A functional spiking-neuron model of activity-silent working memory in humans
based on calcium-mediated short-term synaptic plasticity

Matthijs Pals¹, Terrence C. Stewart², Elkan G. Akyürek³, & Jelmer P. Borst¹*

¹Bernoulli Institute, University of Groningen, Groningen, The Netherlands.
²Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, Ontario, Canada.
³Department of Experimental Psychology, University of Groningen, Groningen, The Netherlands.

*Corresponding author:
E-mail: j.p.borst@rug.nl (JPB)
Abstract

In this paper, we present a functional spiking-neuron model of human working memory (WM). This model combines neural firing for encoding of information with activity-silent maintenance. While it used to be widely assumed that information in WM is maintained through persistent recurrent activity, recent studies have shown that information can be maintained without persistent firing; instead, information can be stored in activity-silent states. A candidate mechanism underlying this type of storage is short-term synaptic plasticity (STSP), by which the strength of connections between neurons rapidly changes to encode new information. To demonstrate that STSP can lead to functional behavior, we integrated STSP by means of calcium-mediated synaptic facilitation in a large-scale spiking-neuron model. The model simulated a recent study that measured behavior and EEG activity of participants in a delayed-response task. In this task, a visual grating had to be maintained in WM, and compared to a subsequent probe. It was demonstrated that WM contents could be decoded from the neural activity elicited by a task-irrelevant stimulus that was displayed during the activity-silent maintenance period. In support of our model, we show that it can perform this task, and that both its behavior as well as its neural representations correspond to the human data. We conclude that information in WM can be effectively maintained in activity-silent states by means of calcium-mediated STSP.

Author Summary

Mentally maintaining information for short periods of time in working memory is crucial for human adaptive behavior. It was recently shown that the human brain does not only store information through neural firing – as was widely believed – but also maintains information in activity-silent states. Here, we present a detailed neural model of how this could happen in our brain through short-term synaptic plasticity: rapidly adapting the connection strengths between neurons in response to incoming information. By reactivating the adapted network, the stored information can be read out later. We show that our model can perform a working memory task as accurate as human participants.
can, while using similar mental representations. We conclude that our model is a plausible and effective neural implementation of human working memory.
Introduction

The ability to temporarily hold information in working memory (WM) is a crucial part of day-to-day life: it is what allows us to remember someone’s name at a cocktail party, what ingredients to buy at the supermarket for dinner, and which platform we need to go to when changing trains [1,2].

The maintenance of information in WM is often studied by means of a delayed-response task, in which a briefly presented memory item is followed by a delay period [3,4]. The delay period ends with the presentation of a probe that the participants need to compare to the memorized item. The maintenance of information during the delay period of such tasks was long thought to be mediated by continuously spiking neurons [5,6]. Although neural spiking is certainly important for WM, it was recently shown that spiking activity during delay periods can be intermittent or even absent [7–11]. This suggests that information may be stored instead using activity-silent mechanisms, for instance through transient connectivity patterns in the brain [2,12,13]. The spiking activity observed previously might reflect the initial phase necessary to initialize new synaptic weights, active maintenance of the focus of attention [14–17], or the read-out of information from working memory [13,18].

One of the candidate mechanisms for storing information in activity-silent states is short-term synaptic plasticity [STSP;19], which entails rapid changes in the strength of connections between neurons to reflect new information being presented to the network [12]. Indeed, it was previously shown that synapses in areas implicated in WM can be facilitated [20,21], potentially as a consequence of residual calcium building up in presynaptic terminals [19,22]. In this way residual calcium effectively leaves a ‘synaptic trace’ of what is currently stored in WM. An elegant implementation of activity-silent storage by means of STSP was proposed by Mongillo and colleagues [12], who developed a model that can maintain information through calcium-mediated synaptic facilitation in recurrent networks of simulated spiking neurons. In response to a particular input to the network, a subset of the neurons fires, with the result that their outgoing connections are facilitated. Subsequently, stored information can be read out by applying a network-wide non-specific input that will be mostly subthreshold for non-facilitated neurons but leads to firing of facilitated neurons.
In the current study, we show that the mechanism proposed by Mongillo and colleagues [12] not only results in efficient and robust storage, but also in effective, functional human behavior. We integrated their calcium-mediated STSP mechanism in a large-scale spiking-neuron model that can perform a delayed-response task. To evaluate this model, we used a previously reported electroencephalography (EEG) dataset of a visual WM task, in which activity-silent memory states were measured [3]. To this end, Wolff and colleagues developed an innovative method to probe activity-silent brain states [3,23]. They showed that when the WM network is perturbed by a high-contrast task-neutral stimulus during maintenance, ensuing neural activity reveals what is currently held in an activity-silent state.

