Towards a neural-level cognitive architecture: modeling behavior in working memory tasks with neurons

Zoran Tiganj (zorant@bu.edu)
Nathanael Cruzado (nac0005@bu.edu)
Marc W. Howard (marc777@bu.edu)
Center for Memory and Brain
Boston University

Abstract

Constrained by results from classic behavioral experiments we provide a neural-level cognitive architecture for modeling behavior in working memory tasks. We propose a canonical microcircuit that can be used as a building block for working memory, decision making and cognitive control. The controller controls gates to route the flow of information between the working memory and the evidence accumulator and sets parameters of the circuits. We show that this type of cognitive architecture can account for results in behavioral experiments such as judgment of recency, probe recognition and delayed-match-to-sample. In addition, the neural dynamics generated by the cognitive architecture provides a good match with neurophysiological data from rodents and monkeys. For instance, it generates cells tuned to a particular amount of elapsed time (time cells), to a particular position in space (place cells) and to a particular amount of accumulated evidence.

Keywords: Cognitive architecture; Neural-level modeling; Working memory; Cognitive control; Decision making; Judgment of recency; Probe recognition; Delayed-match-to-sample

Introduction

Behavioral experiments provide important insights into human memory and decision making. Building neural systems that can describe these processes is essential for our understanding of cognition.

Here we propose a neural-level architecture that can model behavior in different working memory based cognitive tasks. The proposed architecture is composed of biologically plausible artificial neurons characterized with instantaneous firing rate and with the ability to: 1) gate information from one set of neurons to the other (Hasselmo & Stern, 2018; Bhandari & Badre, 2018; Sherfey, Ardid, Miller, Hasselmo, & Kopell, 2019) and 2) modulate the firing rate of other neurons via gain modulation (Salinias & Sejnowski, 2001). The architecture is based on a canonical microcircuit that represents continuous variables via supported dimensions (Shankar & Howard, 2012; Howard et al., 2014). The microcircuit is implemented as a two-layer neural network. The same microcircuit prototype is used for maintaining a compressed memory timeline, evidence accumulation and for controlling the flow of actions in a behavioral task. Here we demonstrate that this architecture can be used for modeling behavioral responses and neural activity in a variety of working memory tasks.

A neural architecture for cognitive modeling

We sketch a neural cognitive architecture and apply it to three distinct working memory tasks. The architecture is composed of multiple instances of a canonical microcircuit (Figure 1). This microcircuit represents vector-valued functions over variables. These functions can be examined through attentional gain field and then used to produce a vector-valued output. We first discuss the properties of the microcircuit.

Function representation in the Laplace domain

The microcircuit consists of two layers. The first layer approximates the Laplace transform of $\mathbf{f}(t)$ (a vector across the input space) via set of neurons which can be described as leaky integrators $\mathbf{F}(t,s)$, with a spectrum of rate constants s. Each neuron in $\mathbf{F}(t,s)$ receives the input and has a unique rate constant:

$$\frac{d\mathbf{F}(t,s)}{dt} = \alpha(t)\left[-s\mathbf{F}(t,s)+\mathbf{f}(t)\right], \qquad (1)$$

where $\alpha(t)$ is an external signal that modulates the dynamics of the leaky integrators. If $\alpha(t)$ is constant, $\mathbf{F}(t,s)$ codes the Laplace transform of $\mathbf{f}(t)$ leading up to the present. It can be shown that if $\alpha(t) = dx/dt$, $\mathbf{F}(t,s)$ is the Laplace transform with respect to x (Howard et al., 2014). We assume that the probability of observing a neuron with rate constant s goes down like $1/s$. This implements a logarithmic compression of the function representation.

The second layer $\tilde{\mathbf{f}}(t,\overset{*}{x})$ computes the inverse of the Laplace transform using the Post approximation. It is implemented as a linear combination of nodes in $\mathbf{F}(t,s)$: $\tilde{\mathbf{f}}(t,\overset{*}{x}) = \mathbf{L}_k^{-1}\mathbf{F}(t,s)$. The operator $\mathbf{L}_k^{-1}$ approximates kth derivative with respect to s. Because $\mathbf{L}_k^{-1}$ approximates the inverse Laplace transform, $\tilde{\mathbf{f}}(t,\overset{*}{x})$ provides an approximation of the transformed function. It turns out (Shankar & Howard, 2012) that the width of the activity of each unit in $\tilde{\mathbf{f}}(t,\overset{*}{x})$ depends linearly on its value of $\overset{*}{x}$ with a Weber fraction that is determined by the value of k.

