Computational models of memory

Marc W. Howard
Department of Psychology
Syracuse University

Computational models of memory are constrained from a number of directions. A physically accurate model of memory must simultaneously describe the details of behavioral performance as well as the mechanistic neural processes that support such performance. Such a model should simultaneously explain both changes in behavior due to manipulations of the neural substrate, such as lesions or pharmacological manipulations, as well as describing the detailed neurophysiology of the underlying structures, as measured by fMRI, EEG/ERP and the activity of single neurons. We are far from such a comprehensive view of the neural basis of the types of memory typically studied by cognitive psychologists. However, some progress has been made towards this ideal in several domains. This work can be divided into content areas, distinguished by the types of memory and brain region supporting them, or level of approach. This article describes the levels of approach that are possible in constructing computational models of memory and reviews progress in one content area—computational models of working memory.

Approaches to modeling memory

An end-state comprehensive neural model of cognition must be formulated at several levels of analysis, from the molecular to the very abstract. The question though, is where to start in this vast endeavor. Presumably one can start from an accurate projection of the correct comprehensive model at any particular level of description and work at redescribing the model at adjacent levels of description. Appropriately, then, computational models of memory have been constructed at several levels of description.

Formal or mathematical models of memory are typically focused on describing behavioral performance in as near complete detail as possible. Typically models in this tradition start from some simple but abstract low-level structure from which asymptotic or average behavior can be inferred. Although it is clearly desirable that this implementation be interpretable as some neural process, concerns of neural realism have not historically been considered diagnostic in this tradition. For instance, stimulus sampling theory, an early example of this approach developed in the 1950's, postulated that a stimulus was represented by a set of abstract stimulus element. Learning of an S-R association consisted of binding a subset of the elements corresponding to a stimulus to a response. With repeated presentation of a stimulus, a new sample of elements is conditioned to the appropriate response, resulting in learning that gradually approaches an asymptote. An elaborated version of this simplified model was developed to describe changes in the strength of retention with different retention intervals and presentation schedules by postulating that the set of stimulus

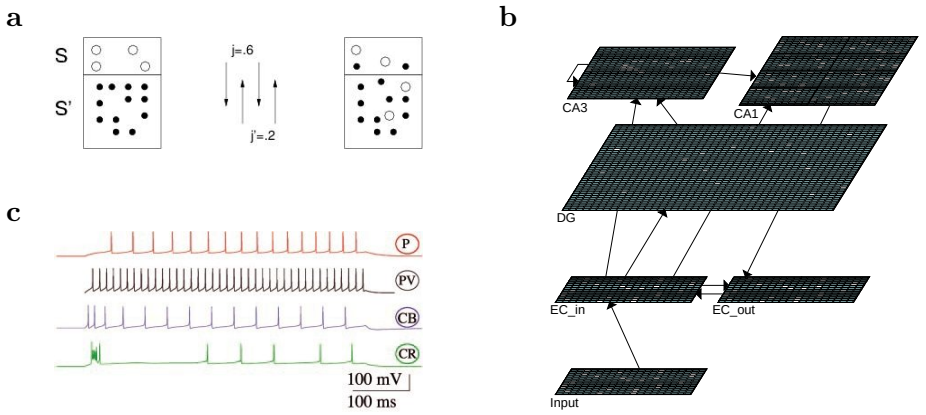

Figure 1. Approaches to computational models of memory. **a.** Mathematical models, here instantiated by Estes' stimulus sampling model generate behavioral predictions from a small set of mathematical rules that need not map onto actual neural processes. **b.** Connectionist models, here instantiated by Norman and O'Reilly's model of item recognition describe memory performance using a set of quasi-neural processing elements. Adapted from Norman and O'Reilly (2003), copyright APA, with permission. **c.** Biophysical models of memory attempt to provide a realistic description of the cellular processes that contribute to memory.

elements available to be conditioned fluctuated gradually over time. Although the abstract processing elements of stimulus sampling theory do not map simply onto any known neural process, or even brain region, closed-form analytic solutions can be derived for a number of important quantities that map onto behavioral observables. Mathematical models of memory have grown more elaborate since the 1950s, such that computer simulation is usually required to develop complete behavioral predictions.

