Dual Perspectives

Dual Perspectives Companion Paper: Persistent Spiking Activity Underlies Working Memory, by Christos Constantinidis, Shintaro Funahashi, Daeyeol Lee, John D. Murray, Xue-Lian Qi, Min Wang, and Amy F.T. Arnsten

Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not

Mikael Lundqvist,[1] Pawel Herman,[2] and Earl K. Miller[1]

[1]Picower Institute for Learning & Memory and Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and [2]Computational Brain Science Laboratory, Department Computational Science & Technology, KTH Royal Institute of Technology, Stockholm, Sweden 11428

Persistent spiking has been thought to underlie working memory (WM). However, virtually all of the evidence for this comes from studies that averaged spiking across time and across trials, which masks the details. On single trials, activity often occurs in sparse transient bursts. This has important computational and functional advantages. In addition, examination of more complex tasks reveals neural coding in WM is dynamic over the course of a trial. All this suggests that spiking is important for WM, but that its role is more complex than simply persistent spiking.

Key words: **working memory; transient dynamics; persistent activity; computational models**

Introduction

Working memory (WM) refers to our ability to volitionally hold and manipulate a limited amount of information in mind. Because of its fundamental role in goal-directed behavior and fluid intelligence, WM has been the focus of a great deal of research (Miller, 1956; Just and Carpenter, 1992; Engle et al., 1999; Miller and Cohen, 2001; Vogel and Machizawa, 2004; Chatham and Badre, 2015).

Our understanding of WM's neural basis began with the pioneering work of Fuster, Goldman-Rakic, and colleagues (Fuster and Alexander, 1971; Goldman-Rakic, 1995). They found that neurons in higher-order cortex, including and especially the PFC, show spiking during memory delays of WM tasks (Fuster and Alexander, 1971; Goldman-Rakic, 1995). Everything we know suggests that this "delay activity" is central to WM.

The question is not whether delay activity is a WM mechanism. It clearly is. The question here is how, exactly, it retains memories. The dominant model since the 1970s has been that spiking keeps an ensemble in an active state available for processing. This also seems largely true. What we and others aim to add is a small, yet important, detail: that spiking is more sparse than persistent and that, in between spikes, memories are carried by temporary changes in synaptic weights, "impressions" left in the network. In other words, the brain saves energy (spikes cost energy) by keeping ensembles in an active state with help of impressions instead of continual spiking.

This is not just energy-saving. There are functional implications. Purely persistent spiking as a memory mechanism has shortcomings. It is labile. It is easy to disrupt with additional inputs and poor at retaining more than one memory simultaneously. Having synaptic weights help carry the memories is more robust, less labile. It also adds an additional level of control. By controlling the time spent in the active state, it can prevent a given ensemble from taking command of behavior until needed and allow independent control of different items held in WM (Lundqvist et al., 2018). To be clear, we are not throwing the baby out with the bathwater. We can all agree spiking is an important WM mechanism. Rather, we are doing what one does with important models. We are offering an update.

Persistent activity: how did we get here?

A task often used, in monkeys, to argue for persistent activity is the oculomotor delayed response (ODR) task. While the monkey stares at a central dot, a visual cue is flashed somewhere in the periphery. Central fixation is maintained for a few seconds, during the so-called memory delay. Then the monkey looks in the direction of the remembered location of the cue.

ODR typically produces robust delay activity. There is, however, an issue to consider when using ODR to argue that persistent activity underlies WM. In ODR, the motor response is known when the cue appears. Thus, an action is being planned (and inhibited) over the memory delay. These premotor signals

Received March 11, 2018; revised May 17, 2018; accepted May 19, 2018.

This work was supported by National Institute of Mental Health R37MH087027, National Institute of Mental Health R01MH091174, ONR (Office of Naval Research) MURI N00014-16-1-2832, and the Massachusetts Institute of Technology Picower Institute Innovation Fund.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Earl K. Miller, Picower Institute for Learning & Memory and Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139. E-mail: ekmiller@mit.edu.

DOI:10.1523/JNEUROSCI.2485-17.2018

 0270-6474/18/387013-07$15.00/0

can, and do, contribute to delay interval spiking. However, many do not consider motor preparation to be WM. At the least, it is a special case of WM. Any proposed WM mechanism needs to be able to hold information in the absence of motor preparation because WM can operate in the absence of action planning.

As such, many investigators use tasks that do not specify the motor response until after the delay, thus removing the premotor signals. Examples are matching tasks (e.g., Fuster and Alexander, 1971; Miller et al., 1996; Romo et al., 1999). Subjects choose a stimulus that matches one seen before the memory delay. Thus, the action is not specified until after the memory delay, at the moment of choice. Delay activity under these circumstances is often less robust and less sustained (Shafi et al., 2007). Another example is a common WM test used in humans: Change Detection (e.g., Luck and Vogel, 1997; Luria et al., 2016). A scene or sequence of colored squares appears in different locations and reappears in the same locations after the memory delay. The subject indicates whether a square changed color. Importantly, this task demonstrates the defining characteristic of WM, its severely limited capacity (approximately four squares, on average). Thus, in our recent work, we used Change Detection (Lundqvist et al., 2016). But we also used ODR and WM for object sequences; all showed evidence of sparse, not persistent, activity (Lundqvist et al., 2016, 2018; Bastos et al., 2018; see below).

