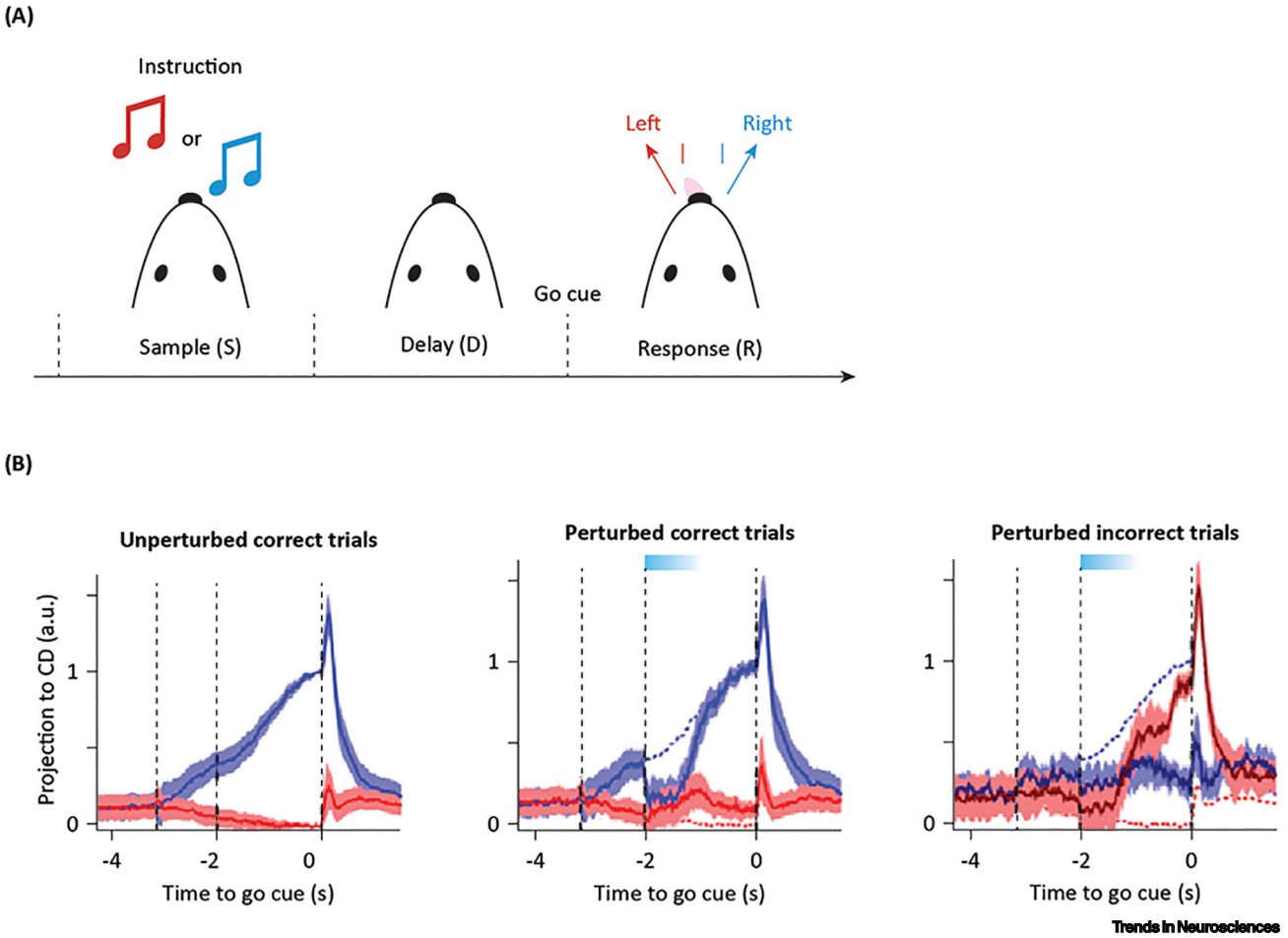


Figure 1. Mnemonic activity in mouse performing a delayed response task. (A) One of two sensory cues is presented briefly, which can be a high or low tone in an auditory task, or a near or far location of an object on the whisker in a somatosensory task. The two stimuli are mapped to left and right licking 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 1D 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 the middle panel but this time optogenetic manipulation induces an incorrect response (left, red). ALM decoding still predicts the erroneous movement direction, demonstrating its correlation with behavior performance. Abbreviations: ALM, anterior lateral motor cortex; a.u., arbitrary units; CD, coding direction. Figure reproduced, with permission, from [62].

fixed endpoints in the state space of recorded neural population activity, in support of discrete attractor models (Figure 1B) [63]. Fourth, optogenetic inactivation during the delay period revealed that thalamocortical connections are important for the maintenance of delay period activity in the ALM [64].

In this task, because the sensorimotor 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, such as delayed match to sample (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 [66–68]. 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 heterogeneous delay activity

Although the attractor model has received theoretical and empirical support, it has been challenged on the grounds 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 more depending on the precise recording location [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 [16,75], in support of its importance at the behavioral level.