To provide a more concrete idea of how such ideas might be actualized, Figure 1a shows a figure illustrating some of the basic assumptions of stimulus sampling theory, a prominent early mathematical model of memory. A set of abstract elements, represented by filled or open circles are available to represent a stimulus on each trial. The set of such elements used to represent the stimulus on a trial is represented by the set S. The remainder of the set of stimulus elements, which are unavailable is referred to as S'. On the first learning trial, shown on the left, all of the elements in S are conditioned to the response. Conditioned elements are shown as large open circles whereas unconditioned elements are shown as small filled circles. Over a delay, abstract processing elements shift from S to S', and vice versa, with transition probabilities defined by j and j'. These fluctuations lead to a forgetting of the conditioned response as conditioned elements used to describe the stimulus transition from the active set to the inactive set.

Connectionist approaches to memory modeling adopt a somewhat intermediate level of description. Rather than completely abstract processing elements, connectionist models rely on sets of interconnected units that are often mapped explicitly onto specific brain regions. These units typically have a scalar activation at each time step that corresponds to firing rate, perhaps averaged over a group of physical neurons and connections, perhaps modifiable, between units that correspond to synapses, or at least pathways between brain

regions. Although one can find counterexamples to all such generalizations, connectionist models of memory focus less on a quantitative description of behavior than do mathematical models. These models are evaluated as much on the basis of their neural plausibility given what is known about the anatomical circuits and physiology underlying the regions. These constraints can come from correctly describing the qualititative effects of physiological manipulations, such as brain lesion or pharmacological manipulations.

Figure 1b shows an example of a connectionist model of item recognition memory, a memory task in which the subject is given a probe item and asked whether or not it was presented as part of a study episode. In this model, a set of interconnected abstract processing elements are identified with brain regions. In this case, an input pattern is presented to a set of processing elements identified with input layers of the entorhinal cortex (EC). These processing elements are connected to other elements identified with other brain regions—in this case the dentate gyrus (DG) and the subfields of the hippocampus (CA3 and CA1). The model's memory for a recognition probe can be assessed by examining the degree to which the pattern in the output layer of the EC (EC_{out}) resembles the input pattern presented to the input layer (EC_{out}) when the pattern is re-presented as a probe item after learning.

At a yet more detailed physical level, biophysical models of memory attempt to describe memory processes using models of neural activity that are not merely plausible, but as close to physically accurate as possible. Typically model neurons in these simulations fire individual spikes, with a set of realistic conductances taken from experimental data. The goal of biophysical models of memory is typically not to model behavioral data *per se*, but rather to show how a mechanism believed to relate to some cognitive mechanism supporting some type of memory performance could be physically implemented. These models are typically evaluated in terms of their ability to accurately describe the currents and firing rates observed in neurons during performance of a task believed to be related to the cognitive processes one is studying.

Figure 1c shows a biophysical model designed to capture aspects of working memory maintenance, a mechanism believed to be required to retain information in attention for a short period of time. In this simulation, there are several subtypes of simulated neurons that are interconnected to each other. Figure 1c shows the response of each of the types of model neurons in response to a transient injected current. Simulated pyramidal cells (P) are interconnected with three types of interneurons. The pyramidal cells are modeled as three-compartment neurons—that is at the subcellular level. The pyramidal cells and interneuon populations differ in their intrinsic properties, synaptic properies and network connectivity. In this simulation, the traces labeled PV, CB and CR correspond to three different subpopulations of interneurons. This model is primarily evaluated on the basis of its ability to describe the detailed behavior of neurons of the different classes observed in the prefrontal cortex during the maintenance phase of a working memory task in which a monkey must remember a spatial location.

