Alternative models of working memory maintenance: how memory information is stored in a dynamic manner?

Jingtai Liu

Michigan State University

Abstract

Persistent activity has been interpreted as the neural substrate underlying working memory for decades. However, recent studies suggested that these persistent activity findings might be an "artifact" of trial-averaging. In single-trial analyses, activity often occurs in sparse, synchronous bursts, both for single neurons and local networks. Alternative working memory models depending on the short-term synaptic plasticity and rhythmicity of discharges were examined and concluded that each of these models could explain a range of memory-guided behaviors which are hard to be explained with persistent activity models. Some attempts to reconcile the discrepancy of persistent activity and dynamic coding frameworks were introduced. In the end, I pointed out some directions which are worth considering in the future.

Working memory (WM) refers to the cognitive ability to store and manipulate information in mind, over a time span of seconds, for future use. It is a core component in a variety of high-level cognitive functions such as language, reasoning, problem solving, and abstract thought (Baddeley, 1992). WM itself has different components including encoding, maintenance, and retrieval of information. Of these components, the ability to maintain information in the absence of sensory input has always been an active topic of WM research.

To understand the mechanism of WM maintenance, it is essential to know its neural basis. Many neurophysiological studies in nonhuman primates which identified single neuron activities (Fuster & Alexander, 1971; Kubota & Niki, 1971; Miller, Erickson, & Desimone, 1996), and neuroimaging studies (e.g., functional Magnetic Resonance Imaging (fMRI) and electroencephalography/magnetoencephalography (EEG/MEG)) of humans which recorded population activities from large cortical assemblies (Druzgal & D'Esposito, 2003; Leung, Gore, & Goldman-Rakic, 2002) have suggested that persistent activity during retention after a stimulus was no longer present is critical for maintaining WM information. Sreenivasan, Curtis, and D'Esposito (2014) gave a detailed definition of persistent neural activity: the above-baseline neural activity begins during the sample presentation and remains stable and elevated throughout the delay, and returns to baseline at the end of the trial. In this review, I consider persistent activity as a sustained and stationary pattern of neural activity. It is worth noting that persistent activity is not equivalent to perfectly stationary activity during the delay interval, for example, some researchers have indicated that content-specific delay activity could vary with the task relevance and exhibit a ramping activity in anticipation of the response (Constantinidis et al., 2018; Stokes, 2015). This persistent activity model has been predominant for decades.

However, recent studies challenged this model. Studies have suggested that stable, persistent activity may not be necessary for WM maintenance. Instead, delay activity of content-specific neurons is dynamic and evolved over time. A related class of attractor models also suggested that information is only expressed as spiking during short-lived attractor states (Lundqvist et al., 2016). The term ‘attractor' describes a system which consists of interacting units (e.g. neurons) evolves over time towards a stable state, given a fixed input (Wang, 2009). And between active states, information is held by selective synaptic changes in the recurrent connections (Mongillo, Barak, & Tsodyks, 2008). Accordingly, alternative frameworks/models have been proposed. Of these frameworks, I will mainly introduce two popular categories: dynamic coding framework, especially the activity-silent model, which is a representative of nonspiking models dependent on synaptic mechanisms instead of persistent spike generation; and rhythmic framework, which suggested that WM information is not persistent but can be conveyed by the frequency and phase of oscillatory activity. In the review, I take a position in favor of the alternative frameworks, because it seemed the persistent activity model does not represent the actual properties of neural activities when reevaluated with new analytic techniques, and the alternative frameworks get support across different anatomical levels, suggesting that they might be a more generic framework for WM.

Persistent activity model and its problems

The very first line of evidence which suggested the importance of persistent activity during WM delay can be traced back to the 1970s. Fuster and Alexander (1971) found that when monkeys performed a manual delayed-response task, neuronal activation in the prefrontal cortex (PFC) exhibited persistent, sustained firing during the delay period compared with that in the intertrial interval. In another study, Kubota and Niki (1971) reported a similar finding during the delay period in the PFC when monkeys performed a delayed alternation task.

Following these pioneer studies, many researchers have found similar persistent activity in the PFC during WM delays. For example, during the delay intervals of a delayed match-to-sample task, activities of PFC neurons of two monkeys were recorded. When comparing the average firing rate across the delay intervals with the spontaneous firing rate before the start of the trial, Miller and colleagues (1996) found more than half of the PFC neurons showed significantly higher activity during the delay intervals. Furthermore, half of the neurons showed a flat delay activity profile when comparing the activity of the first and second half of the delay interval. More recent event-related fMRI studies of humans also suggested a sustained activity pattern during a retention interval. Leung, Gore, and Goldman-Rakic (2002) measured cortical activations with fMRI as human subjects maintained visuospatial memoranda over 18- and 24-sec delay periods. As the single-unit studies in nonhuman primates suggested, they found sustained hemodynamic signal in the middle frontal gyrus throughout both delay periods.

Persistent activity is not merely an epiphenomenon of working memory; it is closely related to working memory performance. The most extensively used paradigm to study visual working memory involves the oculomotor delayed response (ODR) task (Figure 1A). In a typical ODR task, monkeys are presented with a brief stimulus, and after a delay period (≥ 1s), they are required to make an eye movement to previously remembered location. Zhou et al. (2013) used the ODR task and found that prefrontal cortex in peripubertal monkeys generated robust persistent activity in the delay period of the task. Interestingly, this persistent activity was associated with behavioral performance: diminished sustained delay period activity tended to result in errors, whereas neuronal activity in the delay period in the correct trials was robust and was not eliminated by a distracting stimulus.

Another line of evidence linking behavioral performance and persistent activity came from studies which suggested the persistent activity is scaled with working memory load. That is, as the number of items that needs to be maintained increases, the level of delay period activity also increases until close to the limits of the individual's short-term memory capacity (Curtis & Lee, 2010). For example, in a fMRI study, Druzgal and D'Esposito (2003) parametrically varied the mnemonic load of a face delayed recognition task to test the roles of the PFC and the fusiform face area (FFA) in face processing. Their results suggested that activity during the retention period increased parametrically with memory load in both the PFC and FFA. Altogether, this body of work provided a detailed picture of the relationship between persistent activity and working memory-related behavioral performance.

The results I presented so far were concentrated in the PFC. The primary role of PFC is proposed to keep the internal representations of relevant sensory events online, and beyond that, PFC is also crucially responsible for integrating separate representations which all contingent on each other (Curtis & D'Esposito, 2003). Other models have emphasized more of a role for the PFC in manipulating information rather than storing information in WM. These models proposed that PFC provides the top-down control over more posterior areas where the sensory information is actually maintained (Curtis & D'Esposito, 2003; Postle, 2006). A recent lesion study, using monkeys with lateral PFC lesion and a delayed-match-to-sample WM task, provided some support to this view (Pasternak, Lui, & Spinelli, 2015). The researchers found that lateral PFC lesions caused deficits in monkeys' ability to compare current and remembered directions of two moving stimuli, separated by a delay. These deficits were independent of motion coherences of stimuli signaling motion direction, but were most pronounced when the task required rapid reallocation of spatial attention, i.e., the comparison stimulus appeared in an unpredictable visual field location compared to remembered stimulus. Thus, lateral PFC's role is more likely to be attending and accessing the preserved motion signals rather than their storages. Whether the role of PFC is to preserve sensory information per se or reflects control processes such as monitoring and selection, it is no doubt the PFC has a predominant role in WM maintenance. In fact, the PFC has long been thought to be the most important substrate for WM (Curtis & D'Esposito, 2003).

However, persistent activity is not exclusively found in the PFC during working memory maintenance. Persistent activity has also been reported in the parietal cortex, different visual areas of the inferotemporal cortex, early sensory cortices including primary visual cortex and somatosensory cortex (Leavitt, Mendoza-Halliday, & Martinez-Trujillo, 2017), as well as subcortical regions including the basal ganglia and thalamus (Riley & Constantinidis, 2016). For instance, Woloszyn and Sheinberg (2009) reported robust delay activity in the inferior temporal cortex of monkeys performing a delayed match-to-sample task, and there remained a small fraction of neurons that represented WM information even in the presence of the distracting stimulus. But, even though persistent activity seemed prevalent across brain regions, the number of studies reporting sustained activity in areas of the PFC exceeds that of any other region (Leavitt, Mendoza-Halliday, & Martinez-Trujillo, 2017).

There are several possible, and not mutually exclusive, models to explain how persistent activity is generated and controls WM (Curtis & Lee, 2010). The most accepted model proposed that persistent activity emerges from reverberations in a recurrent network (Compte, 2006; Wang, 2001). More specifically, persistent activity might be the result of mutual excitation between neurons that would be self-sustained via dense reciprocal synaptic connections, the activity would then be self-maintained for a much longer time than the biophysical time constants (tens of milliseconds) of fast electrical signals in neurons and at synpases. This reverberation can be observed in local neural circuit such as fontal lobes. For example, Goldman-Rakic (1995) proposed a columnarly organized cortical network model for the PFC, in which persistent activity arises from reverberatory excitation, and stimulus selectivity is formed by recurrent inhibition. This reciprocal excitation can also be observed within a network of areas including both cortical and subcortical areas, such as the cortico-striato-thalamo-cortical loop (Hikosaka, Takikawa, & Kawagoe, 2000; Wang, 2001;Watanabe & Funahashi, 2004). Consistent with this argument, thalamic and caudate neurons showed eleaveted persistent activity whereas ouptut neurons from basal ganglia showed sustained inibition during the delay period of a WM task (Wang, 2001). Another model proposed that persistent activity relies on intracellular signals. It suggested that specific membrane currents or cumulative changes in the concentration of intracellular calcium might account for persistent activity in individual cortical neurons (Curtis & Lee, 2010).