Crucially, 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 analyze if attractors are steady-states [37,38,41,76]. However, attractor states do not need 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]. Persistent activity may also exhibit chaotic dynamics [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 the stimulus-selectivity of a recorded neuron may only be detectable in a brief portion of the delay period, and each cell shows statistically significant selectivity at different times ([79,80]). 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 another time t', thus the quality of decoding is shown in a 2D '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 on the upper right corner, demonstrating that working memory representation is relatively stable over time.

Studies using cross-temporal classification analysis have 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 are time-varying when information processing and manipulation are required during the delay period; sometimes a code is stable in a time-window, and 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, principal component analysis (PCA) was applied to PFC neural trajectories using data from ODR [16] and VDD [23] monkey experiments

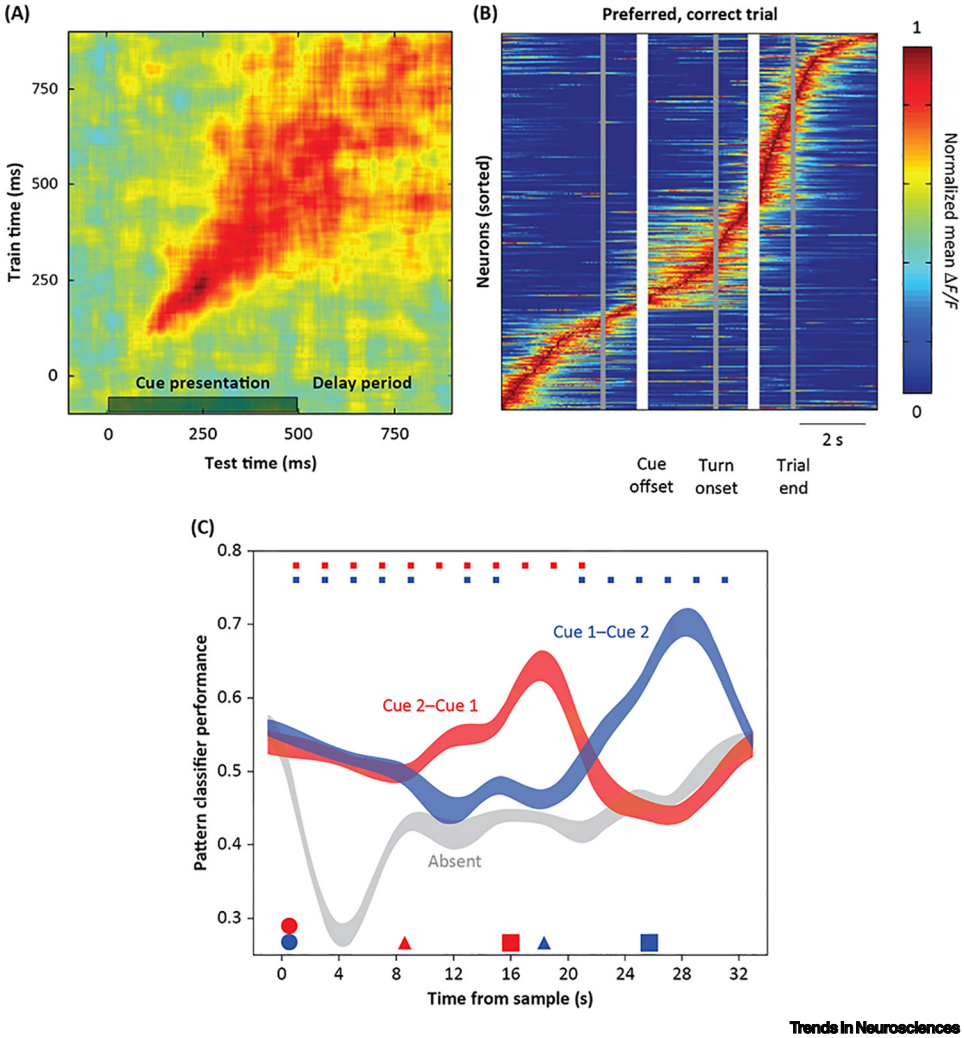
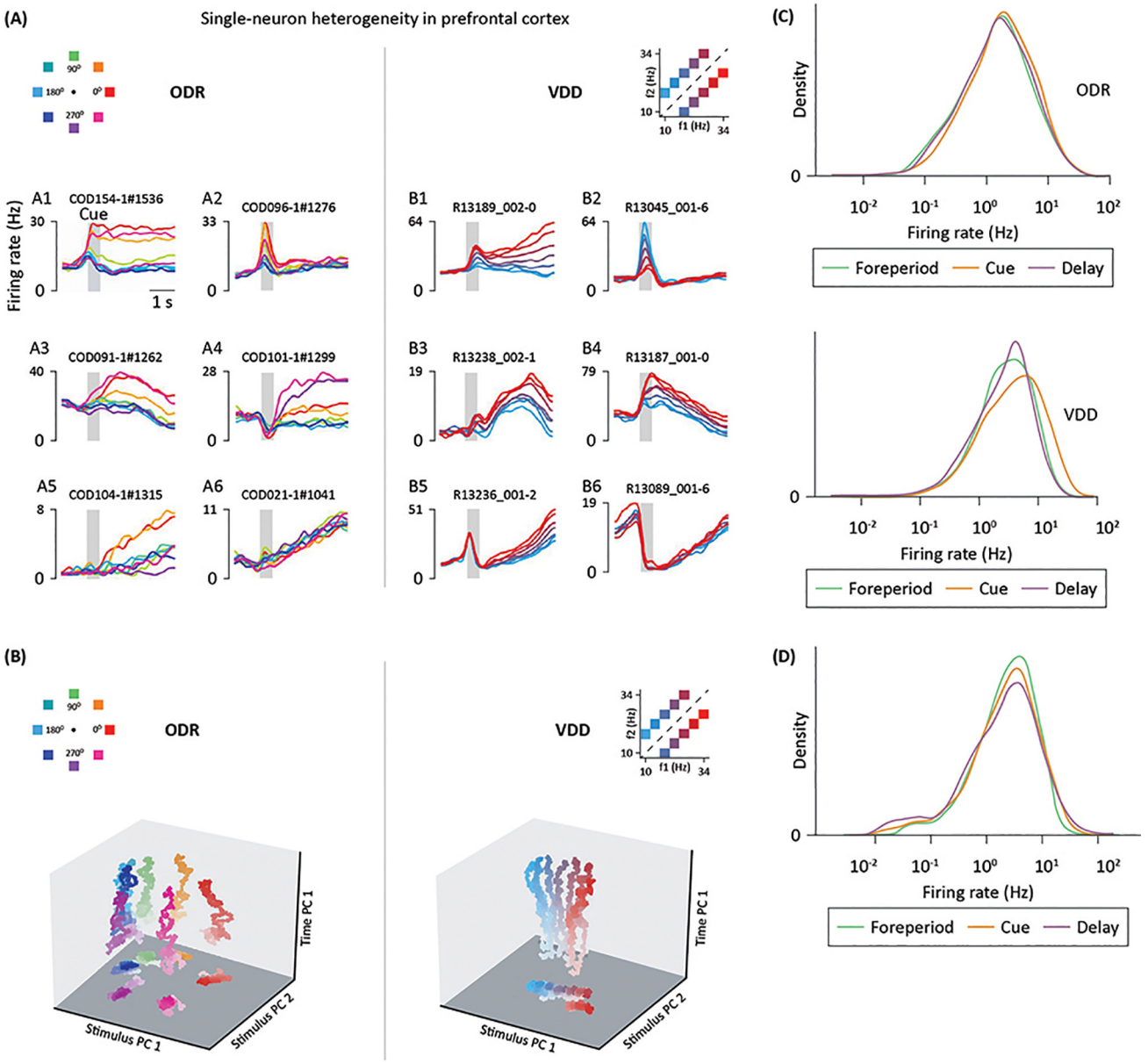