This evidence resulted from a different way of looking at data. Most prior studies of WM averaged spiking across trials. The assumption is that spike rate is important, but timing is not. This is a valid assumption if one studies spike rate. However, averaging across trials can create the appearance of persistent spiking even when, in real time (on individual trials), spiking is actually sparse (Lundqvist et al., 2016). Therefore, we and others (Shafi et al., 2007; Stokes and Spaak, 2016) have argued the importance of examining activity on individual trials. As we will see, this revealed sparseness, not persistence, both for single neurons and local networks; activity occurs in sparse, synchronous bursts. We contend that trial-averaged data cannot be used to demonstrate persistent activity.

But there are examples of single neurons that seem to show persistence on individual raster plots of individual trials, right? Yes, but it is important to keep in mind that many of these examples come from the ODR task (which has a motor component; see above) and single-electrode experiments. When one records from one electrode at a time, one is naturally biased toward examining the neurons that seem to have the property of interest, skewing the sampling. Plus, "example" neurons are often "best of," not a typical member of the population. Multiple-electrode recoding mitigates sampling bias because it allows recording from up to 100s of neurons simultaneously. Multiple-electrode studies have reported that the bulk of neurons spike sparsely, even when trial-averaging is used (Shafi et al., 2007; Hussar and Pasternak, 2012; Schmitt et al., 2017).

In sum, previous work has provided valuable data supporting a central role for delay activity in WM. However, much of it cannot be considered strong evidence for persistent activity per se. This is because of the use of tasks with a motor component, uneven sampling of neuron populations, and, especially, because activity was averaged across trials. Persistence versus sparseness is an issue that can only be truly resolved on individual trials.

Issues with persistent activity for WM storage

Persistent activity has been modeled in neural attractor networks where excitatory connections between units allow the activity ignited by an input to linger in a stable, self-sustaining state (Amit and Brunel, 1997; Wang, 1999; Compte et al., 2000; Renart et al., 2007; Barbieri and Brunel, 2008; Lundqvist et al., 2010; Wimmer et al., 2014). Models based on known PFC connectivity rely on fixed-point (Amit and Brunel, 1997; Compte et al., 2000; Lundqvist et al., 2010) or line attractor (Druckmann and Chklovskii, 2012) dynamics. One of these, the bump attractor model (Compte et al., 2000), made a number of detailed, successful predictions relating to error trials and the precise spike patterns observed (Wimmer et al., 2014). Overall, models of persistent activity can reproduce the irregular firing patterns observed experimentally (Renart et al., 2007; Barbieri and Brunel, 2008; Lundqvist et al., 2010).

A central idea in these models is that spiking is asynchronous: individual neurons may spike sparsely; but by doing so at different times, together they fill gaps in time over which the memories are held. There are some issues with this idea, however. First, persistent spiking is metabolically expensive. Second, in attractor dynamic models, the memory tends to be lost when activity is disrupted. A distracting sensory input may knock the dynamics out of the stable attractor state; and without any additional mechanisms to retain the memory, it cannot be recovered. Third, persistent attractor dynamics models have difficulty storing more than one item at a time. It is true that WM is capacity-limited and can only hold a few items at a time, but attractor dynamic models have difficulty, with even two items. Bump attractor models, originally proposed for visuospatial WM, have been shown to store multiple locations if there is no overlap in their neural representations and thus they are held by different networks (Almeida et al., 2015). Nearby attractors, however, tend to melt into one. This has been proposed to explain the capacity limitations of WM (Edin et al., 2009). But it is still problematic for representations that are highly overlapping as has been seen in the PFC, at least for nonspatial information (Warden and Miller, 2010; Rigotti et al., 2013). Any universal model of WM thus needs to deal with overlapping representations. Otherwise, it is only a special-case model.

Experimental observations and models of nonpersistent activity

Two major observations argue against the idea that the brain simply latches onto a sensory input and maintains it by persistent spiking per se. First, the activity of neurons during WM tasks is sparse and not asynchronous. Local networks exhibit brief coordinated bursts of WM-related activity followed by extended periods of quiescence. Second, neuron populations do not maintain whatever pattern of activity was initiated by a sensory input. The population code evolves and changes with time, additional inputs, and task demands. Below, we provide experimental support for these two ideas.