In their experiment [3], each trial started with the display of two randomly oriented gratings (Fig 1). After an 800 ms fixation period, this was followed by a cue indicating which of the two stimuli had to be maintained in memory. In order to examine the contents of WM during the subsequent delay part of the trial, an impulse stimulus was presented 900 ms later. At the end of each trial, participants had to indicate whether a probe stimulus was rotated clockwise or counter-clockwise with respect to the cued memory item. To track the contents of WM, a decoding analysis was applied to the EEG data [3]. It was shown that decoding accuracy quickly dropped to chance level after presentation of the memory items, but returned when the probe was presented. This indicates that between the presentation of the memory items and the probe, information is maintained in an activity-silent (or at least quiescent) state. In addition, it was shown that it is possible to decode the orientation of the cued memory item from the EEG data in response to the impulse stimulus. Thus, when the WM
network was perturbed by a task-neutral stimulus, the ensuing signal allowed for decoding of the current contents of the activity-silent state. Interestingly, already after the presentation of the cue, only the orientation of the cued memory item could be decoded, indicating that the uncued stimulus was quickly forgotten, or actively cleared from memory.

To test our model, we let it perform the same experiment – including the application of the impulse perturbation method – and compared both our model’s performance as well as its mental representations and underlying spiking behavior to the human data.

Results

Model Architecture

In order to implement a functional spiking-neuron model of WM we used Nengo, a framework for building large-scale brain models that link single cell activity to demonstrative cognitive abilities [24–26]. In this framework, information is represented by vectors of real numbers that can be encoded and decoded from the collective spiking activity of a population of neurons. Connections between neural populations allow for both communication and transformation of information. Here, Nengo acts as a ‘neural compiler’: given a desired function, the connection weight matrix between populations is calculated so that this function is approximated. Besides pre-calculating connection weights, plasticity can be introduced by making use of biologically plausible learning rules [27].

To account for short-term synaptic plasticity, we integrated the calcium kinetics mechanism proposed by Mongillo and colleagues [12] in the model. Accordingly, synaptic efficiency between two neurons is dependent on two parameters: the amount of available resources to the presynaptic neuron (reflecting neurotransmitters) and the presynaptic calcium level. Each time a neuron fires, the amount of available resources decreases, reducing synaptic efficiency. As resources are quickly replenished (in the order of 200 ms), this results in short-term depression of firing rates. However, at the same time calcium flows into the presynaptic terminals, increasing synaptic efficiency. Because
calcium is much slower to return to its baseline levels than the resources, the synaptic connection is facilitated in the long-term, for about 1.5 seconds. This STSP mechanism was applied to the recurrent connections of two working memory populations in our model. As described above, the aim is to simulate a dataset of a delayed-response task in which the orientation of two memory items has to be compared to a probe ([3]; Fig 1). In this task, significant EEG lateralization was observed at posterior electrodes after presentation of the cue. We therefore hypothesized that distinct populations of neurons are responsible for processing visual stimuli presented in the left and right visual field. Correspondingly, the model was divided into two independent modules, each responsible for perceiving and representing one of the two incoming stimuli (Fig. 2).

In order to demonstrate that our model is able to deal with real-world input, the stimuli from [3] were presented to the model. The sensory populations use two-dimensional Gabor filters as encoders [25,28,29]. As a result, the information present in the gratings – including their direction – is encoded into 24-dimensional vectors that are passed on to the memory populations. That is, the information encoded into the neurons is a compressed representation of the input image, using the top 24 singular values as per SVD (see Methods for more details). The memory populations contain recurrent connections exhibiting STSP, in line with previous models of WM and anatomical areas.
associated with WM [e.g., 21,30,31]. Consequently, the first stimulus during a trial will drive facilitation of recurrent connections representing this stimulus. Neural activity resulting from subsequent stimuli will be affected by this change in connectivity.

This enables implementing decision making as a match-filter process [8,13]. To decide on the orientation change of the probe compared to the relevant memory item, both the sensory and memory populations communicate the orientation of the gratings to a comparison population. When a probe is presented, the orientation received from the sensory population is driven entirely by the incoming stimulus. However, the orientation of the memory population is driven by a dynamic combination of activity resulting from the incoming stimulus and activity from facilitated connections as a result of the encoded memory item (i.e. hysteresis). In other words, the orientation represented by the memory population reflects the orientation of the probe ‘tuned’ by the orientation of the memory item stored in facilitated synapses, over time reverting to the new probe stimulus. To estimate the orientation difference between the memory item and the probe, the outgoing connections from the comparison layer subtract the two represented orientations. The resulting one-dimensional value reflects the signed difference between the orientation of the memory item and that of the probe stimulus.