Accessing the function

The representation described above stores working memory as a vector-valued approximation of a function over an internal variable. We assume that this entire function cannot be accessed all at once, but that one can compute vector-valued integrals weighted by attentional gain over the function. The microcircuit includes an attentional gain function $\mathbf{G}(\overset{*}{x})$ that is

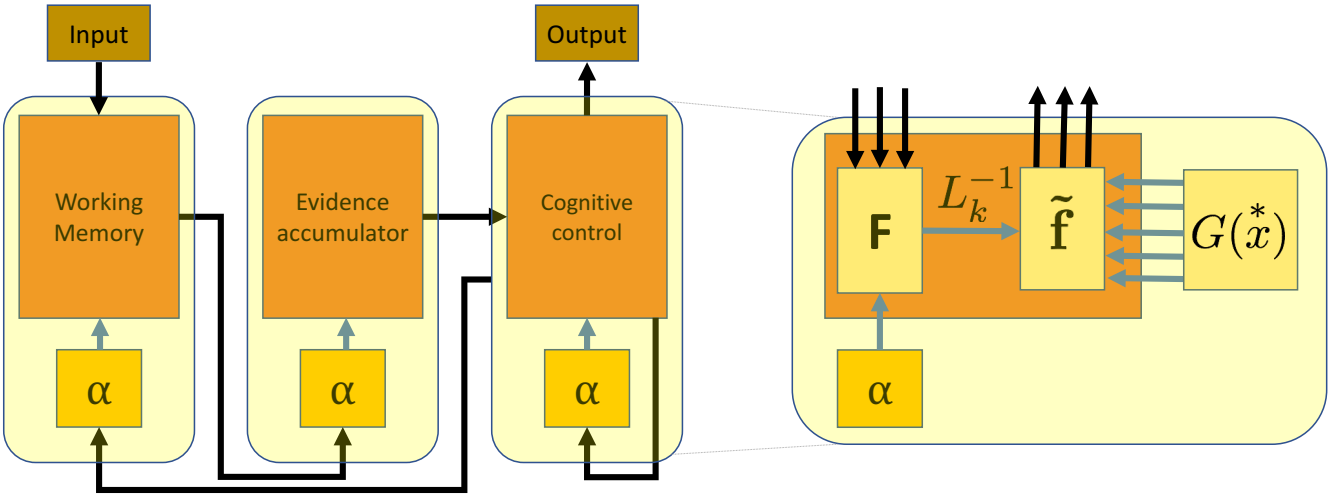

Figure 1: **A schematic of a neural-level circuit that can be used to model different behavioral tasks. This circuit was used to implement all the tasks described here.** The diagram on the left-hand side displays a configuration of the circuit composed of three blocks: working memory, evidence accumulator and cognitive control. The cognitive control block executes a sequence of actions. While a particular action is executed (e.g. waiting for a probe) the sequence is paused by setting its own α to 0. To move to the next action α is set to -1. Some actions will access working memory and feed the memory output to the evidence accumulator (e.g. to compare the probe with the content of memory). Output of the evidence accumulator is sent to the cognitive control block where it is used to trigger an appropriate action (e.g. press the left button). Each of the three blocks on the left-hand side is implemented with the microcircuit shown on the right-hand side. The microcircuit takes a vector input (fed into $\mathbf{F}(t,\overset{*}{x})$) and outputs a vector of the same size (through $\tilde{\mathbf{f}}(t,\overset{*}{x})$) selected by the attentional gain field $\mathbf{G}(\overset{*}{x})$ (multiple arrows from $\mathbf{G}(\overset{*}{x})$ represent that it can select different $\overset{*}{x}$ from $\tilde{\mathbf{f}}(t,\overset{*}{x})$). Depending on the initialization and inputs, this multipurpose microcircuit can run a predefined sequence in a self-modulating manner (by modulating its own α), store a compressed memory representation through sequential activation in $\tilde{\mathbf{f}}(t,\overset{*}{x})$ or encode functions of variables (e.g. accumulated evidence) for which a temporal derivative is available.