These three examples illustrate a number of differences between mathematical, connectionist and biophysical approaches. They can be seen as lying on a continuum in terms of several variables. The spatial scale over which these models are specified obviously varies, from the behavior of the whole organism, in the case of mathematical psychological models, to sets of cells identified with brain regions in the case of connectionist models, to sub-

cellular processes in the case of the biophysical model. These approaches also differ with respect to the temporal scale over which the models are specified. Whereas mathematical models often treat memory in units of stimulus presentations, which may be several seconds, connectionist models typically measure time in tenths of seconds. In contrast, biophysical models often describe somatic voltage at the sub-millisecond time scale.

In addition to the (often considerable) difficulty associated with getting a model to work within a particular level of description, there is an additional level of difficulty in constructing models that make contact across levels of description. We use the term memory to encompass a very broad range of phenomena, including everything from an *aplysia* learning to retract its gill to a person consciously recollecting the sounds and sights of a specific event given, say, a familiar odor as a cue. These phenomena, and others, are typically studied separately, with the range of phenomena under examination organized by a putative memory system, or perhaps the specific brain region (or regions) believed to be crucial in a particular task. A great many such problems that have been addressed using compuational models of memory at one or more level of investigation sketched out here. Some of the more notable topics that have benefitted from computational models of memory include categorization, delay and trace conditioning, cognitive control, associative learning, and hippocampal-dependent memory (see especially Koene & Hasselmo chapter, this volume). A thorough survey of these diverse problems is impossible given the space constraints of this article. Instead, we will restrict the content of our discussion severely and discuss computational models of working memory maintenance. This area is perhaps particularly well-suited to give the reader a flavor of the different approaches that are possible using computational models of memory. The models that have been developed to study working memory are also perhaps notable in that a number of attempts have been made to bridge multiple levels of analysis.

Models of working memory

To cognitive psychologists, the term working memory is used to describe the set of cognitive processes required to store information actively in attention for a short period of time and make decisions based on that information. Ideas about working memory have evolved from previous conceptions of short-term memory. These models pay differing amounts of attention to the necessity to, on the one hand, store information for a short period of time, and, on the other, control what information enters into short-term storage and how it might be manipulated.

The 1960's and 1970's saw development of very successful mathematical models of short-term memory. The most prominent of these was the Atkinson and Shiffrin (AS) buffer model (Figure 2a). In this model, attentional control processes gate information into and out of a labile but limited capacity short-term store (STS). Activity in short-term store is closely associated in this view with the process of rehearsal, in which subjects actively and consciously repeat to-be-remembered materials to themselves. For instance, suppose that someone tells you a phone number that you must remember for a few seconds before you can write it down. A common strategy in this situation is to repeat the digits to yourself to keep them "fresh" in your memory. Subjects in human verbal learning studies frequently adopt a rehearsal strategy to help them remember the study items. Formally, STS was conceived of as a small number of slots, typically two to four, into which study items could

be copied. There are a small, integral number of slots and an item can either be present or absent in the buffer—no partial activation was allowed. The relatively small capacity of STS means that decisions must be made about what information to prioritize. In the AS model, these decisions are assumed to be under conscious strategic control, although the rules for deciding what information should be maintained or discarded were not explicitly modeled but rather assumed to reflect specific strategies appropriate to a particular task. Finally, STS serves as a gateway for information to be stored in long-term store (LTS). The longer an item remains active in STS, the better the chances that it will be successfully stored and later recovered from LTS.

The AS model, and its successors, were extraordinarily successful at precisely describing human behavioral data in a relatively wide variety of experimental paradigms, including a continuous paired-associates task and the free recall task, in which subjects are presented with a list of words and then instructed to recall them in any order they think of them. As a consequence, ideas about short-term memory from the AS model, including the notion of a short-term store that retains an integral number of stimuli in a discrete fashion, have persisted and been implemented in a variety of biophysical and connectionist models of working memory.