Neural representations

The model simulated the experiment reported in [3] and illustrated in Fig 1. In the original paper, it was shown that decoding accuracy quickly dropped after presentation of the memory items but returned again during presentation of the probe – as would be expected for an activity-silent maintenance mechanism. Correspondingly, we examined the spiking activity and quality of representations of our model during the task, in order to validate that any maintenance of information in our model is realized in activity-silent states and not by persistent firing.

Fig 3 shows the spiking activity of the neurons in the memory populations of both modules during one trial (A: cued module, B: uncued module), together with the mean amount of resources ($\alpha$) and calcium ($\mu$) in these populations. In both modules, there is spiking activity during and shortly after presentation of the memory items, the impulse stimulus, and the probe, but not in between. The
spiking activity causes the amount of available resources and the calcium level to decrease and
increase, respectively. The resulting short-term depression can directly be observed as the amount of
spiking declines after the onset of a stimulus, although it periodically reactivates.

In the original experiment, a retro-cue that indicated which of the two previously presented
items needed to be memorized was presented 800 ms later, which was followed by significant
lateralization at posterior electrodes. To mimic this, the memory population of the cued module is briefly reactivated by means of a non-specific population wide input [cf. 12]. This not only re-activates the memory item, but also helps to maintain the stimulus for a longer time period, as reactivation of facilitated synapses will lead to re-facilitation of those connections.

Next, we analyzed the vectors represented by the memory populations of both the cued and uncued module. Fig 3 (bottom of each panel) shows the absolute cosine similarity between the vector represented by the memory populations and the ideal vectors of potential representations. To clearly illustrate the difference between the two modules, the mean cosine similarity was calculated over 100 trials in which both modules were presented with the same memory item and probe, with a rotation of 0° and 42°, respectively. Note that in the simulation of the real experiment, the cued and uncued modules are never presented with the same memory item.

During presentation of the initial memory item of 0°, the vectors represented by both modules are very similar to the ideal 0° vector. In addition, the cosine similarity is inversely correlated with the angular difference between the represented vector of 0° and potential representations, indicating that similar stimuli are represented by similar vectors and firing patterns. As was the case in the original experiment, during the delay periods we could not decode what is being represented by the neural populations as there is no spiking activity – indicating activity-silent memory. However, in response to the non-specific reactivation of the cued model at 1050 ms, there was spiking activity that clearly represents the originally encoded vector. It therefore appears that neural connections representing the memory item were indeed facilitated, and that mainly those connections and neurons get activated in response to the non-specific reactivation elicited by the cue.

One of the main results of the original study was that the EEG activity in response to the impulse stimulus contained the orientation of the cued memory item, and not of the uncued item [3]. This was taken to show that a stimulus is only maintained in an activity-silent state if it is still needed for the task. If not, it is quickly forgotten or actively cleared from the network. To see if our model has both the same storage and forgetting capabilities, we examined the vectors represented by the memory populations of the cued and uncued module during presentation of the impulse (Fig 4; cf. Fig
A SPIKING-NEURON MODEL OF ACTIVITY-SILENT WORKING MEMORY

In both modules, the memory populations start representing the impulse stimulus. When the facilitated recurrent connections of the cued and uncued items become activated, both modules also represent the original 0° memory item. However, only for the cued module does the represented vector become (very briefly) more similar to the ideal memory item than to the impulse vector, offering a potential explanation of why only the cued, and not the uncued memory item, could be decoded after the impulse [3].

To summarize: in both the cued and the uncued modules, STSP encodes the initial stimulus. In the cued model, facilitated connections are re-facilitated at the moment of cue, counteracting the gradual calcium decay that goes on in both modules (Fig 3). As a result, once the impulse arrives, only the cued model has sufficiently facilitated connections specific to the memory item to generate a response larger than the impulse representation (Fig 4). Note that the uncued memory population was not actively cleared, but that the calcium levels of the facilitated synapses simply decayed away as it was not reactivated at the moment of the cue.
Behavior

In order to see if our model not only matches neural activity, but also gives rise to functional behavior similar to human participants, we evaluated its performance. First, to see if the information maintained in the facilitated synapses can be used to produce a relevant response, we inspected the value represented by the decision population in the cued module. This population receives the angular difference between the memory item and the probe from the comparison population, and thus represents a measure of difference between the orientations decoded from the sensory and memory populations. Fig 5 shows the represented value for the possible orientation differences between memory items and probes, averaged over all simulated trials. First, it takes a moment for the probe information to reach this population. Second, the facilitated synapses become activated, reactivating the memory item, thereby leading to different representations in the sensory and memory populations, and thus to a difference in the decision population. Finally, the probe starts overriding the memory representation, reducing the difference until both populations represent the probe and the difference has disappeared. Overall, both the sign and magnitude of the orientation difference are clearly represented in the decision signal.