Transient activity: observations

A key question is whether neuron populations together fill the memory delay with spikes, thus providing persistent activity on the population level. The cortex operates in a regimen with approximate balance between excitation and inhibition, leading to highly irregular spiking from single neurons (Compte et al., 2003). However, cortex is thought to be organized into local recurrently connected clusters of neurons with similar response properties (e.g., Goldman-Rakic, 1995). Incorporating this in models operating in a balanced regimen can lead to neurons in each local cluster taking turns firing (i.e., asynchronously) with the cluster, as a whole, showing persistent activity (Renart et al., 2007; Barbieri and Brunel, 2008; Lundqvist et al., 2010). Thus, the question at hand is whether this actually occurs in the brain,

whether neural spiking is truly asynchronous and persistent in local clusters of neurons.

It is therefore essential to examine activity in "real time" on single trials (Shafi et al., 2007; Lundqvist et al., 2016). In past WM studies, activity has been averaged with regard to external events like a sensory input. Thus, the timing of spiking that is not time-locked to external events has not been considered. The brain has its own internal dynamics that are not time-locked to external events, and varies from trial to trial, particularly in higher-order cognition. Averaging such activity across trials can give the impression of persistent activity, even when the underlying signal is actually sparse (Lundqvist et al., 2016).

We also need to measure activity in local networks, not just single neurons. This can reveal whether local populations of neurons spike asynchronously or, instead, show sparse, coordinated, spiking. Local field potentials provide a measure of local network activity (Legatt et al., 1980; Singer and Gray, 1995). One interest is narrow band gamma oscillations (40–100 Hz) because they have been associated with sensory signals in sensory cortex (Gray and Singer, 1989; Fries et al., 2008) and the encoding (Howard et al., 2003; Sederberg et al., 2003) and maintenance (Pesaran et al., 2002; Jensen et al., 2007; Honkanen et al., 2015; Lundqvist et al., 2016; Wutz et al., 2018) of sensory information in WM. Importantly, these oscillations have also been closely associated with spiking carrying information about WM memoranda (Lundqvist et al., 2016). Thus, the gamma oscillations can be used as a measure of whether WM-related spiking is synchronous or asynchronous at the local network level. Even when averaged, it is clear that gamma oscillations are not stationary during WM retention. They wax and wane, often modulated by lower-frequency oscillations in the theta and δ bands (Canolty et al., 2006; Axmacher et al., 2010). This suggests that the associated spiking is sparse and periodic, not asynchronous.

Single-trial analyses reveal even more sparseness (Lundqvist et al., 2016, 2018; Kucewicz et al., 2017; Bastos et al., 2018). There are brief bursts of elevated gamma surrounded by periods of relative silence. The gamma bursts are narrow-band and thus not simply a reflection of the spike waveforms. They reflect coordinated activity in local networks. These gamma episodes co-occur with increased spiking and stimulus information in spiking (Lundqvist et al., 2016, 2018; Bastos et al., 2018). Thus, although it may appear so with trial averaging, WM-related activity is not asynchronous and persistent over a delay. Local networks of neurons instead show coordinated bursts of activity that are transient and sparse. Indeed, we re-created the appearance of persistent activity by averaging across trials, even though the underlying activity was anything but (Lundqvist et al., 2016). We have found sparse WM-related spiking/gamma-bursting using a variety of WM tasks: ODR (Lundqvist et al., 2016; Bastos et al., 2018), Change Detection (Lundqvist et al., 2016), and WM for object sequences (Lundqvist et al., 2018).

The rate of these transient gamma bursts/spikes is correlated with WM functions, such as encoding of information, its readout, clearing out of WM, and switching the contents of WM (Lundqvist et al., 2018). For example, spiking tends to "ramp up" toward the end of WM delays (Chafee and Goldman-Rakic, 1998; Shafi et al., 2007; Watanabe and Funahashi, 2007; Barak et al., 2010; Warden and Miller, 2010). This has been interpreted as a "turning up of the volume" of persistent activity. But others have argued that it indicates that delay activity is more of a readout or preparatory, rather than memory storage mechanism (Stokes, 2015). We have shown that this ramp up is mediated by an increase in the rate of sparse, transient, bursts of activity on single trials, not by a gradual increase in WM-related activity. Related observations have been made in studies of decision-making (Latimer et al., 2015). By measuring deviations from these burst rates, we could predict forthcoming errors in more detail than from spiking alone (Lundqvist et al., 2018).

Thus, WM maintenance seems to involve sparse, transient coordinated activations of local neurons rather than asynchronous persistent activity. This is consistent with models where "packets" of information are reactivated pseudo-randomly (Mongillo et al., 2008; Lundqvist et al., 2011). Indeed, neural information about different items in WM phase-lock to different phases of slower oscillations (Siegel et al., 2009; Bahramisharif et al., 2017). In other words, different items are multiplexed in time, as if the brain were juggling them, at odds with persistent activity models.