Figure 2. Analysis of information coding by delay period activity. (A) Cross-time classification matrix of recorded neurons for a remembered task. Classifiers are trained to discriminate trial type at one timepoint (*y* axis) and are tested at another timepoint (*x* axis). (B) In a delayed response task, calcium imaging of choice-specific cells (one cell per row) in the posterior parietal cortex of a behaving mouse. Traces were normalized to the maximal activity of each cell on preferred trials and sorted by the peak time. (C) Decoding from human fMRI blood oxygen level-dependent (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 indicate 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 presentation of the second cue, decoded evidence for the two categories was reversed for the remainder of the trial. Abbreviation: *ΔF/F, change in fluorescence intensity*. Panel (A) is reproduced, with permission, from [82], (B) from [83], and (C) from [84].

[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) as proposed in [89]. This observation was reproduced in attractor network models [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.



Trends in Neurosciences

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 oculomotor delayed response (ODR, left) and vibrotactile delayed discrimination (VDD, right) tasks. Different colors correspond to different stimuli. (B) Demixed principal component analysis (PCA) of prefrontal cortex (PFC) 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 (time PC1). (C) Firing rate distributions of PFC neurons in behaving monkeys, plotted with a logarithmic scale along the *x* axis and a 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, with permission, from [88], panel (D) was generated, with permission, from data provided by Nuo Li [63].

Firing activity during 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

during 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, analysis of peak times of spiking activity of recorded neurons did not support sequential activation underlying delay period information coding [62], 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 thus the computational benefit of such a scheme in comparison with a stationary code in the first place remains unclear.

Some 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 behavioral response [94–96]. Corroborative evidence was also reported in a delayed response task in mice 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 [62]. Other temporal changes require different explanations, some of which may be related to uncontrolled factors in an experiment, such as micro-behavior that is not necessary to perform the task [97].

Independently of temporal variations in the firing of a cell, 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 recently elaborated attractor models display considerable cell-to-cell heterogeneities [78,98–102]. In the brain, heterogeneity could arise from variations of biological properties across individual cells in a well-defined population, or/and because several subtypes of neurons are being recorded [103]. Heterogeneity across neurons may also be understood in terms of desirable functions such as mixed-selectivity that is 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 elevated activity should eventually disappear if the delay period is sufficiently long.