To translate this decision signal into a response, we integrated the decision activation after the presentation of the probe. Integrating the evidence corresponding to two distinct decisions has been
widely used before in accumulator models of perceptual decision making [e.g., 32]. We did not model motor processes, but simply interpreted a positive result as a clockwise response and a negative result as a counter-clockwise response. Fig 6 shows that the model’s proportion of clockwise responses across orientation differences follows a similar S-shape as the human responses.

**Discussion**

We developed a functional spiking-neuron model to explain recent theories of activity-silent human working memory. Whereas incoming information is encoded through spiking, maintenance of information was realized by short-term synaptic plasticity based on calcium kinetics [12]. This mechanism can maintain information effectively for short periods of time without requiring neural spiking. In support of the model, we simulated a recent EEG study that applied an innovative impulse perturbation method [3] to reveal the content of activity-silent WM. Both the model’s choice behavior, as well as its mental representations corresponded well to the human data. This demonstrates that calcium-mediated STSP not just results in robust maintenance of arbitrary stimuli, as shown earlier [12], but can also simulate effective human behavior based on real-world stimuli.

With regard to localization, the model was used to simulate data from Wolff and colleagues [3], who reported posterior EEG effects. However, WM is often attributed to prefrontal areas [e.g., 14,46]. Activity-silent maintenance has likewise been found in both posterior [3,8,47] and frontal
[18,44] regions. It appears that especially sensory working memory should be attributed to the relevant sensory systems [47,48], instead of to a centralized system. While the exact function of the different regions implicated in WM might differ, the neural substrate and mechanisms might be similar, and could potentially all be explained by the proposed STSP mechanism.

A number of design choices warrant discussion. First, the employed neurons do not have a baseline firing rate, as is evident by the lack of any spiking activity during the delay-period of a trial (Fig 3). In order to clearly demonstrate activity-silent maintenance of information, we defined the tuning curves of the sensory and memory neurons so that they only fire when presented with input. However, background firing could simply be added to the model without affecting the functioning of our model, as has been done in the past ([e.g., 12,33]; see also below). Second, the number of neurons per population and the number of dimensions used to represent the stimuli were set to reflect human behavior. In general, adding more neurons will improve the representation of vectors and the approximation of the functions computed over those. Increasing the number of dimensions expands what can represented [25,34,35]. Thus, changing the number of neurons and dimensions will change the quality of the representations and will influence the number of errors made during the task. Here, we estimated parameters to roughly match human performance; we do not have a principled reason either for using 1000 or 1500 neurons per population or 24 dimensions. Finally, information was represented using Nengo’s default vector representation, which provides an intuitive method to link neural spiking to representation and function [25]. However, representing information differently should not affect the basic functioning of the model as all connections and the STSP mechanism are implemented at the neural level.

**Representations in WM**

As discussed above, in the current model information is maintained without any intermittent firing (Fig 3). This directly contradicts the original analysis of the dataset [3], where the represented stimulus could be decoded for some time after its offset. In addition to full activity-silent maintenance, Mongillo and colleagues [12] observed a bi-stable regime in their model: with added
A spiking-neuron model of activity-silent working memory

Background noise, neurons with facilitated connections reactivate spontaneously. Consequently, due to the dynamics of $u$ and $x$, the reactivated neurons will be briefly depressed before being facilitated again, leading again to reactivation. In this regime, the time between subsequent reactivations is on the scale of $\tau_D$ – the time constant of the available resources – as it is controlled by the recovery from the synaptic depression. A brief exploratory analysis shows that such a bi-stable regime can also be added to our model, as is illustrated in Fig 7. This provides the model with an additional method of maintaining information, possibly over a longer period of time. It also indicates a potential explanation for the delay-activity observed in the original analysis [3]: non-specific background or recurrent input after presentation of a stimulus might temporarily have pushed the network into this regime.

A functional role for delay activity in WM might be tracking the focus of attention [15–17,36]. Wolff and colleagues [3] performed a second experiment, similar to the one discussed in this paper. Again, two memory items were presented at the start of the trial. However, in this version both were eventually probed, and participants were aware of which of the two items would be probed first. Before the first probe, the prioritized item could be much better and longer decoded than the un-prioritized item, a pattern that reversed after the first probe. This could be simulated with our model by giving the prioritized item a persistent firing state as in Figure 7, while the un-prioritized item
might be completely maintained in an activity-silent state as in Figure 3. This also provides a functional explanation for psychological theories that state that a single focal WM item can be used without any time cost [16,37–39], while other items in working memory incur a cost estimated at 200 ms [15,40–43] – the latter potentially being due to the costs of reactivating the information from a non-active state.