We have focused on the activity of local networks. We argue that this is the critical level because much of the brain's computations takes place on a local level and the cortex is thought to be organized into local clusters with shared tuning properties (Kritzer and Goldman-Rakic, 1995; Constantinidis et al., 2001). One might argue, however, that, even though activity is not locally persistent, it may be on a more global scale. In other words, transient local activity could be counterbalanced by activity in other parts of a larger network so that global activity is persistent.

This is, of course, hard to completely rule out. There is, however, evidence against this argument. Even when activity is averaged over trials, it is apparent that seemingly persistent spiking is labile and easily disrupted. Distracting animals during a memory delay by having them attend to task-irrelevant stimuli abolished WM-related spiking (Watanabe and Funahashi, 2014; Spaak et al., 2017). Yet, when attention returned to the WM (i.e., during their readout), elevated spiking returned. Noninvasive EEG recordings of global activity have revealed that, for extended periods of time, information held in WM cannot be decoded. However, when the area is "pinged" by a task-irrelevant stimulus, the network "rings" back with the WM information (Stokes et al., 2013; Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2017), suggesting that sustained activity is not necessary to maintain the memories.

Thus, while delay activity spiking does seem to play a role in WM maintenance, its inability to truly bridge gaps in time suggests that other mechanisms of storing information are also at play.

Transient activity: models

Models of transient dynamics explain the sparseness in spiking. WMs are maintained between sparse episodes of spiking by spike-induced changes in synaptic plasticity (Sandberg et al., 2003; Mongillo et al., 2008; Lundqvist et al., 2011) or cellular mechanisms (Lisman and Idiart, 1995). In other words, in these "synaptic attractor" models, spikes leave an "impression" in the networks that maintains WM information between spiking.

Not only is this metabolically less expensive, these models are better equipped to handle multiple items in WM because overlapping representations can be multiplexed in time (taking turns being active or silent). The limitation in WM capacity is explained by limitations in replay time (Lundqvist et al., 2011; Mi et al., 2017) as items need to be reactivated to "refresh" the decaying synaptic changes. Increased rate of gamma-bursting with memory load is consistent with this (Lundqvist et al., 2016). In a model by Lisman and Idiart (1995), instead, each WM item is held in a different gamma cycle. The number of gamma cycles is limited by a cycle of an underlying theta rhythm. In contrast to most WM models, both these models account for network oscillations in addition to spiking and thus offer explanations for why gamma-band power increases with WM load (Howard et al., 2003).

Synaptic attractor models offer other benefits. Because memories are not exclusively held in spiking, synaptic attractor models are resistant to disruption by additional sensory inputs. The time multiplexing of different WM items aids in information readout because each ensemble is dynamically separated in time. This avoids the problem of reading out information from mixed activity from the superposition of multiple items, which may be problematic when novel combinations of items are considered. These models have not yet explicitly incorporated dynamic coding, the change in population coding of WM content.

Dynamic coding: observations

Another argument against a straightforward persistence model is that WM-related activity evolves over time. In the strong model of persistent activity, a sensory input is maintained by the same ensembles that were activated by the input. However, during WM tasks, neural population codes change over time as well as with task demands (Barak et al., 2010; Stokes et al., 2013; Cavanagh et al., 2017; Spaak et al., 2017).

There are two neural behaviors that have been linked to an evolving population code. First, information is carried by a sequence of brief activations of single neurons so that one neuron activates another with similar tuning properties forming a cascade of activity (Shafi et al., 2007; Barak et al., 2010; Cromer et al., 2010; Harvey et al., 2012; Hussar and Pasternak, 2012). Second, individual neurons can change their tuning. For example, a neuron that responds to a certain location or stimulus during sensory input may instead respond more strongly in the delay to other stimuli/locations. Experimental evidence shows that both these mechanisms contribute to dynamic changes in population codes (Barak et al., 2010; Cavanagh et al., 2017; Parthasarathy et al., 2017; Spaak et al., 2017).

Changes in a population code are evaluated by training decoders on population spiking. The simultaneous spiking of all neurons is used as input, typically combining neurons across multiple recording sessions (Stokes et al., 2013). If a decoder trained on data at one time point does not perform well at another time point, there has been a change in the population code. This revealed that the population code can quickly change after the initial sensory input is removed (Stokes et al., 2013; Cavanagh et al., 2017; Murray et al., 2017; Parthasarathy et al., 2017; Spaak et al., 2017). During the middle of a memory delay, the population code tends to settle into a relatively stable state and decoders generalize significantly across neighboring time points. However, overall decoding performance (information) is often relatively low. Multivariate analysis of human MEG activity revealed a similar pattern, with a slowly, yet continuously evolving, code during a 4 s delay (Trubutschek et al., 2017).