externally controllable. The output of the microcircuit at any moment is:

$$\mathbf{O}(t) = \sum_{i=1}^{N} \mathbf{G}(\overset{*}{x}_i)\tilde{\mathbf{f}}(t,\overset{*}{x}_i), \tag{2}$$

where N is the number of values of $\overset{*}{x}$ used to implement the function approximation $\tilde{\mathbf{f}}$. In models used here we restrict $\mathbf{G}(\overset{*}{x})$ to be unimodal across $\overset{*}{x}$. Attentional gain field can be made narrow and then activated sequentially, allowing a scan of the function representation or it can be made broad to sum across the $\overset{*}{x}$. This enables one to construct cognitive models based on scanning (e.g., Hacker, 1980) or to construct global matching models (e.g., Donkin & Nosofsky, 2012).

Working memory: Functions of time

When $\alpha(t)$ is constant, $\tilde{\mathbf{f}}$ maintains an estimate of $\mathbf{f}(t)$ as a function of time leading up to the present and we write $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$. If the input stimulus was a delta function at one point in the past, the units in $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ activate sequentially with temporal tuning curves that are broader and less dense as the stimulus becomes more temporally remote (Figure 2A). Neurons with such properties, called time cells, have been observed in mammalian hippocampus (MacDonald, Lepage, Eden, & Eichenbaum, 2011) and prefrontal cortex (Tiganj, Kim, Jung, & Howard, 2017). Furthermore, different stimuli trigger different sequences of cells (Tiganj et al., 2018), Figure 2B. Taken together at any time t, $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ can be understood as a compressed memory timeline of the past. The application of the Laplace transform in maintaining working memory in neural and cognitive modeling has been extensively studied (e.g., Shankar & Howard, 2012; Howard, Shankar, Aue, & Criss, 2015).

Evidence accumulation: Functions of net evidence

In simple evidence accumulation models, the decision variable is the sum of instantaneous evidence available during the decision-making process. In these models, a decision is executed when the decision variable reaches a threshold. By setting $\alpha(t)$ to the amount of instantaneous evidence for one alternative, we can construct the Laplace transform of the net amount of decision variable since an initialization signal was sent via the input $f(t)$. If no new evidence has been observed at a particular moment then $\frac{dF(t,s)}{dt} = 0$, thus all the units remain active with sustained firing rate. Large amount of evidence will, on the other hand, mean a fast rate of decay. Inverting the transform results in a set of cells with receptive fields along a "decision axis" (Howard, Luzardo, & Tiganj, 2018) consistent with recent findings from mouse recordings (Morcos & Harvey, 2016).

Cognitive control: Functions of planned actions

The program flow control activates a sequence of actions necessary for completion of a behavioral task. For instance, a typical behavioral task may consist of actions such as attend-

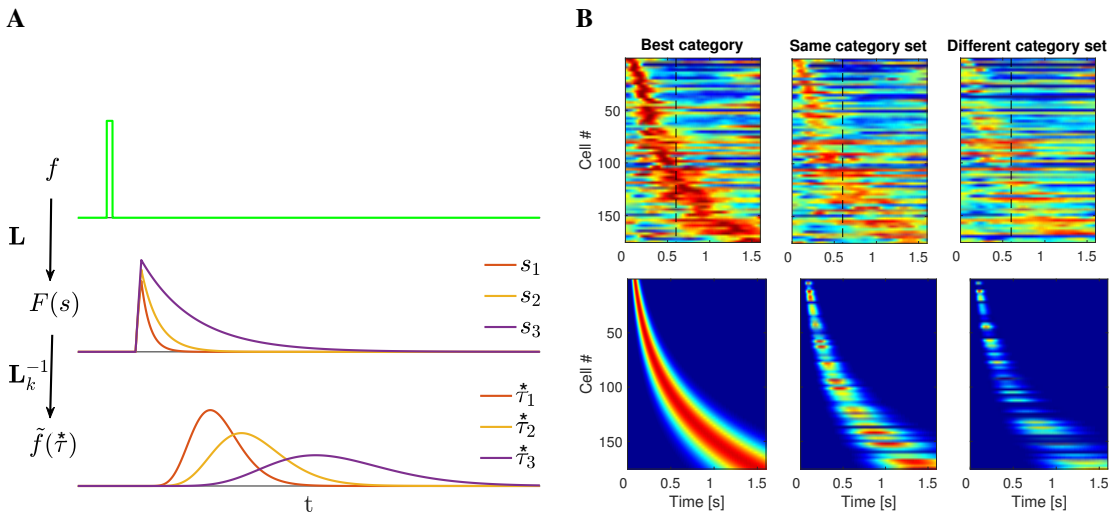

Figure 2: **A scale-invariant compressed memory representation through an integral transform and its inverse: model and neural data**