Even though the persistent activity model is clearly important and has been successful in the field of working memory in past decades, some recent results have complicated this view. Lundqvist, Herman, and Miller (2018) pointed out this model's most significant shortcomings: 1) it is not energy-saving and is labile. This model assumes the mnemonic information is stored via persistent activity which might be metabolically expensive, and this pattern is not resistant to distractors or lacks the compatibility with storage of multiple items, because the memory tends to be lost when the activity is disrupted; 2) many early persistent activity findings used the so-called ODR task, which might confound delay persistent spiking with motor planning activity; 3) importantly, classical findings of persistent activity were mainly based on averaged results of many individual trials, and thus might not reflect the real property of individual neurons. The following paragraphs will focus on the second and third points to explain why the persistent activity account might be shaken when examined with new analytical techniques and observations.

The evidence researchers used to argue for the persistent activity usually involved the ODR task. However, recent researchers (Leavitt, Mendoza-Halliday, & Martinez-Trujillo, 2017; Lundqvist, Herman, & Miller, 2018) suggested that this paradigm had an issue because the action planning activity was also involved during the delay period in addition to the activity of interest for WM maintenance. For example, in a recent study, Markowitz, Curtis, and Pesaran (2015) used the ODR task in combination with the large-scale recording from neurons across the lateral PFC of macaque monkeys and found that WM is composed of three anatomically specific modes of persistent activity. The first two modes encode early and late forms of memory storage, and the third encodes response preparation. Information encoded in the response preparation directly influences the timing and accuracy of planned behavior. They also used a database of recordings from chronically implanted movable electrode arrays to map the spatial organization of persistently active neurons, and the result showed that response network was concentrated in anterior PFC, proximal to area 46. This result suggested that prefrontal neurons which were believed only engaged in representing stimulus properties might represent motor preparation instead.

More carefully controlled experiments have been conducted to remove the confounding premotor signals, these experiments tried not to specify motor response until after the delay. Classical examples included delayed match-to-sample task, which only requires subjects to make a response when the probe stimulus reappears at previously remembered cued location (Figure 1B); and its variant, match-nonmatch task, which requires subjects to saccade to either a green or a blue response target depending on whether the probe target matches the previously presented stimulus or not (Figure 1C). It has been shown that with these modified tasks delay activity is often less robust and less sustained (Lundqvist, Herman, & Miller, 2018; Shafi, Zhou, Quintana, Chow, Fuster, & Bodner, 2007).

The most critical and controversial argument Lundqvist, Herman, and Miller (2018) made might be that they suggested persistent activity is an artifact of averaging across trials. When spiking was analyzed on individual trials, it actually occurred in sparse, synchronous bursts, both for single neurons and local networks. The rationale was when activity was averaged with regard to external events, the brain's internal dynamics, which were not time-locked to external events, were ignored. Thus, averaging produced the artifact of persistent activity even though the real activity was sparse.

To support this argument, Lundqvist et al. (2016) developed a novel method to quantify the temporal structure of gamma band (45-100Hz) and beta band (20-35Hz) activity at the local network level on individual trials during working memory. The gamma band was chosen because of its close association with spiking carrying information about memory items (Lundqvist, Herman, & Lansner, 2011). The authors trained two monkeys to maintain multiple colored squares (two or three items, each in unique location) over a short period of time, then, asked the monkeys to make a saccade to the testing squares when they changed colors relative to the encoding squares at the same location. They suggested that when holding WM information in mind, ensembles of neurons in the prefrontal cortex are active in brief bursts and WM information is stored in synaptic changes between bursts. For burst extraction, they used two comparable methods (i.e., band-pass filtering techniques and multi-taper analysis) to estimate temporal profiles of the local field potentials (LFPs) spectral contents within the gamma and beta bands. To quantify the gamma or beta burst, they defined a burst as an increase in power of two standard deviations above the trial mean spectral power for that particular frequency, and with the duration of lasting at least three cycles. Having the burst intervals extracted for gamma oscillations from each trial, they found these gamma waves occurred sporadically (i.e. the occurrence of gamma burst was not periodic, instead, it was widely scattered in frequency because it reflected a transient attractor state). They also proposed that each ensemble of neurons encoded a specific item and produced a different burst of gamma waves. Beta oscillation also occurred in brief, irregular bursts but reflected a default state, as suggested by the anti-correlated relationship between gamma and beta bursts. van Ede, Quinn, Woolrich, and Nobre (2018) conceptualized this burst view: the underlying pulses might be rhythmic, but there exists a threshold to determine whether any given pulse will result in a measurable burst-event. If single pulses cross the threshold, isolated bursts occur. However, when this activity was averaged across trials, it appeared as a familiar sustained and long-lasting gamma activity (Figure 2), but it might be an artifact of averaging across trials.

In a more recent study, Lundqvist and colleagues (2018) reanalyzed a multiple-electrode dataset from a previously published experiment (Warden & Miller, 2007). In that experiment, two monkeys were trained to determine whether a test sequence of two objects matched a sample sequence presented earlier. This task, as with other delayed match-to-sample tasks, carefully controlled the motor planning activity. When the authors reanalyzed the LFPs recorded in prefrontal cortex using single-trial analyses (Lundqvist et al., 2016), they found brief narrowband oscillatory bursts of varying central frequency on single trials, while previous results reported persistent activity using the trial-averaging method.

This transient activity model has some advantages over persistent activity models. Because it suggested that WM items are only expressed as spiking during active states, between active states, items are maintained by short-term synaptic facilitation mediated by increased residual calcium levels at the presynaptic terminals of the neurons that code items (Mongillo, Barak, & Tsodyks, 2008). Spikes (in the form of brief bouts) can induce temporary (< 1s) changes in synaptic weights for memory retention (Constantinidis et al., 2018; Miller, Lundqvist, & Bastos, 2018), and memories stored in synaptic states can be transformed into spiking activity again as a result of global reactivating input (like attention) to the network or through intrinsic network dynamics (Mongillo, Barak, & Tsodyks, 2008). Thus, spiking and short-term synaptic plasticity work together to maintain WM information.

This hybrid mechanism is supposed to be metabolically less expensive, because it does not require continuous generation of spikes to retain memories. During the silent period, the neural circuit responsible for holding information is decoupled from other brain regions because no spike is discharged. Such a decoupling might be the underlying mechanism of modular brain systems, in which different sensory modalities are encoded in different modules (Fusi, 2008). This synaptic weight theory can better handle multiple items because memories are stored in synaptic states, thus there are no overlapping neural activities for different memories. Synaptic weights are also more resistant to interference. Because as less time spent in active attractor states, working memories are less likely to be disrupted by new sensory input (Miller, Lundqvist, & Bastos, 2018).

To sum up, persistent spiking has been thought as the neural substrate of working memory maintenance, and many studies from single neuron/LFP recording, and neuroimaging methods have provided evidence for this claim. However, Lundqvist, Herman, and Miller (2018) pointed out that virtually all of the evidence supporting the persistent spiking model (at least in single neuron and LFP recordings) came from studies that averaged spiking across time and trials. It is argued that when the trials were analyzed individually, activity actually occurred sporadically. Even though the activity might occur in sparse transient bursts in single neurons as well as local networks (e.g., multiple simultaneously recorded neurons and LFPs), some might argue it could still be persistent on a more global scale (i.e., combine enough neurons across highly distributed networks), because transient local activity in different parts of a larger network could be counterbalanced to make the global activity persistent (Lunqvist, Herman, & Miller, 2018). Nevertheless, recent neuroimaging recordings of global activity from EEG and fMRI were not consistent with this argument. In the next section, I will present evidence to demonstrate that activity is also not persistent on the level of populations of neurons.

Dynamic coding in working memory

In the strict model of persistent activity, WM-related activity should keep constant and the information encoded in a given state at time *t* may be encoded in the same state at time *t*+1 (Sreenivasan, Curtis, & D'Esposito, 2014). However, many recent findings suggested this is not the case. Instead, WM-related activity evolves over time; that is, the activity containing information drifts over the course of the trial, but stable representation can still be preserved via this dynamic trajectory through the activity of the neurons in the population (Sreenivasan, Curtis, & D'Esposito, 2014).

Changes in a population code can be evaluated by training a classifier on population spiking. If the population code is persistent, it should not matter whether a classifier is trained on one time point and tested on another, however, if it is dynamic, a classifier trained on one time point will probably only perform well at the same or close time points (Spaak, Watanabe, Funahashi, & Stokes, 2017; Stokes, Kusunoki, Sigala, Nili, Gaffan, & Duncan, 2013). Using this temporal generalization logic, accumulating evidence suggests an important role of dynamic population coding in the maintenance of working memory information in PFC (Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Spaak, Watanabe, Funahashi, & Stokes, 2017; Stokes, Kusunoki, Sigala, Nili, Gaffan, & Duncan, 2013).

For example, Spaak, Watanabe, Funahashi, and Stokes (2017) applied multivariate pattern analysis to explore the population dynamics in lateral PFC in macaques during three variants of the classic ODR task (i.e., the standard ODR task and its variant with varied delays, and a dual task which required monkeys to perform the ODR task and an attention task simultaneously). They observed significant dynamic population coding during both the cue period and the early part of the subsequent maintenance period in the ODR tasks. Furthermore, they used simulated neural populations based on the observed dataset to study the relative contributions of two factors in driving the dynamic coding: different subpopulations of neurons are involved at different time points, and location selectivity in individual neurons changed over time. These results indicated that it is likely that a combination of two components was contributing to the observed dynamic population code.