Despite population code changes, it is possible to find a linear combination of neurons that will then maintain a stable code during memory delay (Murray et al., 2017). This has been deemed a "stable subspace." It has been taken as evidence for line attractors congruent with persistent activity hidden in the overall heterogeneous neural dynamics. In other words, there can be stable readout from delay activity from subsets of population activity.

However, as noted above, a major challenge for models of persistent activity is its lack of compatibility with distractors or storage of multiple items. The stable subspaces have been demonstrated with "empty" memory delays without additional inputs. WM in the real world, in contrast, involves potential distractions as well as encoding of additional items. These additional inputs and demands result in drastic changes in the population code, even in the middle of a WM delay. Classifiers trained on time points before the additional inputs or distractors do not perform well on time points following it (Cavanagh et al., 2017; Parthasarathy et al., 2017). This change in coding is consistent with mixed selectivity (Warden and Miller, 2010; Rigotti et al., 2013) where individual neurons are sensitive to the combination of multiple behavioral conditions and items.

This means that WM needs to also be studied in richer experimental paradigms. Much of the evidence for stable coding comes from simple tasks that only require memory for a single item over a blank delay. While simple tasks would favor the persistent activity model, the model cannot explain all aspects of the observed activity during richer tasks. Dynamic coding seems to dominate during more complex tasks with multiple stimuli. A comprehensive model has to account for these more complex situations more like those encountered in the real world.

Dynamic coding: models

Models that use dynamic coding include random, recurrent networks relying on "chaotic" dynamics. These networks do not maintain a fixed activity in response to inputs. Instead, activity evolves chaotically following inputs (Barak et al., 2013). Chaotic activity is not random but deterministic, dependent on starting conditions. This allows a trace of past events that have perturbed the system to persist, even if the initial state has vanished. This can lead to neurons that change tuning over time, as observed experimentally (Barak et al., 2010; Cavanagh et al., 2017; Spaak et al., 2017).

The evolution of activity in chaotic networks poses a challenge for readout of information. Downstream decoding neurons would have to dynamically adjust readout over time. This problem can be mitigated for a well-defined memory task with a consistent structure so that the decoder can adapt in a predictable fashion. But relying on trained classifiers for WM readout may still be problematic. It would require retraining of the decoder if novel items or new combinations of familiar items are stored. Humans and other animals have no problem-solving WM tasks with novel items.

In addition to models relying on chaotic activity, synfire chain networks have been proposed (Prut et al., 1998; Goldman, 2009; Rajan et al., 2016). In these models, activity is transferred from different subpopulations that share tuning. This results in a chain of activity, unique to each stimulus held in WM. Synfire chains have been used to explain observations that neurons fire over a small portion of a delay (Cromer et al., 2010; Hussar and Pasternak, 2012; Schmitt et al., 2017; Lundqvist et al., 2018) and the occurrence of repeating, precise temporal firing patterns (Prut et al., 1998). However, individual neurons that change preference over time (which has been observed), or after an additional stimulus, is not a straightforward prediction of these models. In short, experimental observations of dynamic coding can be partly reproduced but pose a challenge for many existing WM models.

Principles for further work

Several experimental observations have given support to various competing models (Wimmer et al., 2014; Lundqvist et al., 2016; Bahramisharif et al., 2017; Murray et al., 2017; Schmitt et al., 2017; Trubutschek et al., 2017). In our view, no model has yet explained all the neurophysiological observations. Clearly, more work is needed. We here aim to suggest some guidelines for future work on WM. They can provide the necessary data needed to distinguish between different models and whether or not WM depends on persistent activity.

1. *Assessment of network dynamics*. Different models of WM posit either asynchronous activity (persistent activity) or

brief, coordinated, transient bursts of activity (transient dynamics). Other models posit shifts in population code, including different chains of neurons activating for different memoranda. One cannot observe these dynamics studying one neuron at a time. Our ability to record an increasing number of neurons simultaneously will be critical. Multielectrode recordings will allow this assessment, especially when used in conjunction with recordings of local field potentials whose activity provides an index of the dynamics of the networks in which the neurons are embedded.

2. *Single-trial analyses*. Different models make distinct claims about the details of the dynamics of neural activity. One cannot assess them by averaging activity across trials and different recording sessions. This obscures the very dynamics needed to test the models (Stokes and Spaak, 2016). Do not get us wrong — averaging across trials is useful and necessary for many questions. But one cannot make claims, for example, about whether WM-related activity is persistent or sparse by only using trial-averaged data. Averaging can create the appearance of persistent activity if there is none. Similarly, due to cofluctuations, dynamic coding is best evaluated on single trials from simultaneously recorded neurons, rather than combining neurons across recording sessions.
3. *Complex WM tasks*. In the classic WM test, there is a single memorandum and a "blank" memory delay of ≥1 s. This provided us with a wealth of information about the fundamentals of WM physiology. The pioneers who introduced and used these tests have been rightly lauded for their groundbreaking work. However, inevitably, the more we learn about any phenomenon, the more complex it turns out to be. We know now that requiring that multiple items be held in WM, the requirement to ignore distractions, etc. has a big effect on neural activity. The classic tests were never meant to approach the complexity of WM in the real world. But they have taught us enough so that we now add more real-world elements to our experimental efforts.
4. *Taking neural rhythms into account*. Some models assume asynchronous spiking or rate coding. However, the brain oscillates across a wide range of frequencies. A large number of experimental observations have shown that these oscillations are modulated during WM and, in turn, modulate spiking. This poses important constraints to WM models. Models should ideally provide a comprehensive explanation of the neurophysiology associated with WM, including associated oscillatory dynamics.