Connectionist models of working memory have emphasized the importance of maintenance of information as a subprocess in the service of cognitive control. For instance in the Stroop task, subjects are presented with a color name presented in a colored font that may be inconsistent with the named color; for instance, the word RED written in blue. In this task, subjects are instructed to read the name of the word rather than report its actual color. Subjects are much slower to read the name, and make more errors, when the physical color is inconsistent with the written color name (e.g. the word RED written in blue) than when it is consistent (e.g. the word RED written in red). Neuropsychological data indicates that the ability to inhibit the tendency to say the name of the physical color depends on an intact prefrontal cortex. A number of closely-related connectionist models of working memory describe the ability to control the inappropriate response by means of active attentional maintenance of a rule representation, i.e. "name." This rule representation modulates, or gates, the appropriate behavioral response to the physical stimulus. For the rule representation to remain active, it must be actively maintained in a short-term buffer in prefrontal cortex. In these connectionist models, then, the concept of active maintenance is retained, but the emphasis is more on the maintenance of rules rather than information to be committed to long-term memory.

Biophysical models of working memory have focused on the mechanisms of active maintenance rather than either the role of this maintenance in executive processing or its potential role in transfer to long-term memory. A variety of biophysical mechanisms have been proposed for working memory maintenance in at least two distinct regions. These models can be roughly divided into two classes—those that rely on network properties to allow persistent firing, and those that rely on intrinsic properties of individual neurons that allow them to maintain firing over long durations. A goal of many of these efforts has been to implement bistability—a population of neurons that have two stable states, one with a high sustained firing rate and one with a much lower firing rate. In this sense, these models have retained the idea of all-or-none activation present as early as the AS buffer model.

Network models rely on the idea that a population of neurons is divided into cell