Studies above demonstrated that without intervening input, population code can change over the memory delay, which is consistent with findings at local networks discussed in the previous section. Another line of evidence further demonstrated that neural coding is not fixed, but depends on the exact context of the current task. Even though the persistent activity model suggests that memory performance will be compromised when activity does not persist during retention (Curtis & Lee, 2010), more and more studies using paradigms which required subjects to perform two tasks together or hold multiple objects showed that persistent activity is not necessary for WM-guided behavior. For example, in a study, Warden and Miller (2007) trained monkeys to memorize a sequence of two objects across a short delay while recording activity of neurons from the lateral PFC. They found that neurons encoding the first stimulus were suppressed during the presentation of the second stimulus; however, after the offset of the second stimulus, the suppressed activity for the first stimulus was 'reactivated'.

Similar findings were also observed in human subjects with fMRI and EEG studies (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). In an fMRI study of multistep delayed-recognition task, Lewis-Peacock and colleagues (2012) presented two sample stimuli concurrently. After the offset of the stimuli and an initial delay period, a retro-cue was presented, indicating which sample was relevant for the first probe, followed by a second delay and the initial recognition probe. After the first probe, a second retro-cue, followed by the third delay, appeared to indicate whether the same item (repeat trials) or the previously uncued item (switch trials) would be tested in the following probe. Thus, during the second delay (before the first probe), subjects had to keep both items in memory, but the third delay would only require the retention of the cued item because subjects knew the uncued item would never be tested. Using decoding algorithms, the authors found that classifier evidence for both stimuli was apparent at trial onset and remained until the onset of the first retro-cue. After the first retro-cue, classifier evidence for the uncued item dropped precipitously to the baseline. However, if the second cue was a switch cue (i.e., previously uncued item would be probed), classifier evidence for previously uncued item was reinstated, while evidence for previously cued item dropped to the baseline. To summarize, these findings suggested that persistent delay activity is not necessary for the process of maintaining WM information. Instead, neural coding exhibits context-dependent responses.

Stokes (2015) further proposed that unattended items do not seem to have a corresponding activity state; instead, they remained in an activity-silent state (i.e., hidden state), even though they were still preserved in memory (Figure 3). However, once attention is directed back to them, the activity state becomes apparent again. This hidden neural state was usually undetected (e.g., classifier evidence for uncued items dropped to chance levels) because recording techniques typically measure activity states only (Stokes, 2015). Nevertheless, Wolff, Ding, Myers, and Stokes (2015) developed a novel way to reveal this hidden state of memory. They employed an analogy of sonar system to describe this method: underwater objects are invisible to sailors, but when a ship uses sonar to emit pings of sound, they can “see” the objects when they receive the sound waves reflected back. Particularly, in the hidden state memory example, if working memories are hiding in an active-silent network of altered synaptic weights, any input (functions as sonar sound) will trigger a unique pattern that depends on the physical properties of input stimulus and the initial state of synaptic weights (underwater objects). Thus, when the input stimulus remains unchanged (sonar sound is fixed), if we can still record neural activity changes, this change should be attributed to the change of hidden state pattern. Critically, they used high-contrast visual stimulus as the ping to the brain. The ping could be three big circles shown side by side with plain white or filled with black-and-white dartboards. The precise feature of a ping probably is not important, as long as it is a high-contrast visual stimulus which can target the memory network properly and is neutral to the information being remembered (Rademaker & Serences, 2017). It was believed that this method can reveal the actual contents of WM (Wolff, Jochim, Akyurek, & Stokes, 2017), rather than the focus of attention (as revealed by conventional decoding approach without this perturbation approach), from EEG signal as a function of time.

This functional perturbation approach combined with multivariate pattern analysis was applied to a working memory task to characterize the functional dynamics of the hidden state for WM (Wolff, Jochim, Akyurek, & Stokes, 2017). In their experiment, two memory items were presented, and participants were instructed to remember both items. Both items were ultimately tested in each trial. However, their priorities were manipulated by blocking the order in which items would be probed early. Decoding results showed that both items were decodable immediately following memory items onset, even though the prioritized item (tested early) was more prominent. However, decoding of the deprioritized item quickly dropped to chance level after item presentation, while the prioritized showed significant decoding until the end of the epoch. This result was consistent with previous evidence which showed only cued item could be decoded. However, what makes this method unique is when a ping was presented during the delay period, both prioritized and deprioritized items became decodable again, which suggested that WM for a temporarily deprioritized item was stored in a hidden state and could be revealed by functional perturbation approach.

At a cognitive level, this activity-silent model resonates well with the state-based models of working memory, especially Oberauer’s three-embedded-components model (Oberauer, 2002, 2009). This model, which is an extension of Cowan’s embedded model (Cowan, 1995), consists of three components: the activated part of long-term memory (LTM), the region of direct access, and the focus of attention. The activated part of LTM keeps all information that could be relevant to the task. The region of direct access contains a subset of the activated LTM, and the focus of attention is a selection device which might hold only one item or chunk at a time. According to the activity-silent model, cued or prioritized items are maintained in an active state, which might also be kept in the focus of attention. However, uncued or deprioritized items, which are preserved in a hidden state, might be temporarily pushed out of the focus of attention, but they are not lost because they might still be in the region of direct access, and can be restored by retro-cues or by transcranial magnetic stimulation (TMS) pulses which can also function as a sonar to ping the brain (Rose et al., 2016).

In previous examples, even though memory item might enter a deprioritized state which dissipated related neural activity, subjects were still consciously aware of the input. However, recent studies had found even when subjects indicated not having seen the memory target, they could still recall the target much better than chance (Soto, Mantyla, & Silvanto, 2011; Trubutschek et al., 2017), and interestingly, the activity-silent mechanism also constitute a plausible neural mechanism for this non-conscious working memory (Trubutschek et al., 2017). In a spatial delayed-response task combined with MEG, Trubutscheck et al. (2017) assessed working memory performance under varying levels of subjective visibility. Their finding suggested that an unseen but correct response stimulus that failed to cross the threshold for sustained activity and subjective visibility could still possess enough activity in high-level cortical circuits to modify short-term synaptic weights, thus could be maintained in an activity-silent state.

The activity-silent state could have many possible neurobiological bases, the common theme of these possible mechanisms is that they do not assume an unbroken chain of sustained spiking (Myers, Stokes, & Nobre, 2017). As discussed previously for single neurons and local neural circuits, the most prominent candidate mechanism for activity-silent state on global scale is short-term synaptic plasticity (Mongillo, Barak, & Tsodyks, 2008). Short-term synaptic change is not only a theoretical framework, direct physiological evidence for this framework also exists. For example, Fujisawa and colleagues (2008) examined large-scale recording of neural activity in the medial prefrontal cortex of the rat during a working memory task at timescales of milliseconds and seconds. Their observations indicated that for a given interneuron, increased activity from one presynaptic neuron can reduce or increase that neuron's control of the interneuron, this *in vivo* evidence was consistent with synaptic facilitation and depression, which argued in favor of synaptic mechanisms of working memory.

Even though converging evidence from univarate and multivariate techniques have supported the activity-silent model, there are some concerns whether this hidden state really exists. Most evidence argued in favor of activity-silent state are based on null effects, and it remains a possibility that researchers just missed some content-related activities. For example, Watanabe and Funahashi (2014) trained monkeys to attend to a specific spatial location when performing an ODR task, and found that during this dual-task period, WM-specific delay activity in prefrontal neurons was attenuated but still noticeable, even under the presence of the most difficult attention task conditions, suggesting that memory information was not completely lost during delay period. In the study of Lundqvist et al. (2016), even though researchers reported that beta and gamma activities occurred in brief and irregular bursts during WM delay on single-trial level, they particularly only chose those activities with high spectral powers as bursts. Setting a threshold might be able to help to eliminate noisy burst-like oscillatory dynamics, but it might also ignore those persistent signals with lower power. Thus, depending on how they set the thresholds, the results might suggest opposite conclusions. Another line of evidence questioned the usefulness of activity-silent WM storage. In a simulation study, Schneegans and Bays (2017) implemented a neural model of working memory simply based on the principle of sustained activity in neural populations. More specifically, they formulated this neural model in a form of recurrent neural network that generates continuous time courses of neural activity patterns. With this sustained neural activity model, they could reproduce both behavioral and fMRI results of a spatial recall task with retro-cues, which were previously taken as evidence for activity-silent working memory state (Sprague, Ester, & Serences, 2016).

To sum up, even though previous studies suggested the neural activity persist during the delay period of WM, more recent studies in combination with decoding techniques indicated that WM-related activity actually evolves over time. Importantly, the persistent activity is not necessary for storing memory information as studies suggested that a memory item can be stored in an activity-silent state. Even though there are some debates on whether this activity-silent memory really exists (Xu, 2017), more and more evidence seems to support its existence (Rose et al., 2016; Sprague, Ester, & Serences, 2016; Wolff, Jochim, Akyurek, & Stokes, 2017). As indicated by the activity-silent model, working memory maintenance does not solely depend on sustained neural firing. Instead, persistent neural activity only reflects sustained attention to the currently task-relevant item, whereas other items can be held in an activity-silent state. However, context changes (e.g., retro-cues) can refresh the synaptic weights and reactive dormant representations. Because of the feature that target-related activities can disappear and reappear intermittently, this model also suggests a possibility that target information might be maintained in a rhythmic manner, i.e., they wax and wane throughout the delay. In the next section, I will provide evidence to show how rhythmic activity has been implicated in WM.