In conclusion, past work has revealed that spiking activity in the absence of sensory stimuli plays a critical role in WM. Nothing we have said here changes that. Instead, new analytical techniques, observations, and models have given us additional insights. What naturally follows is an updating of our existing models.

Response From Dual Perspectives Companion Author–Christos Constantinidis

Lundqvist et al. argue that persistent activity cannot maintain information and that gamma-band bursting is the critical neural correlate in the delay period of WM tasks, instead. We present the opposing view in the companion article but wish to respond here to four arguments that are particularly problematic.

1. **First, Lundqvist et al. (2018) claim that robust persistent activity is generated only in the ODR task, which confounds stimulus presentation with motor preparation. They cherry-pick the literature and ignore every report of persistent activity in tasks that dissociate the two factors, such as match-nonmatch tasks. Studies sampling hundreds of neurons in an unbiased fashion reveal robust persistent activity, qualitatively similar with that generated in the ODR task (Meyer et al., 2011; Mendoza-Halliday et al., 2014).**
2. **Lundqvist et al. (2018) suggest that gamma-band bursting is advantageous because it is sparse and metabolically inexpensive. First, modulation in the local field potential during the delay period does not imply that the underlying spiking activity is "sparse." Even if that were true, Lundqvist et al. argue here that gamma-band bursting ought to be the neural correlate of WM but offer little evidence that it actually is. The rate of gamma-bursting in the delay period was not predictive of correct or erroneous recall in any WM task tested by the authors. In those instances where error data are presented, gamma-band bursting in the delay period is indistinguishable in correct and error trials (Lundqvist et al., 2018). Counterevidence exists showing that gamma power is not predictive of WM performance (Ma et al., 2018).**
3. **Lundqvist et al. (2018) also suggest that attractor models that stimulate the generation of persistent activity have difficulty representing multiple items stored in memory or filtering distractors. This is an argument against simple, recurrent-network models rather than the role of persistent activity per se, but Lundqvist et al. miss how successful such models have been, precisely because they link persistent activity with behavioral performance.**
4. **Finally, Lundqvist et al. (2018) claim that studies successfully decoding stimulus information from delay period firing rate are flawed because they rely on neurons recorded asynchronously. Persistent activity has been documented in multielectrode recordings, and simultaneous firing can encode information equally well (Leavitt et al., 2017). Lundqvist et al. also fail to provide evidence that information can be decoded from gamma-bursting and that this measure is superior to firing rate.**

References

Leavitt ML, Pieper F, Sachs AJ, Martinez-Trujillo JC (2017) Correlated variability modifies working memory fidelity in primate prefrontal neuronal ensembles. Proc Natl Acad Sci U S A 114: E2494-E2503. CrossRef Medline

Lundqvist M, Herman P, Warden MR, Brincat SL, Miller EK (2018) Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nat Commun 9:394. CrossRef Medline

Ma L, Skoblenick K, Johnston K, Everling S (2018) Ketamine alters lateral prefrontal oscillations in a rule-based working memory task. J Neurosci. Advance online publication. Retrieved Feb. 2, 2018. doi: 10.1523/JNEUROSCI.2659–17.2018. CrossRef Medline

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

Meyer T, Qi XL, Stanford TR, Constantinidis C (2011) Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. J Neurosci 31:6266–6276. CrossRef Medline

References

Almeida R, Barbosa J, Compte A (2015) Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. J Neurophysiol 114:1806–1818. CrossRef Medline

Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb Cortex 7:237–252. CrossRef Medline

Axmacher N, Henseler MM, Jensen O, Weinreich I, Elger CE, Fell J (2010) Cross-frequency coupling supports multi-item working memory in the human hippocampus. Proc Natl Acad Sci U S A 107:3228–3233. CrossRef Medline

Bahramisharif A, Jensen O, Jacobs J, Lisman J (2017) Serial representation of items during working memory maintenance at letter-selective cortical sites. bioRxiv. Advance online publication. Retrieved August 2, 2017. doi:10.1101/171660.