A. A response of the network to a delta-function input. Activity of only three nodes in each of the two layers is shown. Nodes in $\tilde{f}(\overset{*}{\tau})$ activate sequentially following the presentation of input stimulus f. The width of the activation of each node scales with the peak time determined by the corresponding $\overset{*}{\tau}$, making the memory scale-invariant. Logarithmic spacing of the $\overset{*}{\tau}$ makes the memory representation compressed. **B.** Top: During DMS task sequentially activated cells in monkey lPFC encode time conjunctively with stimulus identity (firing rate encodes visual similarity of the stimuli - stimuli in "Best category" were visually more similar to stimuli in the "Same category set" than to stimuli in the "Different category set"). The three heatmaps show neural activity during the stimulus presentation (first 0.6 s) and the delay period (following 1 s) averaged across trials. (Taken from Tiganj et al. (2018)). Bottom: Activity of the units in the working memory block of the architecture resembles the neural data.

ing to stimuli, detecting the probe, accumulating evidence and taking an appropriate action depending on which of the available choices accumulated more evidence. These operations require the ability to route information to and from the working memory and evidence accumulation modules. For instance, in order to compare a probe to the content of memory, one might route the output of the working memory unit, filtered by a probe stimulus, to the $\alpha(t)$ of an evidence accumulation unit. Because various operations take place in series, we can understand them as a function of future planned actions. Rather than past stimuli, the vectors in $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ can be understood as operations that affect other units (each action has a corresponding two-layer network turning $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ into vectors across the action space).

Different cognitive models correspond to different initial states in $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$. The actions will be executed sequentially by setting $\alpha(t) < 0$, winding the planned future closer and closer to the present. For instance, if the first step of a behavioral task is to wait for a probe, then that action will set the controller's $\alpha(t)$ to 0 until the probe is detected. Once the probe is detected, $\alpha(t)$ will be set to a default value of -1 so the neurons in the first layer will grow exponentially and the sequence loaded in $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ will continue evolving.

Integrating microcircuits into cognitive models

The three blocks described above: working memory, evidence accumulation and cognitive control are all constructed from the same microcircuit (Figure 1 right-hand side). Each circuit has an input, α and output. To demonstrate the utility of this approach, we connected the three blocks such that the program control block gates information from the working memory block to the evidence accumulation block and monitors its output (Figure 1 left-hand side).

Results

We demonstrate performance of the proposed architecture on three classical behavioral tasks: Judgment of Recency (JOR), probe recognition and Delayed-Match-to-Sample (DMS). We compare the results of the model with behavioral data (for JOR and probe recognition) and neural data (for DMS). Critically, even though these three tasks have very different demands, the neural hardware for the models is identical. The only difference is in the initial state of the program block. After initialization, each model runs autonomously and is self-contained.

Judgment of Recency: Sequential scanning of the memory timeline

In JOR subjects are presented with a random list of stimuli (e.g. letters or words) one at a time, and then probed with two stimuli from the list and asked which of the two stimuli was presented more recently. The classical finding is that the time it takes subjects to respond depends on the recency of the more recent probe, but not the recency of the less recent probe (Figure 4A) (Hacker, 1980; Singh & Howard, 2017). This result is consistent with a self-terminating backward scan along a temporally organized memory representation, suggesting