Oscillatory models of working memory

The first suggestion that observed delay activity involved in WM actually has an oscillatory character came from intracranial recordings of the local field potential (Raghavachari et al., 2001). In this study, human subjects were presented with lists of one to four consonants sequentially on the screen, and after a delay period, they had to respond whether a probe item was on the list (i.e., Sternberg task). The researchers found the amplitude of theta (4-8 Hz) oscillations at some cortical and subcortical sites increased at the beginning of the trial, was sustained through the entire trial including the delay, and decreased at the end. However, as described in the previous studies (Lunqvist, Rose, Herman, Brincat, Buschman, & Miller, 2016; Lundqvist, Herman, & Miller, 2018), this elevated and sustained theta rhythm was a result of trial-averaged spectrograms. When analyzed with single-trial spectrograms, the theta oscillations probably would not be sustained, but occur in brief bursts. Brief bursts with interleaved periods of silence might be a way to combine the robustness of persistent activity with more flexible computations, in other words, the periods of silence might be opportunities for the network to evolve and incorporate new information (Lundqvist, Herman, & Miller, 2018).

Several lines of neuronal oscillation studies came from the examination of the firing of individual neurons and LFPs in primates. For example, in one study (Siegel, Warden, & Miller, 2009) monkeys were required to maintain both the identity and the order of two objects over a delay of 1 s, while their LFPs and spikes (multi-unit activity) were both recorded from electrodes implanted in the lateral PFC. During the delay, time-frequency analysis of the LFPs revealed population activity at around 32 Hz. Importantly, there was also a prefrontal spike-LFP synchronization at 32 Hz, and spikes at particular phases relative to the ongoing population oscillations carried the most information about the remembered objects. Moreover, according to the order of stimulus presentation, optimal encoding of the first presented object was significantly earlier in the 32 Hz cycle than that for the second object.

More studies, however, were conducted using EEG and MEG in human subjects. For example, Jensen and colleagues (2002) measured scalp EEG using a Sternberg task which required human subjects to remember a list of consonants. They manipulated working memory load in the task by varying the memory set length. Their results revealed dominant oscillations in the 9-12 Hz alpha band during the interval between the onset of the memory list and the onset of the probe. To determine whether the alpha activity was affected by memory load, they compared power spectra concerning different memory loads, and the results indicated the alpha band power increased with memory load in both posterior and bilateral brain regions during the last 2s of the 2800 ms retention interval. In another MEG study, Jensen and Tesche (2002) recorded neuromagnetic responses while human subjects performing a similar Sternberg task as Jensen and colleagues (2002) implemented. Their results also revealed a spectral peak in the 10-12 Hz alpha band over the back of the head, and a 7-8.5 Hz peak in the theta band over frontal areas which was not discovered in the study conducted by Jensen and colleagues (2002). The frontal theta activity could also increase parametrically with memory load during retention interval.

Altogether, these studies indicated that both the magnitude and the phase of oscillations could be modulated by WM information. This modulation was reported at different frequencies, including theta, alpha (8-13 Hz), and gamma (30-200 Hz) (Roux & Uhlhaas, 2014), and at different anatomical scales, ranging from single neurons/small neuronal populations to large cortical assemblies from the surface of the scalp and brain networks using EEG/MEG recordings (Duzel, Penny, & Burgess, 2010).

In general, it has been argued that neuronal oscillations could provide a temporal reference frame and reflect a feedback mechanism to control cortical excitability and spike timing to influence information held in working memory (Helfrich & Knight, 2016). More specifically, theta activity is mainly involved in the PFC as well as the hippocampal-entorhinal system (Roux & Uhlhaas, 2014). It has been reported that theta activity is increased during the encoding and retention of WM tasks and possibly functions as a gating mechanism controlling relevant information and suppressing irrelevant information (Raghavachari et al., 2004; Sauseng, Griesmayr, Freunberger, & Klimesch, 2010). Theta rhythm is also thought to be important for phase coding of information in WM, and higher frequencies like gamma cycle can be nested into theta cycle to enable the reactivation of the memory representations (Lisman & Idiart, 1995). Gamma frequency, which represents a generic mechanism for the representation of individual WM items, helps to integrate various features of an object by integrating activities from different neuronal populations in cortical and subcortical structures (Hakim & Vogel, 2018; Roux & Uhlhaas, 2014). Alpha frequency, which is most frequently observed in sensory regions and the thalamus (Roux & Uhlhaas, 2014), on the other hand, might not be directly relevant for WM information per se, but reflects the orienting of attention (Foster, Sutterer, Serences, Vogel, & Awh, 2017; Wolff, Jochim, Akyurek, & Stokes, 2017) and protects WM maintenance from irrelevant information or distractors (Bonnefond & Jensen, 2012). For example, WM studies have suggested that decreased alpha activity could facilitate processing in task-relevant brain regions, whereas increased alpha activity may suppress distracting information in task-irrelevant regions (Haegens, Nacher, Luna, Romo, & Jensen, 2011). The gamma and alpha band responses are also considered internally generated because in both cases rhythmic responses can sustain as long as the functionally relevant epoch lasts, even when no temporal structure in the stimulus is presented (Herbst & Landau, 2016).

Many findings have suggested that distinct spectral signatures do not occur in isolation, but are functionally coupled (Helfrich & Knight, 2016). For example, slower and faster rhythms can interact by cross-frequency coupling. The most popular coupling might be phase–amplitude cross-frequency coupling, in which the phase of lower frequency oscillations correlates with the amplitude of higher frequencies (Turi, Alekseichuk, & Paulus, 2018). One intriguing model proposed by Lisman and Idiart (1995) suggested that theta-gamma phase-amplitude coupling provides the necessary neural substrate for limited working memory capacity. According to this model, an individual memory item is represented by each cycle in the gamma rhythm. Because multiple gamma cycles can be entrained within the theta cycle, various items represented by multiple gamma cycles can be activated every theta cycle (Figure 4). As suggested by this model, the capacity limit of working memory emerges because there is only a certain amount of phase space available within a theta cycle to keep multiple items active and separated (Eriksson, Vogel, Lansner, Bergstrom, & Nyberg, 2015), and when storage of more items is attempted, the representations might be overlapped in phase space, which could cause memory errors (Hakim & Vogel, 2018).

The most compelling evidence to support this theta-gamma cross-frequency coupling model came from a study conducted by Bahramisharif and colleagues (2018). In this study, the authors recorded activities from epilepsy patients who had electrocorticography implanted on the surface of their brain. They presented those patients with serial sequences of three random letters which they had to remember during the delay. By examining neural activities at "letter-selective" cortical sites, they found that each of the three letters was encoded as bursts of gamma activity at distinct phases of the theta/alpha band (7-13 Hz). Additionally, the position of each letter in the phase space was dependent upon the presented letter position in the letter list. Thus, the authors provided clear evidence that individual items can be stored in gamma rhythm at distinct periods of the phase of the theta cycle.

One interesting inference can be made from this model is if the theta frequency is lower, the phase of theta cycle will be more prolonged, then it might hold more gamma subcycles, and thus increase WM capacity. This assumption gained support from a study that investigated how working memory load influenced phase coupling (Axmacher, Henseler, Jensen, Weinreich, Elger, & Fell, 2010). The authors found as the number of items (one, two, or four trial-unique novel faces) held in mind increased, the gamma frequency was modulated by lower frequency theta band activity in the human hippocampus. This model has great potential in application; for example, WM capacity could be increased by modulating theta frequency. In an exciting new study, Wolinski and colleagues (2018) used transcranial alternating current stimulation (tACS) to modulate and entrain healthy participants' ongoing network oscillations in the theta frequency range while the volunteers performed a visuospatial working memory task. They found that when tACS at 4 Hz was used to entrain ongoing theta oscillation, participants' working memory capacity was enhanced. However, faster tACS at 7 Hz reduced participants' working memory capacity. In another tACS study of spatial working memory, researchers (Alekseichuk, Turi, de Lara, Antal, & Paulus, 2016) also found that co-stimulation of theta and gamma waves in the prefrontal cortex boosted working memory performance only when gamma rhythms were phase locked to the peaks of theta rhythms. Interestingly, the optimal high gamma frequencies manifested in the 80 to 100 Hz frequency range, when the theta cycle was at 6 Hz.

Another line of evidence to support this rhythmic property of WM maintenance can be inferred from the relationship between internal attention and working memory. In the previous section, I introduced the three-embedded-components model of WM, which assumes the most task-relevant memory item is maintained in the focus of attention. In the time-based resource-sharing model of WM (Barrouillet & Camos, 2012), it further emphasizes the importance of attention in the two main functions of WM: the temporary storage and the processing of information. This model proposed that the maintenance of memory traces depends on their activation through attentional focusing and that working memory will decay as soon as the focus of attention is switched away. Because attention is vital to WM maintenance, it is natural to assume that if the attention itself is a rhythmic process, then it is less likely the WM would be maintained in a stable neuronal coding.