Barak O, Tsodyks M, Romo R (2010) Neuronal population coding of parametric working memory. J Neurosci 30:9424–9430. CrossRef Medline

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

Barbieri F, Brunel N (2008) Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? Front Neurosci 2:3. CrossRef Medline

Bastos AM, Loonis R, Kornblith S, Lundqvist M, Miller EK (2018) Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. Proc Natl Acad Sci U S A 115:1117–1122. CrossRef Medline

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight RT (2006) High gamma power is phase-locked to theta oscillations in human neocortex. Science 313:1626–1628. CrossRef Medline

Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW (2017) Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. bioRxiv. Advance online publication. Retrieved December 14, 2017. doi:10.1101/231506.

Chafee MV, Goldman-Rakic PS (1998) Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memorytask. J Neurophysiol 79:2919–2940. CrossRef Medline

Chatham CH, Badre D (2015) Multiple gates on working memory. Curr Opin Behav Sci 1:23–31. CrossRef Medline

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb Cortex 10:910–923. CrossRef Medline

Compte A, Constantinidis C, Tegner J, Raghavachari S, Chafee MV, Goldman-Rakic PS, Wang XJ (2003) Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J Neurophysiol 90:3441–3454. CrossRef Medline

Constantinidis C, Franowicz MN, Goldman-Rakic PS (2001) Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. J Neurosci 21:3646–3655. CrossRef Medline

Cromer JA, Roy JE, Miller EK (2010) Representation of multiple, independent categories in the primate prefrontal cortex. Neuron 66:796–807. CrossRef Medline

Druckmann S, Chklovskii DB (2012) Neuronal circuits underlying persistent representations despite time varying activity. Curr Biol 22:2095–2103. CrossRef Medline

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A (2009) Mechanism for top-down control of working memory capacity. Proc Natl Acad Sci U S A 106:6802–6807. CrossRef Medline

Engle RW, Tuholski SW, Laughlin JE, Conway AR (1999) Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. J Exp Psychol Gen 128:309–331. CrossRef Medline

Fries P, Womelsdorf T, Oostenveld R, Desimone R (2008) The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. J Neurosci 28:4823–4835. CrossRef Medline

Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173:652–654. CrossRef Medline

Goldman MS (2009) Memory without feedback in a neural network. Neuron 61:621–634. CrossRef Medline

Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14:477–485. CrossRef Medline

Gray CM, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci U S A 86:1698–1702. CrossRef Medline

Harvey CD, Coen P, Tank DW (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484:62–68. CrossRef Medline

Honkanen R, Rouhinen S, Wang SH, Palva JM, Palva S (2015) Gamma oscillations underlie the maintenance of feature-specific information and the contents of visual working memory. Cereb Cortex 25:3788–3801. CrossRef Medline

Howard MW, Rizzuto DS, Caplan JB, Madsen JR, Lisman J, Aschenbrenner-Scheibe R, Schulze-Bonhage A, Kahana MJ (2003) Gamma oscillations correlate with working memory load in humans. Cereb Cortex 13:1369–1374. CrossRef Medline

Hussar CR, Pasternak T (2012) Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. J Neurosci 32:2747–2761. CrossRef Medline

Jensen O, Kaiser J, Lachaux JP (2007) Human gamma-frequency oscillations associated with attention and memory. Trends Neurosci 30:317–324. CrossRef Medline

Just MA, Carpenter PA (1992) A capacity theory of comprehension: individual differences in working memory. Psychol Rev 99:122–149. CrossRef Medline

Kritzer MF, Goldman-Rakic PS (1995) Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. J Comp Neurol 359:131–143. CrossRef Medline

Kucewicz MT, Berry BM, Kremen V, Brinkmann BH, Sperling MR, Jobst BC, Gross RE, Lega B, Sheth SA, Stein JM, Das SR, Gorniak R, Stead SM, Rizzuto DS, Kahana MJ, Worrell GA (2017) Dissecting gamma frequency activity during human memory processing. Brain 140:1337–1350. CrossRef Medline

Latimer KW, Yates JL, Meister ML, Huk AC, Pillow JW (2015) Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science 349:184–187. CrossRef Medline

Legatt AD, Arezzo J, Vaughan Jr HG (1980) Averaged multiple unit activity as an estimate of phasic changes in local neuronal activity: effects of volume-conducted potentials. Journal of neuroscience methods 2:203–217. CrossRef Medline

Lisman JE, Idiart MA (1995) Storage of 7 plus/minus 2 short-term memories in oscillatory subcycles. Science 267:1512–1515. CrossRef Medline

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281. CrossRef Medline

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. J Cogn Neurosci 23:3008–3020. CrossRef Medline

Lundqvist M, Rose J, Herman P, Brincat SL, Buschman TJ, Miller EK (2016) Gamma and beta bursts underlie working memory. Neuron 90:152–164. CrossRef Medline

Lundqvist M, Herman P, Warden MR, Brincat SL, Miller EK (2018) Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nat Commun 9:394. CrossRef Medline