The 'activity-silent state' model posits that a memorandum can be encoded by 'hidden' variables that are 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 at synapses 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 in [109]), which should still be considered to be an attractor rather than an 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 the synaptic machinery [110]. Interestingly, STF and other slow synaptic or cellular processes could induce history-dependence across trials [110–112], 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 types of measurements – ranging from single-neuron physiology and electroencephalography (EEG)/magnetencephalography (MEG) to fMRI blood oxygen level-dependent (BOLD) signal. For instance, in a monkey experiment, local field potential (LFP) displays brief episodes of synchrony in the gamma frequency band (around 40 Hz), and this was interpreted as being inconsistent with the sustained activity model [113,114]. However, persistent activity of single cells often coexists with intermittent and weak LFP rhythms [115–118]. Furthermore, if brief bursts are the neural substrate of working memory representation, this predicts 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 several neural assemblies in a network and therefore induces STF at their interconnections, then a later non-selective global signal (a 'pinging' of the entire network) would selectively 're-awaken' 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 remains in some hidden state that is not detectable by the BOLD signal, with the caveat that fMRI measurements are not directly related to spiking neural activity. Moreover, transcranial magnetic stimulation (TMS) could reactivate the 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].

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

crucial because a requirement for normal working memory function is the ability of the brain to filter out irrelevant distracting sensory flow [22,39,41,123–125]. Modeling work showed that a synaptic memory trace is strongest for the latest shown stimulus because signals of earlier stimuli decay [126]. Therefore, in the absence of some additional control mechanism, such a passive memory trace cannot realize working memory in the face of distractors that are presented after behaviorally relevant stimulation.

One argument for the activity-silent state model is that spikes are costly [127], and 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 was seen 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, whereas 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 grounds 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 [3,4,129,130]. Even simple delay-dependent tasks may entail information manipulation to transform 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 random, and the network is incapable of any function. If a to-be-learnt task involves a mnemonic delay, this approach does not make an *a priori* 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 the 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 connections are dominated by short-term depression (Figure 4B, left) whereas others are dominated by STF (Figure 4B, right). The overall synaptic efficacy is the product of the depression factor and facilitation factor. In a motion-direction 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 maintain a memory trace of the sample, in which case activity is not necessary. However, with gradually prolonged delay duration, the solution found by learning in a trained RNN increasingly depends on decoding by the firing activity of recurrent units (Figure 4C). This is because, when the delay period is long compared to the biological time-constants of STF, in an activity-silent state scenario a passive memory trace would decay away before the end of the mnemonic delay, 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,