In fact, many behavioral studies have provided evidence for rhythmic sampling during spatial attention (Fiebelkorn, Saalmann, & Kastner, 2013; Fiebelkorn, Pinsk, & Kastner, 2018; Landau & Fries, 2012; Song, Meng, Chen, Zhou, & Luo, 2014; VanRullen, Carlson, & Cavanagh, 2007). An early behavioral study that demonstrated a theta-band attentional rhythmicity was conducted by VanRullen and colleagues (2007). They measured human psychometric functions for target detection as a function of target duration at various set sizes, when fitting different models of attention deployment to the data, they concluded that ongoing rhythmic attention process serially sampled targets at a rate of 7 per second, and this rhythmic manner persisted even when only one item was attended. Later studies using a spatial cueing task in combination with more direct spectral analysis measures on both human subjects (Fiebelkorn, Saalmann, & Kastner, 2013) and monkeys (Fiebelkorn, Pinsk, & Kastner, 2018) demonstrated that spatial attention is associated with theta-rhythmic fluctuations in hit rates at the attended location. Thus, it seems that theta-rhythmic sampling is a fundamental property of spatial attention.

Fiebelkorn, Pinsk, and Kastner (2018) further investigated the neural basis of these internally generated rhythms. They recorded local field potentials from two key regions of the macaque frontoparietal network: the frontal eye field (FEF) and the lateral intraparietal area (LIP). Their results indicated that the theta phase in the frontoparietal network shapes behavioral performance through temporally coordinating two rhythmically alternating states: the "good" and "poor" theta phase, which were determined by whether the behavioral performance was enhanced or attenuated at the cued location. When the target occurred during the "good" theta phase, in the FEF, detection accuracy was modulated by beta-band oscillations (16-35 Hz), and in the LIP, a similar periodic modulation of behavior included beta as well as gamma band oscillations (30-40 Hz). However, when target appeared at the "poor" theta phase, only alpha-band (9-15 Hz) LIP oscillations modulated behavioral performance. Altogether, these results suggested that there is rhythmic sampling during spatial attention and it might be linked to dynamic interplays between hubs of the frontoparietal network.

Not surprisingly, several studies also suggested that mental representations, which are conventionally considered stable and invariant during WM delay, are actually rhythmic and transient. For example, in a MEG study, Fuentemilla and colleagues (2010) trained a multivariate pattern classifier during the presentation of indoor or outdoor image. When testing the classifier during the delay period of a working memory task, they observed that decodable representations of memorized images recurred at a theta rhythm. Such periodic refreshing of internal representations could potentially serve as the neural correlate of conscious rehearsal (Trubutschek et al., 2017). A similar rhythmic pattern of mental representation was also found in our recent behavioral study of WM (Liu, Liu, & Ravizza, 2018, under review). In that study, we sampled behavioral performance densely in time using a delayed-estimation WM task. In the experiment, participants memorized orientations of two Gabor patches followed by a retro-cue with different probabilities (i.e., 100%, and 50%) indicating the likelihood that a particular orientation would have to be recalled (Figure 5). Memory performance was densely sampled by systematically varying the delay interval between the retro-cue and the probe stimulus. We observed rhythmic patterns mainly in the theta band in the time course of memory recall with both probabilistic retro-cues. These findings indicated that there are worse and better moments during working memory maintenance. Just like results reported in attention studies (Fiebelkorn, Pinsk, & Kastner, 2018), if the target item appeared during a ‘good’ phase, it is more likely to be precisely recalled. Otherwise, memory precision for the target item in a ‘bad’ phase might be degraded.

A similar finding was revealed by Peters, Rahm, Kaiser, and Bledowski (2018). Unlike in our study, the authors were mainly interested in whether object-based attention fluctuated during WM maintenance. In their task, participants were required to memorize four positions located at the endpoints of two objects. During the retention interval, a cue appeared to indicate the position that would most likely to be probed in a delayed match-to-sample decision. The cue was 75% valid, in the remaining 25% of the trials, either the uncued memory position that was located on the same object or on the different object adjacent to the cued position would be probed. They calculated the reaction time difference for correct responses between the same-object position and different-object position conditions as an indicator of object-based attention. By varying the cue-to-probe intervals, they also found the time course of object-based attention in WM oscillated in the theta range at 6 Hz.

Oscillatory dynamics might also support the activity-silent encoding of task-relevant contexts and rules (Helfrich & Knight, 2016). For example, gamma frequency might be required for driving synaptic plasticity (Harris, Csicsvari, Hirase, Dragoi, & Buzsaki, 2003; Miller, Lundqvist, & Bastos, 2018; Munk, 2016) which is considered as the neurobiological substrate of the hidden state. Beta power can disinhibit the recurrent excitation of neurons to keep gamma bursting at a lower level in the delay interval to prevent working memories from prematurely acquiring control of behavior. During working memory readout, beta allows the gamma bursting to increase so that working memories can acquire that control (Miller, Lundqvist, & Bastos, 2018). As indicated by the activity-silent model, focus of attention functions as the medium between hidden and active states. Alpha frequency has been suggested to reflect the orienting of attention (Wolff, Jochim, Akyurek, & Stokes, 2017), in this way, periodic modulation of neuronal activity could possibly function as a selection process to mediate attention which includes directing relevant information into the active state and irrelevant information into the hidden state.

Taken together, in this section, I reviewed evidence to support the notion that both neural and functional architecture of working memory is not persistent but rhythmic. Neuronal oscillation might be a general neural coding mechanism which plays critical role in working memory maintenance and various cognitive functions, such as perception (Spaak, de Lange, & Jensen, 2014; VanRullen, 2016), attention (VanRullen, 2018) and decision making (Wyart, De Gardelle, Scholl, & Summerfield, 2012). This oscillatory neural pattern is expected to manifest itself at the behavioral level if it really exists and has an effect on behavior. Here, evidence has been reviewed to support this assumption. Furthermore, coordinated neuronal activity in distinct frequencies might help to keep WM information online.

Reconciling persistent and dynamic coding of working memory

In the previous sections, two seemingly contradictory frameworks of WM maintenance were reviewed: one claims that WM maintenance needs persistent and above-baseline neural activity, and the other claims that sustained activity is not necessary, but instead, memory item can be stored in an active-silent state and modulated by rhythmic neural activities. Clearly, the dynamic framework has some advantages over the persistent activity model. It is metabolically more economic without persistent spiking, and provides a better explanation for the limitations of working memory capacity via the theta-gamma cross-frequency coupling mechanism.

This present work is not suggesting that classic model of persistent spiking is wrong, instead, it is trying to argue persistent spiking is just part of the mechanisms WM maintenance processes rely on. It is more likely both persistent activity and other complementary mechanisms work together to support WM information. For example, while short-term synaptic plasticity can support memory maintenance, persistent neuronal activity might play crucial role in manipulating information, and tasks requiring greater manipulation required greater levels of persistent activity (Masse, Yang, Song, Wang, & Freedman, 2018). The negative finding of persistent activity in the PFC might also be attributed to the topographical differences in neural activities in the subdivisions of PFC. For example, it has been suggested that neurons with persistent firing to visual features, such as faces, are more ventrally than those with persistent activity to visual space, and those with persistent firing to somatosensory information are even more ventrally in the inferior prefrontal convexity (Constantinidis et al., 2018).

A newly proposed model tried to unite persistent activity with activity-silent synaptic traces in WM (Manohar, Zokaei, Fallon, Vogels, & Husain, 2017). This model suggested that persistent activation serves as the focus of attention that encodes recent activity patterns. However, previously attended items are preserved in activity-silent synaptic traces. They are in a non-privileged state but can be reactivated by partial information.

Furthermore, this model proposed two distinct types of neural representation to account for the focus of attention: fixed feature neurons, and freely-conjunctive neurons. Fixed feature neurons can be observed in posterior cortical areas, and have fixed receptive fields or tuning curves. In contrast, freely-conjunctive neurons are located in PFC, and do not represent a fixed feature or item in memory, but, they are able to rapidly increase or decrease their synaptic connectivity with patterns of fixed feature neurons. When a stimulus is perceived, conjunctive neurons become active in response to the combination of active features, such as color, orientation, and location. Once a conjunctive unit succeeds in activating feature units, attention is focused on the activated features and binds features into a perceptual object. This mutual excitation between feature and conjunction neurons keep the combination of features persistently active, even when the stimulus is no longer present.

When a new stimulus arrives, a new pattern of sensory input could destabilize internal activity, and thus, conjunctive neurons activate again and trigger a shift of attention towards the newly activated features. Importantly, synapses between the previous objects' features and the particular conjunctive unit remained strengthened during this process even though neurons become silent. That is, previously attended items can remain in the background, encoded in activity-silent synaptic traces. Because of this mechanism, presenting any one feature of a previously attended object (e.g., color, spatial location) as a cue will reactivate the corresponding conjunction neurons, and also other features that were associated with the object, which again brings the object back to an attended state supported by persistent activity. This dual functional architecture can explain a wide range of behavioral and neurophysiological data in attention and memory which have been difficult to explain in single persistent activity framework (Manohar, Zokaei, Fallon, Vogels, & Husain, 2017).

Conclusions and future directions

Different levels of neural evidence in this review were provided to support distinct claims of different models. However, the connections between these levels (i.e., individual neurons, LFP, neuroimaging, and behavior) are still poorly understood. For example, in the oscillatory models, a wealth of EEG/MEG recordings reported oscillatory components in active WM maintenance using varied tasks. However, how EEG/MEG oscillations interact with measured dynamics in the LFPs or dynamic mental representations revealed in behavioral studies is largely unknown. This question can only be answered if one has a 'multiscale dataset' (Cohen, 2018), meaning simultaneously recorded neural spikings, LFPs, and EEG/MEG activities. We could get a better sense of WM maintenance mechanism if we could identify the one-to-one mapping between different recording techniques and behavioral measures.