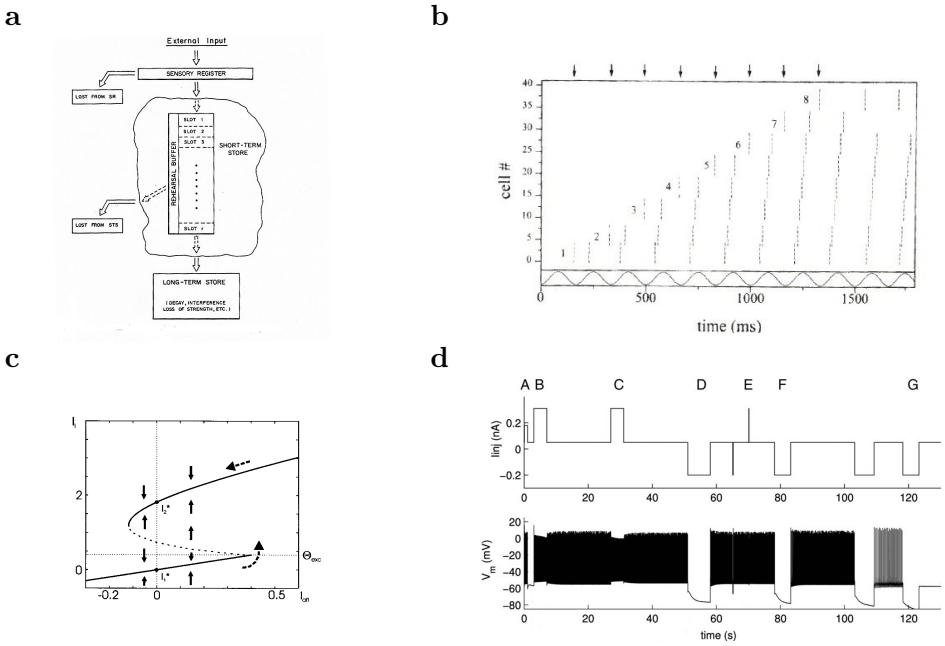

Figure 2. Approaches to models of working memory. **a.** The buffer model of working memory maintenance. Incoming information is stored in one of a discrete number of working memory buffers. Information is either present or absent in these buffer slots. **b.** The Lisman Idiart Jensen (LIJ) model is a biophysical instantiation of the buffer model in which neurons representing different items coactive in the buffer occupy different subcycles of an ongoing theta oscillation. This plot shows a raster plot from a biophysical simulation. Neurons corresponding to each of several items are presented. Patterns of activated items are either in the buffer or not. For instance, item 7 is intitially present in the buffer, but then falls out when item 8 is presented. **c.** Bifurcation plot of dynamics in a multistable biophysical model of working memory maintenance that exploits network properties to implement bistability. The network consists of cell assemblies with excitatory recurrent connections and inhibitory lateral connections. The y-axis plots the firing rate of one of the asemblies and the x-axis plots its external input current. The stable states of the dynamics are shown in solid. When the input current exceeds a threshold (about .4 in the figure), the activity of the assembly jumps from the value on the lower line to the value on the upper line (curved arrow). After this takes place, the cell remains active even after the external current is removed (dashed line). A sufficiently large inhibitory current can cause the assembly to return to rest (left side of figure). **d.** Biophysical simulation of multistability observed in layer V cells in the entorhinal cortex. Under appropriate conditions, these cells respond to a depolarizing input with a sustained non-zero firing rate. As additional inputs are presented, the cell adopts a new stable firing rate (B and C). This property does not depend on recurrent connections as it was observed under conditions of synaptic isolation. In the biophysical simulation shown here, multistability is a consequence of a fixed point attractor, the location of which can be changed by changes in the number of activated channels, a process which is triggered by sufficiently large changes in calcium concentrations. Note the long time scale in the figure.

assemblies with strong exitatory connections within an assembly and inhibitory connections between neurons from different cell assemblies. If the parameters of the model are chosen appropriately the network will have the property that if an assembly is given a sufficiently large external input, then the recurrent connections sustain a high level of firing that can persists even if the external input is removed. The assembly will then remain active until a sufficiently large inhibitory input, either an inhibitory external input or an excitatory input to another assembly, causes it to return to the rest state (Figure 2c). Network models are widely believed to describe the basis of sustained activity observed in neurons in the prefrontal cortex. The anatomical organization of dorsolateral prefrontal cortex, with interconnected pyramidal cells and a spatially distributed network of interneurons, suggests that it may be organized into columns that serve as cell assemblies.

Cellular mechanisms for bistability that do not depend critically on recurrent connections have also been proposed. The Lisman-Idiart-Jensen (LIJ) model proposes that large after-depolarizations observed in some cells may be sufficient to cause a cell to fire again. If an initial spike can cause a second spike, then the cell will continue firing for a macroscopic time interval. In order to allow for multiple items to be maintained simultaneously, the LIJ model proposes that brain oscillations play a role in organizing the firing of patterns of spikes corresponding to separate items that are activated in memory. In the LIJ model, theta oscillations, large 4-12 Hz oscillations in the local field potential that are especially prominent in the hippocampus, serve to provide a temporal frame for the item representations such that spikes corresponding to each of the items active in working memory fire once per theta cycle. The patterns corresponding to different items are kept separate from each other because the patterns are constrained to fire on different cycles of gamma (40-80 Hz) oscillations (see Figure 2b).

Using nested gamma and theta oscillations, the LIJ model has been used to model most of the features of the AS model in a biophysically plausible way. For instance, the biophysical model allows patterns of activity to drop out of the buffer in an all-or-none fashion (note the fate of pattern 7 in Figure 2b). Inclusion of synaptic plasticity via NMDA channels has been shown to lead to long-lasting changes in the strength of the connections between items active in short-term memory, analogous to the transfer from STS to LTS critical in the AS model. The major application of the LIJ model to directly describing cognitive data has been the Sternberg task. In the Sternberg task, the participant studies a list consisting of no more than a handful of stimuli. At the end of the list, the participant is given a probe item and must decide if the probe was part of the list or not. The key finding is that the amount of time it takes the subject to decide if the item was on the list or not increases linearly with the number of items on the list, at least for small list lengths. The LIJ explains this linear slope because of the strict temporal ordering of the items active in memory. In addition to the fact that the slope, typically around 40 ms/item, is on the order of what one would expect if one had to wait for a gamma cycle for the next item to come up, variants of the LIJ model have been shown to describe human performance in the Sternberg task in remarkable detail. Moreover, since the LIJ was first proposed, a wealth of data implicating theta and gamma oscillations in working memory and the Sternberg task in particular have been discovered. The LIJ provides a biophysical implementation of the assumptions of the AS model, and also Sternberg's ideas about how subjects scan through a short list of items to reach a decision on the probe. As such it provides a remarkable

example of a model that bridges multiple levels of description.