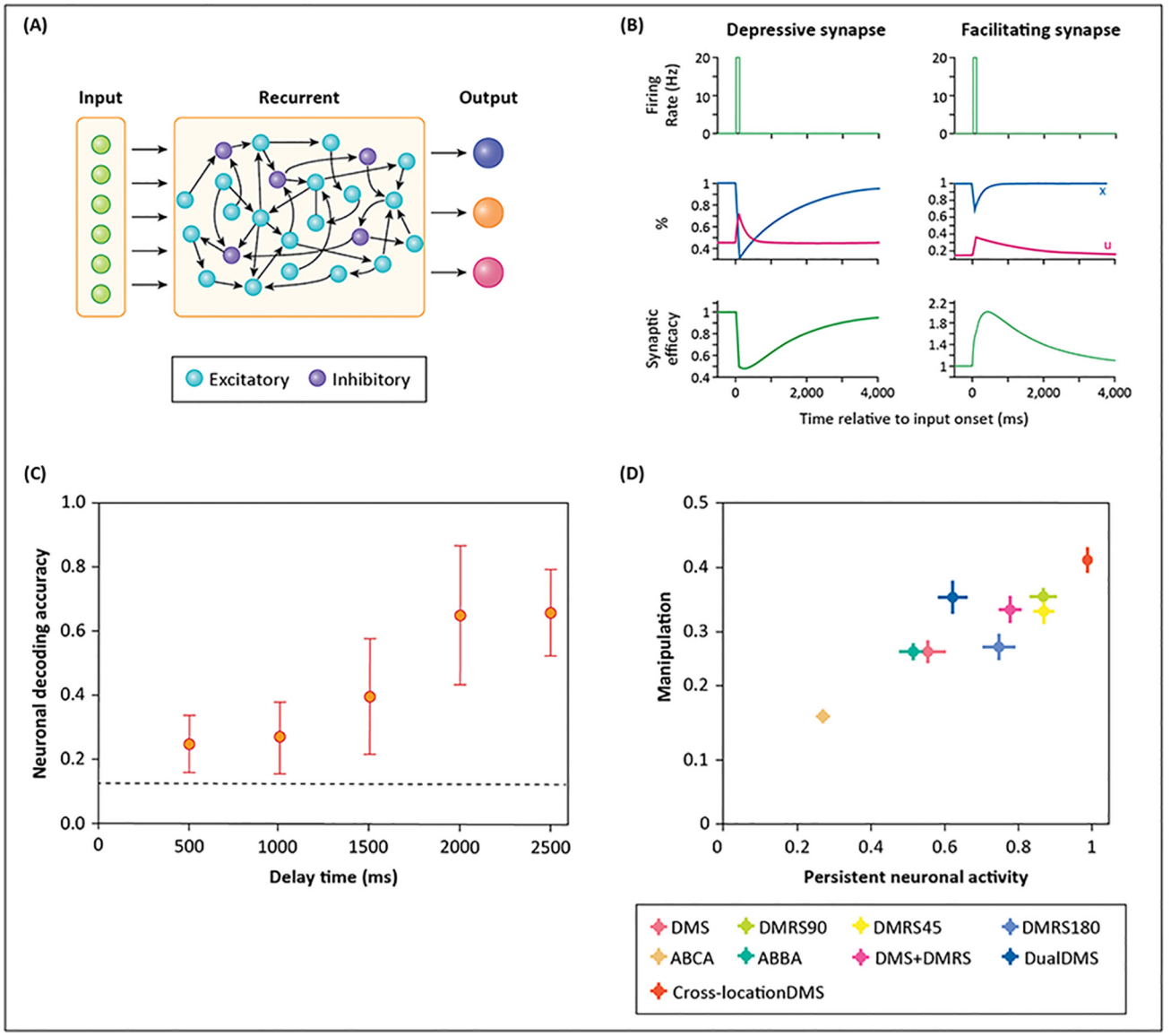


Figure 4. A recurrent neural network trained by machine learning to perform working memory tasks. (A) Model scheme. (B) Short-term facilitation (STF) and short-term depression in response to a pulse input. Variable *u* (red) is the facilitation factor, and *x* (blue) is the depression factor, both defined 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 the delay becomes longer than the biological time-constants of STF. (D) Scatterplot showing 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 of information necessary to perform a delayed response task (*y* axis) across nine different tasks (indicated by colored crosses). Abbreviations: DMRS, delayed match to rotated sample; figures indicate the rotation of the target test in degrees. ABBA and ABCA represent the order of presentation of items (ABBA is more demanding than ABCA because the match response should correspond to the second A but not second B). Adapted, with permission, from [101].

subjects must decide whether the test direction is the same as the sample direction rotated by 90°. In this case, even with a short delay, persistent activity emerged naturally 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, taking advantage of close interactions between experimentation using delay-dependent tasks and biologically based computational models. An important concept running through this research is of '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, although 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 fits naturally into the attractor network model [107,110,112]. Alternatively, if a memory trace is encoded solely by a hidden state such as synaptic efficacy that is endowed with short-term plasticity, physiological experiments should be able to detect the trace [112,136]. 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 for 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, persistent firing 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

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 an active process caused by interference from other external or mental events.

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 reverberations through feedback loops at 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 that are shown after a behaviorally relevant stimulus 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? In the scenario of sequential activation, what is the mechanism that allows a downstream system to read out the stored information at different timepoints?

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 observation that total neural population activity remains unchanged 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 prospective coding take place in a neural circuit?

What is the biological mechanism of the 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?

spatiotemporal mnemonic processes in a working memory circuit, as well as the distinction between working memory and 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 PFC [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 [126,137–140].

Acknowledgments

I thank Bijan Pesaran and Albert Compte for critical reading of the manuscript. This work was supported by the National Institutes of Health (grant R01MH062349), the Office of Naval Research (N00014-17-1-2041), the National Science Foundation (NeuroNex 2015276), and the James Simons Foundation (543057SPI).

Declaration of interests

The author declares no conflicts of interest.

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*.