In our model, we have assumed that the coding of the information itself is static, that is, the same facilitated connections are used repeatedly. However, there has recently been increasing evidence for a dynamic coding framework, which states that information maintained in a WM network traverses a highly dynamic path through neural activation [44]. It is not yet clear how this relates to our model, although a possible clue might come from a model by Singh and Eliasmith [45]. Neural populations in their model represent two dimensions, where one dimension represents time and the other a stimulus. Their model elegantly captures a wide variety of observed neural responses during a WM task; the inclusion of time as a dimension represented by the neurons in the network naturally leads to a dynamic firing pattern over time.

Related Models

Recently, Myers and colleagues [8] described a related non-spiking neural population model with similar functionality as the current model, although they did not match human data directly. Their model consisted of a three-layer architecture: a stimulus layer, a template layer, and a decision layer not unlike the sensory, memory, and comparison population in our model. A critical difference between the two models is that their decision layer only receives input from the template layer, while in our case it receives input from both the sensory and memory populations. The template layer in Myers’ model acts like a match-filter: it is able to maintain a stimulus orientation, and when presented with a subsequent probe orientation convey the signed difference between the two to the decision layer. The memory population in our model can likewise be viewed as a match filter. After onset of the probe, the represented orientation shifts to the orientation of the probe from the direction of the orientation of the memory item. This shift in itself indicates a degree of difference between the two
orientations, including the sign of this difference. One could potentially measure this with a neural population that computes a time derivative with respect to the orientation [49]. However, exploratory analysis indicated this to be less robust than our current method.

Another closely related model was proposed by Barak and colleagues [18]. Their model consisted of a sensory and memory population. After presentation of a stimulus, connections from the sensory population to the memory population will be facilitated. Subsequently, during the delay period, an increasing uniform current is applied to the network which activates the neurons in the memory population that have facilitated incoming connections. During presentation of a subsequent probe, mutual inhibition between the sensory and memory population will guide decision making. This model explains observed ramping up of activity during anticipation of a probe. However, it is not clear whether the gradually increasing external current is essential to extract the information maintained in the facilitated connections in the memory representation. It can be expected that in the brain bottom-up stimulus driven activity might also be able to activate the information stored in connections, for instance when the timing of the probe is unknown.

**Conclusion**

To conclude, our model shows that maintenance of information in WM by means of calcium-mediated STSP can lead to functional behavior. It is broadly consistent with current theories regarding activity-silent storage in human WM and is able to show a variety of effects observed during a visual delayed-response task. Furthermore, it provides a solid basis for exploring a model that incorporates psychological theories on the focus of attention [15–17,37] by combining activity-silent maintenance with storage through persistent firing.

**Methods**

**Model**

*Nengo*. The model was implemented using Nengo, a Python library for simulating large-scale neural models with a clear link between spiking activity and representation [24–26]. Nengo makes use
of a theoretical framework called the Neural Engineering Framework [NEF; 50]. Information is
represented as a vector of real numbers that can be encoded and decoded from the collective spiking
activity of populations of neurons. Encoding is mediated by giving each neuron a non-linear tuning
curve that characterizes their general response to the incoming signal. Decoding is a linear process:
the activity of each neuron in a population is weighted by a constant and summed over time in order
to decode the represented vector. Connections between populations allow for the communication and
transformation of the information. Here the NEF calculates the connection weight matrix between
populations to approximate a desired function. In addition, connection weights can be learned and
adapted through several biologically plausible learning rules, both supervised and unsupervised [27].

**Short-term synaptic plasticity.** Short-term synaptic plasticity was implemented in Nengo
following the calcium kinetics mechanism of Mongillo and colleagues ([12]; available at
https://github.com/Matthijspals/STSP). Because spiking leaky integrate-and-fire (LIF) neurons are
computationally efficient while retaining a degree of biological plausibility, we added this mechanism
to the existing Nengo implementation of LIF neurons. Synaptic efficiency is based on two parameters:
the amount of available resources to the presynaptic neuron ($x$, normalised between 0 and 1) and the
fraction of resources used each time a neuron fires ($u$), reflecting the residual presynaptic calcium
level. When a neuron fires, its resources $x$ are decreased by $ux$, mimicking neurotransmitter depletion.
At the same time, its calcium level $u$ is increased, mimicking calcium influx into the presynaptic
terminal. Both $u$ and $x$ relax back to baseline with time constants $\tau_D$ (0.2s) and $\tau_F$ (1.5s), respectively.
This results in a system where after a neuron fires its outgoing connections will be depressed on the
time scale of $\tau_D$ and facilitated on the timescale of $\tau_F$ as illustrated in Fig 8.