In addition to the multiscale dataset, it is also essential to use appropriate analytic techniques to process these datasets. In previous paragraphs, I have emphasized the necessity of using single-trial analyses to evaluate whether the WM-related activity is persistent or sparse at the neuron level. Similar false notions about sustained activity can also be made when averaging brain rhythms acquired from EEG/MEG signals in the spectral domain. When frequency analysis is applied to a time series signal, the power representation of the signal calculated with typical time-frequency analyses is purely non-negative. Thus, when averaging across trials in the spectral domain, transient bouts of positive spectral power cannot be canceled out, which results in prolonged rhythms in duration (Jones, 2016).

A standard procedure in the study of neural rhythms is to band-pass filter the temporal domain signal into the frequency of interest. This band-pass technique usually uses a sinusoidal-like filter, which forces a sinusoidal shaped waveform of varying amplitude onto the signal. This type of analysis can make non-sinusoidal signal produce peaks in power spectra at the same frequency and power as a sinusoidal signal, even though the neural mechanism underlying non-sinusoidal signal might be fundamentally different. It has been suggested that non-sinusoidal oscillations can lead to misleading phase-amplitude coupling results and phase-phase coupling estimates (Cole & Voytek, 2018). Thus, it is necessary to develop methods to account for non-sinusoidal waveforms when performing spectral analysis.

Lastly, the ODR task and its variants have been very successful in probing the nature of WM maintenance. However, these classical tasks were still too simple. To know more about WM, we have to add more elements to our experiments. Potential candidates include requiring subjects to remember more items, introducing distractors during the delay, and varying retro-cue probability, etc. For example, as we know from the previous discussion, the focus of attention can be biased by retro-cues, and this retro-cue paradigm has contributed a lot in unveiling the mechanism of the activity-silent model. Most of the relevant studies to date used informative cue (i.e., 100% valid) to shift the focus of attention in WM. However, it is less clear how retro-cues with other reliabilities modulate the focus of attention. Would uncued items still be put into the hidden state with highly reliable but not 100% valid cue? To tackle the complexity of WM in the real world, more complex WM tasks are needed (Lundqvist, Herman, & Miller, 2018).

To sum up, several lines of evidence have been given to support competing models in this review: the dominant persistent activity as the neural correlate of WM, and its alternative models, including the activity-silent model which depends on synaptic mechanisms and oscillatory models which convey information based on the phase and frequency of discharges without relying on persistent spiking activity. The arguments against the persistent model pointed out that persistent activity can be highly variable during the course of a trial, but the averaging methods used in earlier studies obscured this dynamics of neural activity. Thus, it is best to use single trials from simultaneously recorded neurons rather than using trial-averaged data when directly evaluating both the persistent activity and dynamic coding models.

On the other hand, recent studies that indicated that WM information can still be stored without being in an active state have questioned the necessity of sustained spiking in keep information online. Furthermore, a large body of evidence suggested that instead of relying on persistent activity, it seems dynamic coding is a more generic mechanism that the brain adopts to maintain working memory information.

References

Alekseichuk, I., Turi, Z., de Lara, G. A., Antal, A., & Paulus, W. (2016). Spatial working memory in humans depends on theta and high gamma synchronization in the prefrontal cortex. *Current Biology*, *26*(12), 1513-1521.

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., & Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proceedings of the National Academy of Sciences*, 200911531.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Bahramisharif, A., Jensen, O., Jacobs, J., & Lisman, J. (2018). Serial representation of items during working memory maintenance at letter-selective cortical sites. *PLoS biology*, *16*(8), e2003805.

Barrouillet, P., & Camos, V. (2012). As time goes by: Temporal constraints in working memory. *Current Directions in Psychological Science*, *21*(6), 413-419.

Bonnefond, M., & Jensen, O. (2012). Alpha oscillations serve to protect working memory maintenance against anticipated distracters. *Current biology*, *22*(20), 1969-1974.

Cohen, M. X. (2017). Where does EEG come from and what does it mean?. *Trends in neurosciences*, *40*(4), 208-218.

Cole, S. R., & Voytek, B. (2017). Brain oscillations and the importance of waveform shape. *Trends in Cognitive Sciences*, *21*(2), 137-149.

Compte, A. (2006). Computational and in vitro studies of persistent activity: edging towards cellular and synaptic mechanisms of working memory. *Neuroscience*, *139*(1), 135-151.

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. (2018). Persistent Spiking Activity Underlies Working Memory. *Journal of Neuroscience*, *38*(32), 7020-7028.

Cowan, N. (1998). *Attention and memory: An integrated framework*. Oxford University Press.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in cognitive sciences*, *7*(9), 415-423.

Curtis, C. E., & Lee, D. (2010). Beyond working memory: the role of persistent activity in decision making. *Trends in cognitive sciences*, *14*(5), 216-222.

Druzgal, T. J., & D'esposito, M. (2003). Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *Journal of cognitive neuroscience*, *15*(6), 771-784.

Düzel, E., Penny, W. D., & Burgess, N. (2010). Brain oscillations and memory. *Current opinion in neurobiology*, *20*(2), 143-149.

Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F., & Nyberg, L. (2015). Neurocognitive architecture of working memory. *Neuron*, *88*(1), 33-46.

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron*, *99*(4), 842-853.

Fiebelkorn, I. C., Saalmann, Y. B., & Kastner, S. (2013). Rhythmic sampling within and between objects despite sustained attention at a cued location. *Current Biology*, *23*(24), 2553-2558.

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2015). The topography of alpha-band activity tracks the content of spatial working memory. *Journal of neurophysiology*, *115*(1), 168-177.

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2017). Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychological science*, *28*(7), 929-941.

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N., & Düzel, E. (2010). Theta-coupled periodic replay in working memory. *Current Biology*, *20*(7), 606-612.

Fujisawa, S., Amarasingham, A., Harrison, M. T., & Buzsaki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature neuroscience*, *11*(7), 823.

Fusi, S. (2008). A quiescent working memory. *Science*, *319*(5869), 1495-1496.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652-654.

Haegens, S., Nácher, V., Luna, R., Romo, R., & Jensen, O. (2011). α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences*, *108*(48), 19377-19382.

Hakim, N., & Vogel, E. K. (2018). Phase-coding memories in mind. *PLoS biology*, *16*(8), e3000012.

Hansel, D., & Mato, G. (2013). Short-term plasticity explains irregular persistent activity in working memory tasks. *Journal of Neuroscience*, *33*(1), 133-149.

Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G., & Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, *424*(6948), 552.

Helfrich, R. F., & Knight, R. T. (2016). Oscillatory dynamics of prefrontal cognitive control. *Trends in cognitive sciences*, *20*(12), 916-930.

Hikosaka, O., Takikawa, Y., & Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological reviews*, *80*(3), 953-978.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral cortex*, *12*(8), 877-882.

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience*, *15*(8), 1395-1399.

Jones, S. R. (2016). When brain rhythms aren't ‘rhythmic’: implication for their mechanisms and meaning. *Current opinion in neurobiology*, *40*, 72-80.

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of neurophysiology*, *34*(3), 337-347.

Landau, A. N., & Fries, P. (2012). Attention samples stimuli rhythmically. *Current biology*, *22*(11), 1000-1004.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: an EEG study. *Journal of Cognitive Neuroscience*, *25*(1), 127-142.

Leavitt, M. L., Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2017). Sustained activity encoding working memories: not fully distributed. *Trends in neurosciences*, *40*(6), 328-346.

Leung, H. C., Gore, J. C., & Goldman-Rakic, P. S. (2002). Sustained mnemonic response in the human middle frontal gyrus during on-line storage of spatial memoranda. *Journal of cognitive neuroscience*, *14*(4), 659-671.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of cognitive neuroscience*, *24*(1), 61-79.

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512-1515.

Lundqvist, M., Herman, P., & Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *Journal of cognitive neuroscience*, *23*(10), 3008-3020.

Lundqvist, M., Herman, P., & Miller, E. K. (2018). Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *Journal of Neuroscience*, *38*(32), 7013-7019.

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature communications*, *9*(1), 394.

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and beta bursts underlie working memory. *Neuron*, *90*(1), 152-164.

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T., & Husain, M. (2017). A neural model of working memory. *bioRxiv*, 233007.

Markowitz, D. A., Curtis, C. E., & Pesaran, B. (2015). Multiple component networks support working memory in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *112*(35), 11084-11089.

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2018). Circuit mechanisms for the maintenance and manipulation of information in working memory. *bioRxiv*, 305714.

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology*, *100*(3), 1407.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of neuroscience*, *16*(16), 5154-5167.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543-1546.

Munk, M. H. (2016). How to Stop Cognitive Processes is as Important as How to Start Them. Commentary on Dipoppa et al. *Advances in cognitive psychology*, *12*(4), 233.

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449-461.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411.

Oberauer, K. (2009). Design for a working memory. *Psychology of learning and motivation*, *51*, 45-100.

Pasternak, T., Lui, L. L., & Spinelli, P. M. (2015). Unilateral prefrontal lesions impair memory-guided comparisons of contralateral visual motion. *Journal of Neuroscience*, *35*(18), 7095-7105.