References

1. Jacobsen, C.F. (1936) Studies of cerebral function in primates. I. the functions of the frontal association areas in monkeys. *Comp. Psychol. Monogr.* 13, 1–682
2. Pribram, K.H. *et al.* (1952) Effects on delayed-response performance of lesions of dorsolateral and ventromedial frontal cortex of baboons. *J. Comp. Physiol. Psychol.* 45, 565
3. Baddeley, A. and Hitch, G.J. (1974) Working memory. In *The Psychology of Learning and Motivation: Advances in Research and Theory* (Bower, G.A., ed.), pp. 47–89, Academic Press
4. Baddeley, A. (2012) Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29
5. Fuster, J.M. and Alexander, G.E. (1970) Delayed response deficit by cryogenic depression of frontal cortex. *Brain Res.* 20, 85–90
6. Fuster, J.M. and Alexander, G. (1971) Neuron activity related to short-term memory. *Science* 173, 652–654
7. Kubota, K. and Niki, H. (1971) Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiol.* 34, 337–347
8. Miyashita, Y. (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335, 817–820
9. Miyashita, Y. and Chang, H.S. (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331, 68–70
10. Miller, E.K. *et al.* (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* 16, 5154–5167
11. Freedman, D.J. *et al.* (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316
12. Wallis, J. *et al.* (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411, 953–956
13. Freedman, D.J. and Assad, J.A. (2006) Experience-dependent representation of visual categories in parietal cortex. *Nature* 443, 85–88
14. Sarma, A. *et al.* (2016) Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nat. Neurosci.* 19, 143–149
15. Gnadt, J.W. and Andersen, R.A. (1988) Memory related motor planning activity in posterior parietal cortex of macaque. *Exp. Brain Res.* 70, 216–220
16. Funahashi, S. *et al.* (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* 61, 331–349
17. Rao, S.C. *et al.* (1997) Integration of what and where in the primate prefrontal cortex. *Science* 276, 821–824
18. Pesaran, B. *et al.* (2002) Temporal structure in neuronal activity during working memory in marcaque pariental cortex. *Nat. Neurosci.* 5, 805–811
19. Pasternak, T. and Greenlee, M. (2005) Working memory in primate sensory systems. *Nat. Rev. Neurosci.* 6, 97–107
20. Vijayraghavan, S. *et al.* (2007) Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. *Nat. Neurosci.* 10, 376–384
21. Wang, M. *et al.* (2011) Neuronal basis of age-related working memory decline. *Nature* 476, 210–213
22. Suzuki, M. and Gottlieb, J. (2013) Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat. Neurosci.* 16, 98–104
23. Romo, R. and Brody, C.D. (1999) Hern´andez, A. and Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* 399, 470–474
24. Bastos, A.M. *et al.* (2018) Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *Proc. Natl. Acad. Sci.* 115, 1117–1122
25. Courtney, S.M. *et al.* (1998) An area specialized for spatial working memory in human frontal cortex. *Science* 279, 1347–1351
26. Finn, E.S. *et al.* (2019) Layer-dependent activity in human prefrontal cortex during working memory. *Nat. Neurosci.* 22, 1687–1695
27. Miller, E.K. and Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202
28. Amari, S. (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 27, 77–87
29. Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* 79, 2554–2558
30. Amit, D. (1995) Hebbian paradigm reintegrated: local reverberations as internal representations. *Behav. Brain Sci.* 18, 617–626
31. Wang, X.-J. (2001) Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* 24, 455–463
32. Strogatz, S.H. (2016) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering* (2nd edn), Taylor and Francis
33. Goldman-Rakic, P.S. (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational