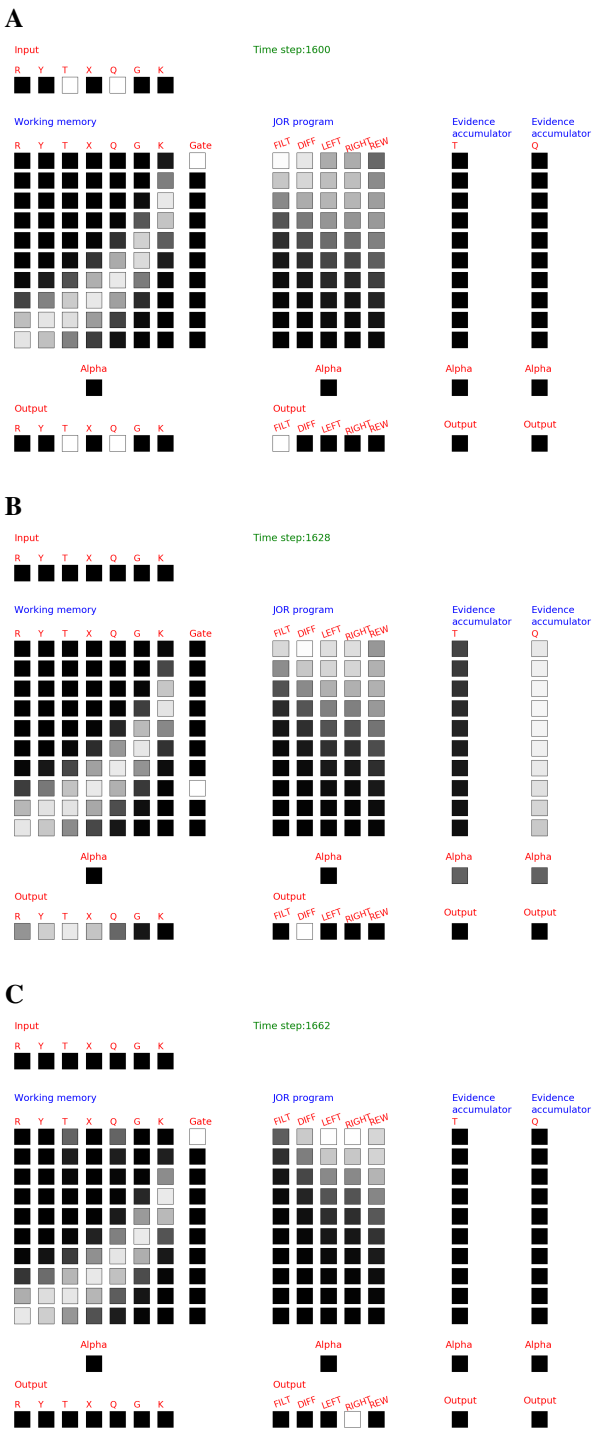

Figure 3: **Example of a JOR task implemented with the proposed architecture.** The implementation is done with microcircuits that correspond to those in Figure 1. Each square corresponds to a single neuron. Squares in the middle layer of each panel correspond to single neurons from $\tilde{\mathbf{f}}(t,\overset{*}{x})$ (neurons from $\mathbf{F}(t,s)$ are not shown). Shading reflects the activity of the neuron at a given time step; darker shading means less activity. **A.** At this time step all the seven items from the test list have been presented and they are stored in the sequentially activated memory. The two probe items T and Q are at the input. **B.** The program (cognitive control) block sequentially gates the information from the working memory into the α neuron of the evidence accumulator (DIFF action in the program block), causing sequential activation in the accumulator. **C.** After the evidence accumulator reaches the threshold, program control continues execution by activating an appropriate action (in this case RIGHT).

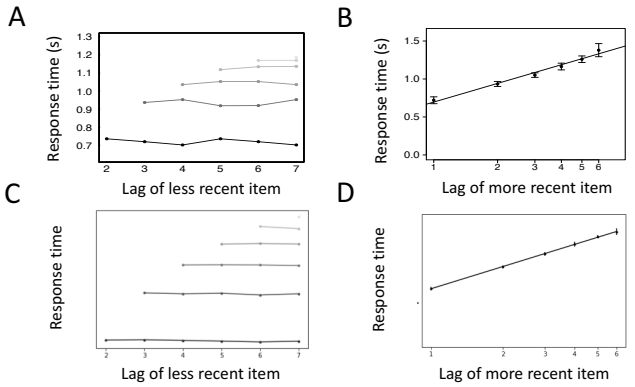

Figure 4: **The model captures behavioral results in the JOR task. A.** In JOR, median response time for correct responses depends strongly on the recency of the more recent probe but not the recency of the less recent probe. Shade of the line denotes lag of the more recent item, with the most recent item shown in black and the most distant item shown in the lightest shade of gray. (From Singh and Howard (2017).) **B.** In JOR, median response time varies sublinearly with recency (x-axis is log-spaced). **C.,D.** Results of the model corresponding to **A** and **B** respectively.

that subjects maintain working memory as a temporally organized, scannable representation. Moreover, the response time is a sublinear function of the lag (Figure 4B) (Singh & Howard, 2017), suggesting that the working memory representation is log-compressed, as proposed by earlier modeling work (Howard et al., 2015; Brown, Neath, & Chater, 2007).