Pereira, J., & Wang, X. J. (2014). A tradeoff between accuracy and flexibility in a working memory circuit endowed with slow feedback mechanisms. *Cerebral Cortex*, *25*(10), 3586-3601.

Peters, B., Rahm, B., Kaiser, J., & Bledowski, C. (2018). Attention samples objects held in working memory at a theta rhythm. *bioRxiv*, 369652.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23-38.

Rademaker, R. L., & Serences, J. T. (2017). Pinging the brain to reveal hidden memories. *Nature neuroscience*, *20*(6), 767.

Raghavachari, S., Kahana, M. J., Rizzuto, D. S., Caplan, J. B., Kirschen, M. P., Bourgeois, B., ... & Lisman, J. E. (2001). Gating of human theta oscillations by a working memory task. *Journal of Neuroscience*, *21*(9), 3175-3183.

Riley, M. R., & Constantinidis, C. (2016). Role of prefrontal persistent activity in working memory. *Frontiers in systems neuroscience*, *9*, 181.

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136-1139.

Roux, F., & Uhlhaas, P. J. (2014). Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct WM information?. *Trends in cognitive sciences*, *18*(1), 16-25.

Rutishauser, U., Ross, I. B., Mamelak, A. N., & Schuman, E. M. (2010). Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature*, *464*(7290), 903.

Sauseng, P., Griesmayr, B., Freunberger, R., & Klimesch, W. (2010). Control mechanisms in working memory: a possible function of EEG theta oscillations. *Neuroscience & Biobehavioral Reviews*, *34*(7), 1015-1022.

Schneegans, S., & Bays, P. M. (2017). Restoration of fMRI decodability does not imply latent working memory states. *Journal of cognitive neuroscience*, *29*(12), 1977-1994.

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., & Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience*, *146*(3), 1082-1108.

Song, K., Meng, M., Chen, L., Zhou, K., & Luo, H. (2014). Behavioral oscillations in attention: rhythmic α pulses mediated through θ band. *Journal of Neuroscience*, *34*(14), 4837-4844.

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), R912-R913.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, *91*(3), 694-707.

Siegel, M., Warden, M. R., & Miller, E. K. (2009). Phase-dependent neuronal coding of objects in short-term memory. *Proceedings of the National Academy of Sciences*, *106*(50), 21341-21346.

Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local entrainment of alpha oscillations by visual stimuli causes cyclic modulation of perception. *Journal of Neuroscience*, *34*(10), 3536-3544.

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *The Journal of neuroscience*, *37*(27), 6503–6516. doi:10.1523/jneurosci.3364-16.2017

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89.

Stokes, M. G. (2015). 'Activity-silent'working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, *19*(7), 394-405.

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364-375.

Stokes, M., & Spaak, E. (2016). The Importance of single-trial analyses in cognitive neuroscience. *Trends in cognitive sciences*, *20*(7), 483-486.

Trübutschek, D., Marti, S., Ojeda, A., King, J. R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *Elife*, *6*, e23871.

Turi, Z., Alekseichuk, I., & Paulus, W. (2018). On ways to overcome the magical capacity limit of working memory. *PLoS biology*, *16*(4), e2005867.

van Ede, F., Quinn, A. J., Woolrich, M. W., & Nobre, A. C. (2018). Neural oscillations: sustained rhythms or transient burst-events?. *Trends in neurosciences*.

VanRullen, R. (2016). Perceptual cycles. *Trends in Cognitive Sciences*, *20*(10), 723-735.

VanRullen, R. (2018). Attention Cycles. *Neuron*, *99*(4), 632-634.

VanRullen, R., Carlson, T., & Cavanagh, P. (2007). The blinking spotlight of attention. *Proceedings of the National Academy of Sciences*, *104*(49), 19204-19209.

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, *24*(8), 455-463.

Wang, X. J. (2009). Attractor Network Models. In Squire L. R. (Ed.), *Encyclopedia of Neuroscience* (pp. 667-679). Oxford: Academic Press.

Watanabe, Y., & Funahashi, S. (2004). Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. I. Cue-, delay-, and response-period activity. *Journal of Neurophysiology*, *92*(3), 1738-1755.

Wolff, M., Ding, J., Myers, N., & Stokes, M. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*, 123. doi:10.3389/fnsys.2015.00123

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature neuroscience*, *20*(6), 864.

Wolinski, N., Cooper, N. R., Sauseng, P., & Romei, V. (2018). The speed of parietal theta frequency drives visuospatial working memory capacity. *PLoS biology*, *16*(3), e2005348.

Woloszyn, L., & Sheinberg, D. L. (2009). Neural dynamics in inferior temporal cortex during a visual working memory task. *Journal of Neuroscience*, *29*(17), 5494-5507.

Wyart, V., De Gardelle, V., Scholl, J., & Summerfield, C. (2012). Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron*, *76*(4), 847-858.

Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences*, *21*(10), 794-815.

Zhou, X., Zhu, D., Qi, X. L., Lees, C. J., Bennett, A. J., Salinas, E., ... & Constantinidis, C. (2013). Working memory performance and neural activity in prefrontal cortex of peripubertal monkeys. *Journal of neurophysiology*, *110*(11), 2648-2660.

Zucker, R. S., & Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual review of physiology*, *64*(1), 355-405.

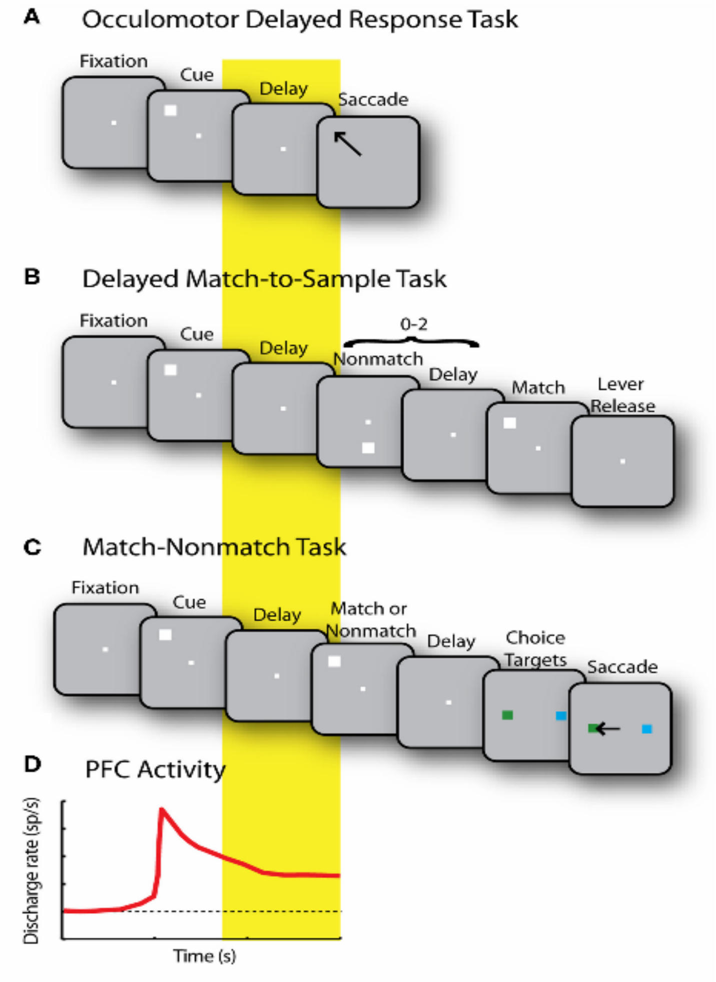

Figure 1. (A) Sequence of events in the Oculomotor Delayed Response task. Subjects are presented with a brief stimulus, and after a delay period, they need to saccade toward the remembered stimulus location. (B) Delayed Match-to-Sample task. Monkeys were firstly presented with a cue stimulus followed by a random number (0-2) of non-match stimuli, separated by delay periods. When a match stimulus appears at the same location as the cue, the monkeys are required to release the lever. (C) Match/Non-match task. Two stimuli are presented in sequence, separated by delay periods. After another delay period, two choice targets (green and blue) are shown. If the second stimulus matched the first stimulus, monkey had to saccade to the green target, otherwise, saccade to the blue target. In the last two tasks, the persistent activity elicited by the stimulus should not be confounded with motor preparation activity, since the response is not known until later in the trial. (D) A schematic diagram of sustained activity in the PFC during the delay periods (in the yellow area) of the previous tasks. Adapted from Riley and Constantinidis (2016).