memory. In *Handbook of Physiology – The Nervous System V* (Plum, F. and Mountcastle, V., eds), pp. 373–417, American Physiological Society
34. Goldman-Rakic, P.S. (1992) Working memory and the mind. *Sci. Am.* 267, 110–117
35. Goldman-Rakic, P.S. (1995) Cellular basis of working memory. *Neuron* 14, 477–485
36. Arnsten, A.F. *et al.* (2010) Dynamic network connectivity: a new form of neuroplasticity. *Trends Cogn. Sci.* 14, 365–375
37. Amit, D.J. and Brunel, N. (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* 7, 237–252
38. Wang, X.-J. (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* 19, 9587–9603
39. Compte, A. *et al.* (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* 10, 910–923
40. Durstewitz, D. *et al.* (2000) Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *J. Neurophysiol.* 83, 1733–1750
41. Brunel, N. and Wang, X.-J. (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* 11, 63–85
42. Tegnér, J. *et al.* (2002) The dynamical stability of reverberatory neural circuits. *Biol. Cybern.* 87, 471–481
43. Wang, M. *et al.* (2013) NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* 77, 736–749
44. van Vugt, B. *et al.* (2020) The contribution of AMPA and NMDA receptors to persistent firing in the dorsolateral prefrontal cortex in working memory. *J. Neurosci.* 40, 2458–2470
45. Yang, S. *et al.* (2021) NMDAR neurotransmission needed for persistent neuronal firing: potential roles in mental disorders. *Front. Psychiatry* 12, 337
46. Wang, X.-J. *et al.* (2004) Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1368–1373
47. Kepecs, A. and Fishell, G. (2014) Interneuron cell types are fit to function. *Nature* 505, 318–326
48. Tremblay, R. *et al.* (2016) GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron* 91, 260–292
49. Krystal, J.H. *et al.* (1994) Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. psychotomimetic, perceptual, cognitive, and neuroendocrine responses. *Arch. Gen. Psychiatry* 51, 199–214
50. Coyle, J.T. *et al.* (2003) Converging evidence of NMDA receptor hypofunction in the pathophysiology of schizophrenia. *Ann. N. Y. Acad. Sci.* 1003, 318–327
51. Wang, X.-J. (2006) Toward a prefrontal microcircuit model for cognitive deficits in schizophrenia. *Pharmacopsychiatry* 39, 80–87
52. Stein, H. *et al.* (2020) Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat. Commun.* 11, 4250
53. Montague, P.R. *et al.* (2012) Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80
54. Wang, X.-J. and Krystal, J. (2014) H. Computational psychiatry. *Neuron* 84, 638–654
55. Wang, X.-J. (2002) Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36, 955–968
56. Roitman, J.D. and Shadlen, M.N. (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* 22, 9475–9489
57. Gold, J.I. and Shadlen, M.N. (2007) The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574
58. Wang, X.-J. (2008) Decision making in recurrent neuronal circuits. *Neuron* 60, 215–234
59. Wang, X.-J. (2013) The prefrontal cortex as a quintessential 'cognitive-type' neural circuit: working memory and decision making. In *Principles of Frontal Lobe Function* (2nd edn) (Stuss, D.T. and Knight, R.T., eds), pp. 226–248, Cambridge University Press
60. Guo, Z.V. *et al.* (2014) Flow of cortical activity underlying a tactile decision in mice. *Neuron* 81, 179–194
61. Egorov, A.V. *et al.* (2002) Graded persistent activity in entorhinal cortex neurons. *Nature* 420, 173–178
62. Inagaki, H.K. *et al.* (2019) Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* 566, 212–217
63. Li, N. *et al.* (2016) Robust neuronal dynamics in premotor cortex during motor planning. *Nature* 532, 459–464
64. Guo, Z.V. *et al.* (2017) Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* 545, 181–186
65. Kopec, C.D. *et al.* (2015) Cortical and subcortical contributions to short-term memory for orienting movements. *Neuron* 88, 367–377
66. Wimmer, K. *et al.* (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* 17, 431–439
67. Piet, A.T. *et al.* (2017) Rat prefrontal cortex inactivations during decision making are explained by bistable attractor dynamics. *Neural Comput.* 29, 2861–2886
68. Finkelstein, A. *et al.* (2021) Attractor dynamics gate cortical information flow during decision-making. *Nat. Neurosci.* 24, 843–850
69. Seung, H.S. (1996) How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13339–13344
70. Lim, S. and Goldman, M.S. (2013) Balanced cortical microcircuitry for maintaining information in working memory. *Nat. Neurosci.* 16, 1306–1314
71. Wei, Z. *et al.* (2012) From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J. Neurosci.* 32, 11228–11240
72. Batuev, A. *et al.* (1979) Unit activity of the prefrontal cortex during delayed alternation performance in monkey. *Acta Physiol. Acad. Sci. Hung.* 53, 345–353
73. Baeg, E.H. *et al.* (2003) Dynamics of population code for working memory in the prefrontal cortex. *Neuron* 40, 177–188
74. Fuster, J.M. (2008) *The Prefrontal Cortex* (4th edn), Academic Press
75. Constantinidis, C. *et al.* (2018) Persistent spiking activity underlies working memory. *J. Neurosci.* 38, 7020–7028
76. Renart, A. *et al.* (2003) Mean-field theory of recurrent cortical networks: working memory circuits with irregularly spiking neurons. In *Computational Neuroscience: A Comprehensive Approach* (Feng, J., ed.), pp. 432–490, CRC Press
77. Barbieri, F. and Brunel, N. (2008) Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? *Front. Neurosci.* 2, 114–122
78. Barak, O. *et al.* (2013) From fixed points to chaos: three models of delayed discrimination. *Prog. Neurobiol.* 103, 214–222
79. Zaksas, D. and Pasternak, T. (2006) Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J. Neurosci.* 26, 11726–11742
80. Mendoza-Halliday, D. *et al.* (2014) Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat. Neurosci.* 17, 1255–1262
81. Meyers, E.M. *et al.* (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* 100, 1407–1419
82. Stokes, M.G. *et al.* (2013) Dynamic coding for cognitive control in prefrontal cortex. *Neuron* 78, 364–375
83. Harvey, C.D. *et al.* (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* 484, 62–68
84. Rose, N.S. *et al.* (2016) Reactivation of latent working memories with transcranial magnetic stimulation. *Science* 354, 1136–1139
85. Meyers, E.M. (2018) Dynamic population coding and its relationship to working memory. *J. Neurophysiol.* 120, 2260–2268
86. Kamiński, J. and Rutishauser, U. (2020) Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Ann. N. Y. Acad. Sci.* 1464, 64–75
87. Cavanagh, S.E. *et al.* (2018) Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat. Commun.* 9, 3498