For all LIF neurons to which we apply STSP, every time step $u$ and $x$ are calculated according
to equation 1.1 and 1.2, respectively:

\[
\frac{dx}{dt} = \frac{1-x}{\tau_D} - ux \delta(t-t_{sp}) \tag{1.1}
\]

\[
\frac{du}{dt} = \frac{U-u}{\tau_F} + U(1-u) \delta(t-t_{sp}) \tag{1.2}
\]
where $x$ represents the available resources, $u$ the residual calcium level, $\tau_D$ the depressing time constant, $\delta$ the Dirac delta function, $t$ the simulation time and $t_{sp}$ the time of a presynaptic spike. In equation 1.2, $\tau_F$ represents the facilitating time constant and $U$ the calcium baseline level. Outgoing connection weights of neurons implementing STSP are determined by both their initial connection weight and their current synaptic efficiency. Initial connections weights are calculated by the NEF, while synaptic efficiency is set to the product of the current value of $u$ and $x$ of the presynaptic neuron, normalised by their baseline value:

$$w_{ij} = \frac{xu}{U} w_{ij}^0$$  \hspace{1cm} (1.3)

where $w_{ij}$ represents the connection weight between neuron $i$ and $j$ and $w_{ij}^0$ the initial connection weight.

**Architecture.** The overall architecture of the model is shown in Fig 2 (the model is available for download at https://github.com/Matthijspals/STSP). The sensory and decision populations consist of 1000 LIF neurons, the memory and comparison populations of 1500 LIF neurons. Biologically
relevant parameters were left to default, which are consistent with neocortical pyramidal cells [24].

Parameters $U$, $t_D$ and $t_F$ were set the same as in [12]. $t_F >> t_D$ and $t_F$ on the order of 1s are consistent with patch-clamp recordings of facilitated excitatory connections in the ferret prefrontal cortex [21].

To describe the relationship between neural representations and real-world stimuli it can be assumed that the brain makes use of a statistical model, not unlikely a parametrized model, where a small number of parameters capture the overall shape of the data [25]. To find such a model we need a set of basis functions that will be good at describing both the incoming images and the encoders of the neurons receiving these images. These basis functions can be found by applying singular value decomposition (SVD) to a matrix containing both the images and the encoders. The images consisted of the stimuli in the experiment, while the encoders were two-dimensional Gabor filters, defined by a sinusoidal plane wave multiplied by a Gaussian function. Gabor filters have previously been shown to accurately describe the response profile of simple cells in the cat [29] and macaque [28] striate cortex and seem to underlie early stages of visual processing. Thus, the SVD mediates a biologically plausible method that results in stimuli being represented by 24-dimensional vectors.

Next, we specified for each possible vector – representing a particular grating – what the corresponding decoded orientation should be. This information was used to define a function that takes a 24-dimensional vector as input and returns the corresponding orientation. The NEF yields the connection matrix at the neural level that approximates this function for the connections from the sensory and memory populations to the comparison populations. Stimulus orientations were not directly decoded as the angle $\theta$, but rather by the sine and cosine of $\theta$. Decoding sine and cosine of $\theta$ is robust, as the ratio between the two determines the stimulus orientation independent of the amplitude, which is not the case when decoding $\theta$ directly. Furthermore, the symmetry of the sine functions provides a natural solution for the symmetric nature of the stimuli, as in the experiment a stimulus with an orientation of -90° contains exactly the same pixels as a stimulus with an orientation of 90° and therefore results in the same neural activity.
**Experimental Simulation**

**Stimuli.** Input to the model consisted of images of 128 by 128 pixels. Stimuli were generated using Psychopy, an open-source Python application [51]. Stimuli consisted of a circle on a grey background (RGB = 128, 128, 128). Memory items and probe stimuli were sine-wave gratings with a diameter of 128 pixels and spatial frequency of 0.034 cycles per pixel. The phase was randomized within and across trial. For each trial, the orientation of the memory items was randomly selected from a uniform distribution of orientations. The angular differences between the memory item and the corresponding probe stimulus were uniformly distributed across seven angle differences (3°, 7°, 12°, 18°, 25°, 33°, 42°), both clockwise and counter-clockwise. The impulse stimulus consisted of a ‘bull’s-eye’ stimulus of the same size and spatial frequency as the memory items. It was presented at twice the contrast compared to the grating stimuli, to each module.

**Procedure.** The model completed the retro-cue delayed-response task from [3]. Each trial started with the presentation of two memory items to the sensory population of the corresponding modules for 250 ms. In the original experiment, a retro-cue that indicated which of the two previously presented items needed to be memorized was presented 800 ms later, which was followed by significant lateralization at posterior electrodes. To mimic this, the memory population of the cued module is briefly reactivated by means of a non-specific population wide input [cf. 12]. After another fixation period, the impulse stimulus was presented to both sensory populations for 100 ms, 1100 ms after the onset of the cue. After another delay of 400 ms, the probe was presented to the sensory populations for 250 ms. To simulate different participants in the experiment, every 1344 trials the random seed was reset and new random Gabor filters were generated to use as encoders for the sensory populations. In total the model performed 30 sets of 1344 trials, reflecting 30 participants in the original experiment.