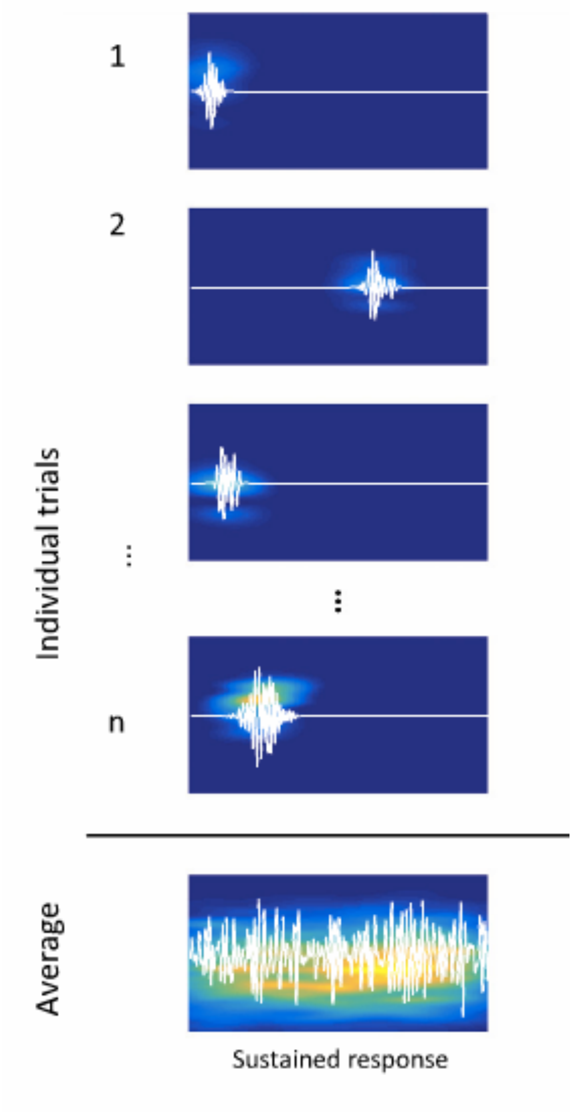

Figure 2. An illustration of trial-average results. The upper panel showed gamma bouts in the prefrontal cortex during individual trials of working memory maintenance. The bottom panel showed a sustained gamma response after averaging single-trial results. Adapted from Stokes and Spaak (2016).

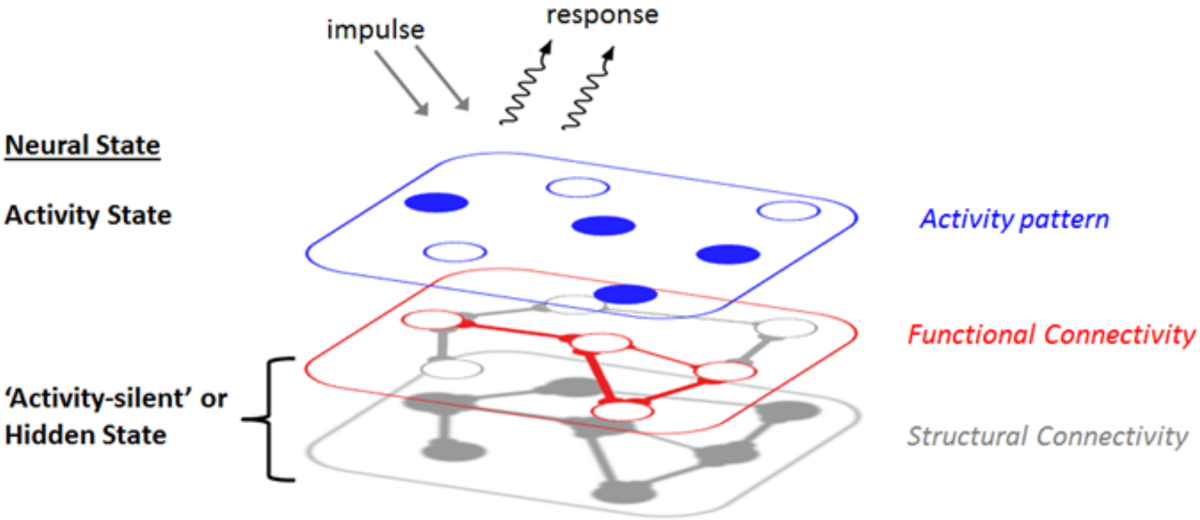

Figure 3. Schematic of layered neural states. The neural state comprises the activity state which is usually maintained with persistent activity and can be measured in typical experiments. The activity-silent state is considered as an unseen structure, which changes in effective connectivity (e.g. short-term synaptic plasticity). Though this kind of state is 'activity-silent', it can still influence subsequent processing. The information contained in this state can be probed using an impulse of activity to drive the network. Adapted from Stokes (2015).

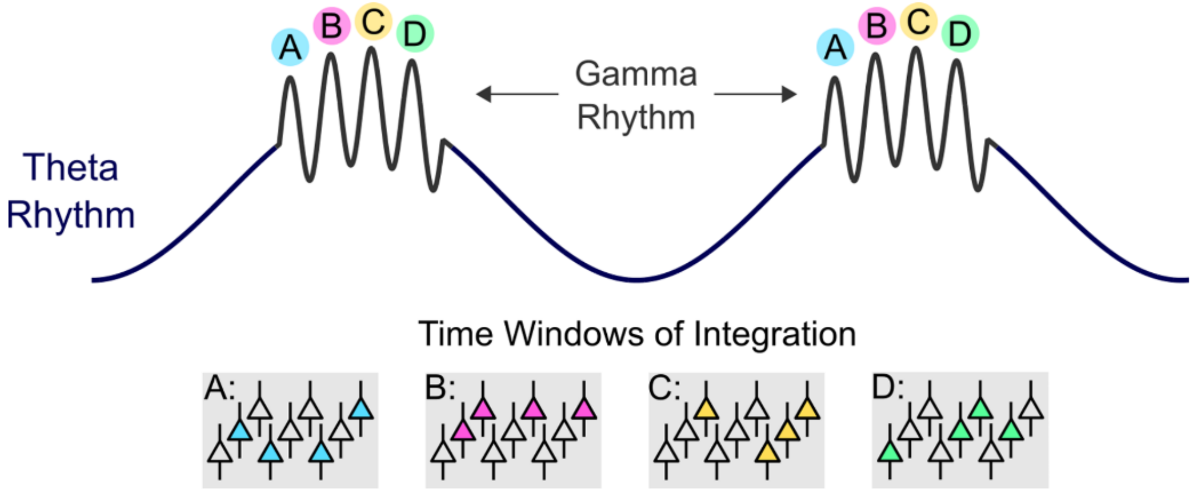

Figure 4. Theta-gamma phase-amplitude coupling scheme. Different memory items (represented as A, B, C, and D) are represented by different groups of active cells. Each gamma cycle encodes a single item, and multiple gamma cycles are nested into the theta rhythm by phase-amplitude cross-frequency coupling, thus, if multiple items are being held, the entire pattern repeats on theta cycles. Adapted from Turi, Alekseichuk, and Paulus (2018).

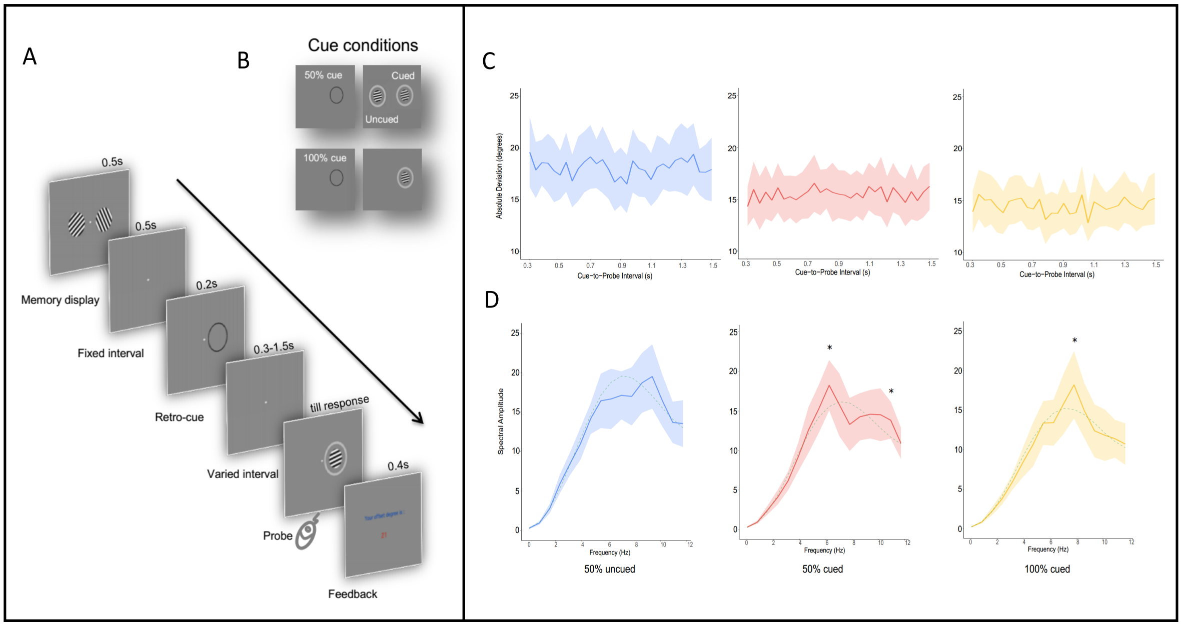

Figure 5. Left panel showed schematic of experimental design in Liu, Liu and Ravizza (in review). (A) Participants were asked to memorize the orientations of two gratings followed by a retro-cue in the form of a circle. After a varying interval from 0.3s to 1.5s, a probe grating occurred either at the cued or uncued position. Participants were required to recall the orientation of the target at the corresponding position by rotating the probe grating with a mouse to match the target orientation in their memory. (B) Retro-cue conditions and corresponding probe positions were listed. Right panel showed the time course and spectral result of each cue condition. (C) Group averaged recall performances as a function of temporal interval between the retro-cue and the probe. Raw time courses were overlaid on the 95% confidence interval bands. (D) Average spectrum for each cue condition (solid lines) was overlaid on its 95% confidence interval band. Dotted lines indicated the average permuted spectral amplitudes for each condition. Asterisks indicated the significant peak frequencies ($p < 0.05$). Adapted from Liu, Liu, and Ravizza (in review).