88. Murray, J.D. *et al.* (2017) Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* 114, 394–399
89. Druckmann, S. and Chklovskii, D.B. (2012) Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol.* 22, 2095–2103
90. Bolkan, S.S. *et al.* (2017) Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat. Neurosci.* 20, 987–996
91. Zhu, J. *et al.* (2020) Transient delay-period activity of agranular insular cortex controls working memory maintenance in learning novel tasks. *Neuron* 105, 934–946
92. Goldman, M.S. (2009) Memory without feedback in a neural network. *Neuron* 61, 621–634
93. Rajan, K. *et al.* (2016) Recurrent network models of sequence generation and memory. *Neuron* 90, 128–142
94. Machens, C.K. *et al.* (2010) Functional, but not anatomical, separation of 'what' and 'when' in prefrontal cortex. *J. Neurosci.* 30, 350–360
95. Markowitz, D.A. *et al.* (2015) Multiple component networks support working memory in prefrontal cortex. *Proc. Natl. Acad. Sci.* 112, 11084–11089
96. Brody, C. *et al.* (2003) Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex* 13, 1196–1207
97. Musall, S. *et al.* (2019) Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* 22, 1677–1686
98. Renart, A. *et al.* (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* 38, 473–485
99. Hansel, D. and Mato, G. (2013) Short-term plasticity explains irregular persistent activity in working memory tasks. *J. Neurosci.* 33, 133–149
100. Chaisangmongkon, W. *et al.* (2017) Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron* 93, 1504–1517
101. Masse, N.Y. *et al.* (2019) Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* 22, 1159–1167
102. Yang, G.R. *et al.* (2019) Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297–306
103. Hirokawa, J. *et al.* (2019) Frontal cortex neuron types categorically encode single decision variables. *Nature* 576, 446–451
104. Rigotti, M. *et al.* (2010) Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* 4, 24
105. Rigotti, M. *et al.* (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590
106. Stokes, M.G. (2015) 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* 19, 394–405
107. Hempel, C.M. *et al.* (2000) Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* 83, 3031–3041
108. Wang, Y. *et al.* (2006) Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* 9, 534–542
109. Mongillo, G. *et al.* (2008) Synaptic theory of working memory. *Science* 319, 1543–1546
110. Pereira, J. and Wang, X.J. (2015) A tradeoff between accuracy and flexibility in a working memory circuit endowed with slow feedback mechanisms. *Cereb. Cortex* 25, 3586–3601
111. Bliss, D.P. *et al.* (2017) Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* 7, 14739
112. Barbosa, J. *et al.* (2020) Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* 23, 1016–1024
113. Lundqvist, M. *et al.* (2016) Gamma and beta bursts underlie working memory. *Neuron* 90, 152–164
114. Lundqvist, M. *et al.* (2018) Working memory: delay activity, yes! persistent activity? maybe not. *J. Neurosci.* 38, 7013–7019
115. Brunel, N. and Hakim, V. (1999) Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput.* 11, 1621–1671
116. Brunel, N. and Wang, X.-J. (2003) What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitationinhibition balance. *J. Neurophysiol.* 90, 415–430
117. Wang, X.-J. (2010) Neurophysiological and computational principles of cortical rhythms in cognition. *Physiol. Rev.* 90, 1195–1268
118. Palmigiano, A. *et al.* (2017) Flexible information routing by transient synchrony. *Nat. Neurosci.* 20, 1014–1022
119. Li, D. *et al.* (2021) Trial-to-trial variability of spiking delay activity in prefrontal cortex constrains burst-coding models of working memory. *BioRxiv* Published online February 1, 2021. https://doi.org/10.1101/2021.01.30.428962
120. Wolff, M.J. *et al.* (2017) Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* 20, 864–871
121. Myers, N.E. *et al.* (2017) Prioritizing information during working memory: beyond sustained internal attention. *Trends Cogn. Sci.* 21, 449–461
122. Christophel, T.B. *et al.* (2018) Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* 21, 494–496
123. Sakai, K. *et al.* (2002) Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nat. Neurosci.* 5, 479–484
124. Gazzaley, A. and Nobre, A.C. (2012) Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135
125. Buschman, T.J. (2021) Balancing flexibility and interference in working memory. *Ann. Rev. Vision Sci.* 7, VS07CH0
126. Froudist-Walsh, S. *et al.* (2021) A dopamine gradient controls access to distributed working memory in the large-scale monkey cortex. *Neuron* 109 Published online September 17, 2021. https://dx.doi.org/10.1016/j.neuron.2021.08.024
127. Attwell, D. and Laughlin, S.B. (2001) An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.* 21, 1133–1145
128. Beukers, A.O. *et al.* (2021) Is activity silent working memory simply episodic memory? *Trends Cogn. Sci.* 25, 284–293
129. Cowan, N. (2008) What are the differences between long-term, short-term, and working memory? *Prog. Brain Res.* 169, 323–338
130. Trübutschek, D. and Marti, S. (2019) Uebersch¨ar, H. and Dehaene, S. Probing the limits of activity-silent non-conscious working memory. *Proc. Natl. Acad. Sci.* 116, 14358–14367
131. Wu, Z. *et al.* (2020) Context-dependent decision making in a premotor circuit. *Neuron* 106, 316–328
132. Ehrlich, D.B. and Murray, J.D. (2021) Geometry of neural computation unifies working memory and planning. *BioRxiv* Published online February 1, 2021. https://doi.org/10.1101/2021.02.01.429156
133. Yang, G.R. and Wang, X.-J. (2020) Artificial neural networks for neuroscientists: a primer. *Neuron* 107, 1048–1070
134. Zylberberg, J. and Strowbridge, B.W. (2017) Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annu. Rev. Neurosci.* 40, 603–627
135. Cogan, G.B. *et al.* (2014) Sensory-motor transformations for speech occur bilaterally. *Nature* 507, 94–98
136. Fujisawa, S. *et al.* (2008) Behavior dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. Neurosci.* 11, 823
137. Leavitt, M.L. *et al.* (2017) Sustained activity encoding working memories: not fully distributed. *Trends Neurosci.* 40, 328–346
138. Christophel, T.B. *et al.* (2017) The distributed nature of working memory. *Trends Cogn. Sci.* 21, 111–124
139. Mejias, J. F. and Wang, X-J. Mechanisms of distributed working memory in a large-scale model of the macaque neocortex. *BioRxiv*. Published online April 2, 2021. https://doi.org/10.1101/760231
140. Wang, X.-J. (2020) Macroscopic gradients of synaptic excitation and inhibition in the neocortex. *Nat. Rev. Neurosci.* 21, 169–178