Lundqvist M, Compte A, Lansner A (2010) Bistable, irregular firing and population oscillations in a modular attractor memory network. PLoS Comput Biol 6:e1000803. CrossRef Medline

Luria R, Balaban H, Awh E, Vogel EK (2016) The contralateral delay activity as a neural measure of visual working memory. Neurosci Biobehav Rev 62:100–108. CrossRef Medline

Mi Y, Katkov M, Tsodyks M (2017) Synaptic correlates of working memory capacity. Neuron 93:323–330. CrossRef Medline

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202. CrossRef Medline

Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J Neurosci 16:5154–5167. CrossRef Medline

Miller GA (1956) The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97. Medline

Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319:1543–1546. CrossRef Medline

Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang XJ (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. CrossRef Medline

Parthasarathy A, Herikstad R, Bong JH, Medina FS, Libedinsky C, Yen SC (2017) Mixed selectivity morphs population codes in prefrontal cortex. Nature neuroscience 20:1770–1779. CrossRef Medline

Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen RA (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat Neurosci 5:805–811. CrossRef Medline

Prut Y, Vaadia E, Bergman H, Haalman I, Slovin H, Abeles M (1998) Spatiotemporal structure of cortical activity: properties and behavioral relevance. J Neurophysiol 79:2857–2874. CrossRef Medline

Rajan K, Harvey CD, Tank DW (2016) Recurrent network models of sequence generation and memory. Neuron 90:128–142. CrossRef Medline

Renart A, Moreno-Bote R, Wang XJ, Parga N (2007) Mean-driven and fluctuation-driven persistent activity in recurrent networks. Neural computation 19:1–46. CrossRef Medline

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585–590. CrossRef Medline

Romo R, Brody CD, Hernández A, Lemus L (1999) Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399:470–473. CrossRef Medline

Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR (2016) Reactivation of latent working memories with transcranial magnetic stimulation. Science 354:1136–1139. CrossRef Medline

Sandberg A, Tegnér J, Lansner A (2003) A working memory model based on fast Hebbian learning. Network 14:789–802. CrossRef Medline

Schmitt LI, Wimmer RD, Nakajima M, Happ M, Mofakham S, Halassa MM (2017) Thalamic amplification of cortical connectivity sustains attentional control. Nature 545:219–223. CrossRef Medline

Sederberg PB, Kahana MJ, Howard MW, Donner EJ, Madsen JR (2003) Theta and gamma oscillations during encoding predict subsequent recall. J Neurosci 23:10809–10814. CrossRef Medline

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 146:1082–1108. CrossRef Medline

Siegel M, Warden MR, Miller EK (2009) Phase-dependent neuronal coding of objects in short-term memory. Proc Natl Acad Sci U S A 106:21341–21346. CrossRef Medline

Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annual review of neuroscience 18:555–586. CrossRef Medline

Spaak E, Watanabe K, Funahashi S, Stokes MG (2017) Stable and dynamic coding for working memory in primate prefrontal cortex. J Neurosci 37:6503–6516. CrossRef Medline

Sprague TC, Ester EF, Serences JT (2016) Restoring latent visual working memory representations in human cortex. Neuron 91:694–707. CrossRef Medline

Stokes M, Spaak E (2016) The importance of single-trial analyses in cognitive neuroscience. Trends Cogn Sci 20:483–486. CrossRef Medline

Stokes MG (2015) 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci 19:394–405. CrossRef Medline

Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic coding for cognitive control in prefrontal cortex. Neuron 78:364–375. CrossRef Medline

Trübutschek D, Marti S, Ojeda A, King JR, Mi Y, Tsodyks M, Dehaene S (2017) A theory of working memory without consciousness or sustained activity. Elife 6:e23871. CrossRef Medline

Vogel EK, Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. Nature 428:748–751. CrossRef Medline

Wang XJ (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. J Neurosci 19:9587–9603. CrossRef Medline

Warden MR, Miller EK (2010) Task-dependent changes in short-term memory in the prefrontal cortex. J Neurosci 30:15801–15810. CrossRef Medline

Watanabe K, Funahashi S (2014) Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nat Neurosci 17:601–611. CrossRef Medline

Watanabe K, Funahashi S (2007) Prefrontal delay-period activity reflects the decision process of a saccade direction during a free-choice ODR task. Cereb Cortex 17 [Suppl. 1]:i88–i100.

Wimmer K, Nykamp DQ, Constantinidis C, Compte A (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci 17:431–439. CrossRef Medline

Wolff MJ, Jochim J, Akyürek EG, Stokes MG (2017) Dynamic hidden states underlying working-memory-guided behavior. Nat Neurosci 20:864–871. CrossRef Medline

Wutz A, Loonis R, Roy JE, Donoghue JA, Miller EK (2018) Different levels of category abstraction by different dynamics in different prefrontal areas. Neuron 97:1–11. CrossRef Medline