**References**

1. Baddeley AD, Hitch GJ. Working Memory. In: Bower GA, editor. The psychology of learning and motivation. New York: Academic Press; 1974. pp. 47–89. doi:10.4324/9781912282418
2. Barak O, Tsodyks M. Working models of working memory. Curr Opin Neurobiol. 2014;25:20–24. doi:10.1016/j.conb.2013.10.008

3. Wolff MJ, Jochim J, Akyürek EG, Stokes MG. Dynamic hidden states underlying working-memory-guided behavior. Nat Neurosci. 2017;20: 864–871. doi:10.1038/nn.4546

4. Rombouts JO, Bohte SM, Roelfsema PR. How attention can create synaptic tags for the learning of working memories in sequential tasks. PLoS Comput Biol. 2015;11: e1004060. doi:10.1371/journal.pcbi.1004060

5. Fuster JMJ, Alexander GE. Neuron Activity Related to Short-Term Memory. Science (80- ). 1971;173: 652–654.

6. Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995;14: 477–485. doi:10.1016/0896-6273(95)90304-6

7. Lundqvist M, Herman P, Miller EK. Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. J Neurosci. 2018;38: 7013–7019. doi:10.1523/JNEUROSCI.2485-17.2018

8. Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes MG, et al. Testing sensory evidence against mnemonic templates. Elife. 2015;4: 1–25. doi:10.7554/eLife.09000

9. Sreenivasan KK, Curtis CE, D’Esposito M. Revisiting the role of persistent neural activity during working memory. Trends Cogn Sci. 2014;18: 82–89. doi:10.1016/j.tics.2013.12.001

10. Watanabe K, Funahashi S. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nat Neurosci. 2014;17: 601–611. doi:10.1038/nn.3667

11. Lundqvist M, Rose J, Herman P, Brincat SL, Buschman TJ, Miller EK. Gamma and Beta Bursts Underlie Working Memory. Neuron. 2016;90: 152–164. doi:10.1016/j.neuron.2016.02.028

12. Mongillo G, Barak O, Tsodyks M. Synaptic Theory of Working Memory. Science (80- ). 2008;319: 1543–1546. doi:10.1126/science.1150769

13. Stokes MG. ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci. 2015;19: 394–405. doi:10.1016/j.tics.2015.05.004
A SPIKING-NEURON MODEL OF ACTIVITY-SILENT WORKING MEMORY

14. Borst JP, Anderson JR. Using Model-Based functional MRI to locate Working Memory Updates and Declarative Memory Retrievals in the Fronto-Parietal Network. Proc Natl Acad Sci USA. 2013;110: 1628–1633. doi:10.1073/pnas.1221572110

15. Borst JP, Taatgen NA, Van Rijn H. The Problem State: A Cognitive Bottleneck in Multitasking. J Exp Psychol Learn Mem Cogn. 2010;36: 363–382. doi:10.1037/a0018106

16. Oberauer K. Access to information in working memory: Exploring the focus of attention. J Exp Psychol Learn Mem Cogn. 2002;28: 411–421.

17. Olivers CNL, Peters J, Houtkamp R, Roelfsema PR. Different states in visual working memory: when it guides attention and when it does not. Trends Cogn Sci. 2011;15: 327–334. doi:10.1016/j.tics.2011.05.004

18. Barak O, Tsodyks M, Romo R. Neuronal population coding of parametric working memory. J Neurosci. 2010;30: 9424–9430. doi:10.1523/JNEUROSCI.1875-10.2010

19. Zucker RS, Regehr WG. Short-Term Synaptic Plasticity. Annu Rev Physiol. 2002;64: 355–405. doi:10.1146/annurev.physiol.64.092501.114547

20. Tsodyks M, Pawelzik K, Markram H. Neural Networks with Dynamic Synapses. Neural Comput. 1998; 821–835.

21. Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat Neurosci. 2006;9: 534–542. doi:10.1038/nn1670

22. Jackman SL, Regehr WG. The Mechanisms and Functions of Synaptic Facilitation. Neuron. 2017;94: 447–464. doi:10.1016/j.neuron.2017.02.047

23. Wolff MJ, Ding J, Myers NE, Stokes MG. Revealing hidden states in visual working memory using electroencephalography. Front Syst Neurosci. 2015;9: 1–12. doi:10.3389/fnsys.2015.00123

24. Bekolay T, Bergstra J, Hunsberger E, DeWolf T, Stewart TC, Rasmussen D, et al. Nengo: a Python tool for building large-scale functional brain models. Front Neuroinform. 2014;7: 48. doi:10.3389/fninf.2013.00048
A SPIKING-NEURON MODEL OF ACTIVITY-SILENT WORKING MEMORY

25. Eliasmith C. How to Build a Brain: A Neural Architecture for Biological Cognition. New York, NY: Oxford University Press; 2013.

26. Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, Tang C, et al. A large-scale model of the functioning brain. Science (80- ). 2012;338: 1202–1205. doi:10.1126/science.1225266

27. Bekolay T, Kolbeck C, Eliasmith C. Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks. Cognitive Science Society; 2013. pp. 169–174.

28. Ringach DL. Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex. J Neurophysiol. 2002;88: 455–463. doi:10.1152/jn.2002.88.1.455

29. Jones JP, Palmer LA. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. J Neurophysiol. 1987;58: 1233–1258.

30. Camperi M, Wang X-J. A model of visuospatial short-term memory in prefrontal cortex: recurrent network and cellular bistability. J Comput Neurosci. 1998;5: 383–405.

31. Miller P, Brody CD, Romo R, Wang XJ. A Recurrent Network Model of Somatosensory Parametric Working Memory in the Prefrontal Cortex. Cereb Cortex. 2003;13: 1208–1218. doi:10.1093/cercor/bhg101

32. Palmeri TJ, Schall JD, Logan GD. Neurocognitive Modeling of Perceptual Decision Making. In: Busemeyer JR, Wang Z, Townsend JT, Eidels A, editors. The Oxford Handbook of Computational and Mathematical Psychology. New York, NY: Oxford University Press; 2015. pp. 320–340.

33. Stewart TC, Bekolay T, Eliasmith C. Learning to Select Actions with Spiking Neurons in the Basal Ganglia. Front Neurosci. 2012;6. doi:10.3389/fnins.2012.00002

34. Rasmussen D, Eliasmith C. A spiking neural model applied to the study of human performance and cognitive decline on Raven’s Advanced Progressive Matrices. Intelligence. 2014;42: 53–82. doi:10.1016/j.intell.2013.10.003

35. Kajić I, Gosmann J, Stewart TC, Wennekers T, Eliasmith C. A Spiking Neuron Model of Word Associations for the Remote Associates Test. Front Psychol. 2017;8: 48.
A SPIKING-NEURON MODEL OF ACTIVITY-SILENT WORKING MEMORY

36. Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR. Neural evidence for a distinction between short-term memory and the focus of attention. J Cogn Neurosci. 2012;24: 61–79. doi:10.1162/jocn_a_00140

37. Oberauer K. Design for a working memory. In: Ross BH, editor. Psychology of Learning and Motivation. Academic Press; 2009. pp. 45–100.

38. Cowan N. Attention and memory: An integrated framework. New York: Oxford University Press; 1995.

39. McElree B. Working memory and focal attention. J Exp Psychol Learn Mem Cogn. 2001;27: 817–835.

40. Anderson JR. How Can the Human Mind Occur in the Physical Universe? New York: Oxford University Press; 2007.

41. Nijboer M, Borst JP, Van Rijn H, Taatgen NA. Contrasting Single and Multi-Component Working-Memory Systems in Dual Tasking. Cogn Psychol. 2016;86: 1–26. doi:10.1016/j.cogpsych.2016.01.003

42. Borst JP, Taatgen NA, Stocco A, Van Rijn H. The Neural Correlates of Problem States: Testing fMRI Predictions of a Computational Model of Multitasking. PLoS One. 2010;5: e12966. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0012966

43. Borst JP, Taatgen NA, Van Rijn H. What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. Proc CHI. Seoul, Korea: ACM Press; 2015.

44. Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J. Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron. 2013;78: 364–375. doi:10.1016/j.neuron.2013.01.039

45. Singh R, Eliasmith C. Higher-dimensional neurons explain the tuning and dynamics of working memory cells. J Neurosci. 2006;26: 3667–3678. doi:10.1523/JNEUROSCI.4864-05.2006

46. Fuster J. The prefrontal cortex. 5th ed. New York, NY: Elsevier; 2015.
47. Sreenivasan KK, Vytlacil J, D’Esposito M. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. J Cogn Neurosci. 2014;26: 1141–1153. doi:10.1162/jocn_a_00556

48. Pasternak T, Greenlee MW. Working memory in primate sensory systems. Nat Rev Neurosci. 2005;6: 97–107. doi:10.1038/nrn1603

49. Tripp BP, Eliasmith C. Population models of temporal differentiation. Neural Comput. 2010;22: 621–659. doi:10.1162/neco.2009.02-09-970

50. Eliasmith C, Anderson CH. Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. Cambridge, MA: The MIT Press; 2002.

51. Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, et al. PsychoPy2: Experiments in behavior made easy. Behav Res Methods. 2019;51: 195–203. doi:10.3758/s13428-018-01193-y