In the model of JOR, the first action was to wait for the probe item to appear (Figure 3A). After that, the gain field over $\overset{*}{\tau}$ was set to scan the memory representation sequentially from more recent towards more distant past. At each step, the value found in the memory was used to drive two evidence accumulators, one independent accumulator for each probe item (Figure 3B). Once one of the two evidence accumulators reached a threshold, the program executed an appropriate action (left or right choice, Figure 3C). Variability in the response times was obtained by adding Gaussian noise to the evidence accumulation process.

Results in Figure 4C indicate that the model captures well the aspect of the data that suggests sequential scanning (Figure 4A): response time depends on the lag of the more recent probe item and does not depend on the lag of the more distant probe item. In addition, the model is consistent with the data regarding compression of the memory representation (Figure 4B - data, Figure 4D - model): the response time grows with the lag of the more recent item.

Old-new probe recognition: Global matching model using the memory timeline

Similarly to JOR, in old-new probe recognition task subjects are presented with a random list of stimuli one at a time. After the list is presented subjects are probed with a single probe that was or was not an item from the list. Subjects choose either *Old* or *New* to indicate their memory. The well-established behavioral results indicate that the response time

increases and accuracy decreases with increasing lag of the probe item (Figure 5A). In other words, if the probe item was further in the past (had larger lag) subjects will take longer to respond and their accuracy will be lower than if the probe was presented less far in the past. Models based on global matching, such as EBRW have managed to capture subjects accuracy and response times (Donkin & Nosofsky, 2012; Nosofsky, Little, Donkin, & Fific, 2011).

Our implementation of probe recognition was similar to JOR, but with several important differences. The main difference between the two tasks was in the way the memory was accessed. Unlike in the implementation of JOR where $G(\overset{*}{x})$ was a delta function resulting in serial scanning, in probe recognition task $G(\overset{*}{x})$ was uniform. This means that the entire memory representation was accessed simultaneously, rather than sequentially scanned. This type of memory access falls under the umbrella of global matching models which includes e.g. EBRW, SAM, Minerva and TODAM (Raaijmakers & Shiffrin, 1980; Murdock, 1982; Hintzman, 1988; Nosofsky et al., 2011).

Figure 5B shows model performance in probe recognition. The two qualitative features observed in the data were captured with the model: response time increased and accuracy decreased as the lag of the probe item increased. Overall, the result of the model resembles the data reported by Donkin and Nosofsky (2012).

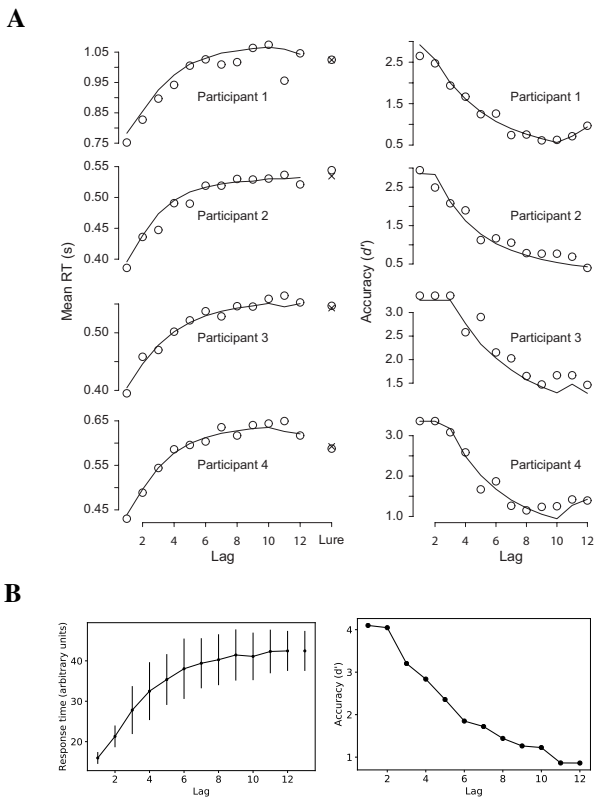

Figure 5: **The model captures behavioral results in the probe recognition task. A.** In probe recognition response times increase and accuracy decreases as the lag of a probe item increases. Circles correspond to data points and solid line is a fit obtained with EBRW model. Taken from (Donkin & Nosofsky, 2012). **B.** Results of the model capture qualitative properties of the data. Response times are shown with standard deviation.

Delayed-Match-to-Sample: Comparing model neurons to empirical evidence for conjunctive coding of what and when

In DMS subjects are presented with a sample stimulus followed by a delay interval, followed by a test stimulus. The action that subjects need to take (e.g. pressing a left or right button) depends on whether the two stimuli were the same or different. We modeled the task with the same components as the JOR task. The only differences were in 1) how the probe item was set (in DMS the second stimulus is by construction the probe, while in JOR the probe is marked by presenting two stimuli at the same time) and 2) what parts of the working memory were gated to the evidence accumulator (in DMS one accumulator accumulated evidence for presence of the probe item in the memory and the other accumulator accumulated evidence that any other item was found in the memory, while in JOR each of the two probe items had its own evidence accumulator). While simple in terms of behavior, DMS task is often done on animals while recording activity of individual neurons. Neural recordings during the delay period of this task show evidence for existence of stimulus-selective sequentially activated cells (Tiganj et al., 2018) that correspond well to the neural activity produced by the sequential memory used here (Figure 2B).

Conclusions

Here we provided an architecture that is based on realistic neural data and that can account for non-trivial behavior. In particular, the behavioral results of JOR task are consistent with the hypothesis that the subjects are scanning along a compressed timeline. The same architecture was used to model DMS task, resulting in neural representation of working memory that closely corresponds to the neural data. Finally, we have also captured qualitative properties observed in probe recognition task by applying an approach analogous to global matching models, but implemented on a neural-level.

Critically, implementation of all three tasks uses the same neural hardware, differing only in the initial condition of the controller. This work is complementary with ongoing efforts of building cognitive architectures such as ACT-R (Anderson, Matessa, & Lebiere, 1997) and SOAR (Laird, 2012). The distinction of the present work is in its attempt to build such architecture with neuron-like units, similar to Spaun (Eliasmith et al., 2012), but with a different type of neural representation. The present work commits to a specific type of representation: variables are represented as supported dimensions via neural tuning curves, tuned to a particular amount of elapsed time, accumulated evidence or a position in a sequence.

Acknowledgments The authors gratefully acknowledge support from ONR MURI N00014-16-1-2832, ONR DURIP

N00014-17-1-2304 and NIBIB R01EB022864.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462.

Bhandari, A., & Badre, D. (2018). Learning and transfer of working memory gating policies. *Cognition*, *172*, 89–100.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-76.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*. doi: 10.1177/0956797611430961

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, *15*, 846-858.

Hasselmo, M. E., & Stern, C. E. (2018). A network model of behavioural performance in a rule learning task. *Phil. Trans. R. Soc. B*, *373*(1744), 20170275.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in multiple-trace memory model. *Psychological Review*, *95*, 528-551.

Howard, M. W., Luzardo, A., & Tiganj, Z. (2018). Evidence accumulation in a laplace domain decision space. *Computational Brain and Behavior*, *1*, 237–251.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, *34*(13), 4692-707. doi: 10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., Shankar, K. H., Aue, W., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24-53.

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT press.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737-749.

Morcos, A. S., & Harvey, C. D. (2016). History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature Neuroscience*, *19*(12), 1672–1681.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609-626.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280-315.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, p. 207-262). New York: Academic Press.

Salinias, E., & Sejnowski, T. (2001). Gain modulation in the central nervous system: Where behavior, neurophysiology, and computation meet. *Neuroscientist*, *7*, 430–440.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134-193.

Sherfey, J. S., Ardid, S., Miller, E. K., Hasselmo, M. E., & Kopell, N. J. (2019). Prefrontal oscillations modulate the propagation of neuronal activity required for working memory. *bioRxiv*. doi: 10.1101/531574

Singh, I., & Howard, M. W. (2017). Recency order judgments in short term memory: Replication and extension of hacker (1980). *bioRxiv*, 144733.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed timeline of recent experience in monkey lateral prefrontal cortex. *Journal of cognitive neuroscience*, 1–16.

Tiganj, Z., Kim, J., Jung, M. W., & Howard, M. W. (2017). Sequential firing codes for time in rodent mPFC. *Cerebral Cortex*, *27*, 5663-5671.