More elaborate forms of persistent neural activity that do not depend on interconnections between cells have been observed in the entorhinal cortex. Cells in layer V of the entorhinal cortex have been shown to be able to produce extremely long-lasting persistent firing in the absence of synaptic input. Moreover, these cells are able to gradually change their firing to reflect the summation of their inputs. That is, a first stimulus might cause the cell to fire steadily at a given rate. A second stimulus would cause the neuron to jump to a new steady firing rate. The mechanisms underlying this remarkable ability are a topic of intense investigation. It is known that persistent firing depends on a calcium-sensitive non-specific cation (CAN) channel that remains open for long periods of time. The firing rates a cell can sustain are apparently graded, rather than discrete and are observed even when intracellular stores of calcium are depleted. One theory of how this remarkable behavior is possible is that at any time the current dynamics obey a fixed point attractor that is insensitive to small perturbations. However, sufficiently large (in terms of amplitude and/or duration) changes in the internal calcium concentration cause the location of this fixed state to gradually change. The exact intracellular mechanisms of how this is accomplished are not known, but a plausible multicompartment model implementing these ideas is able to simulate the behavior of multistable integrator cells in remarkable detail (see Figure 2d).

Intrinsic persistent activity is an extremely important property for a general working memory system to have. If the ability to maintain information in working memory requires a pre-existing synaptic network, as in the recurrent networks supporting bistable cell assemblies like those shown in Figure 2c, then because cell assemblies must support each other, the number of patterns that can be maintained (at different times) must be less than the number of neurons. In contrast, if neurons can maintain their own firing for arbitrarily long periods of time without network connections, then the number of distinguishable patterns that can be maintained at different times can even exceed the number of neurons. This property is extremely useful for a working memory buffer in the spirit of the AS model that should be able to store at least all the words in the English language.

The other unique property of multistable cells is their ability to support a gradual decay of information rather than the precipitous all-or-none dropout rule of the AS model that is implemented in the bistable biophysical models. This gradual decay of information in working memory is a characteristic of a more recent mathematical model of working memory maintenance and storage of information into long-term memory—the temporal context model (TCM). Much like the AS model, TCM was formulated as a mathematical model of experiments where participants learn and recall lists of words. As such it has predicted and explained several findings that would be counterintuitive if the AS model were a literally accurate description of maintenance and storage of recently-presented information. TCM differs from the AS model in two respects. Rather than decaying in an all-or-none fashion, information decays gracefully from the working memory store in TCM, which is referred to as a temporal context vector. This allows TCM to describe gradual forgetting over much longer time scales than the AS model, a property that is consistent with some behavioral data. Second, in TCM the input to the working memory store caused by an item is not fixed over time, as in the AS model, but can change to reflect the temporal context in which the item was presented. This ability "bind" items to the temporal contexts they are encoded in turns out to have several important implications.

Recently, proponents of TCM made a bridging hypothesis to connect the abstract mathematical model with the actual neural substrate. They proposed that the temporal context vector should correspond to the pattern of activity across cells in extra-hippocampal medial temporal lobe regions, in particular the entorhinal cortex. They showed that a population of multistable cells, like those illustrated in Figure 2d, if subject to cortical gain control, could implement the equation for the change of temporal context in such a way as to allow gradual decay of information in the temporal context vector. Continuing the mapping between TCM and the brain, the function of the hippocampus proper is to enable items to be bound to and subsequently recover the temporal context, or state of activation in EC, that obtained when they were studied. This assumption made it possible for the model to account for the effect of hippocampal lesions on learning tasks in rats.

The approach of models exploiting the bistability property introduced by the AS model and the multistability property exploited by TCM differ considerably in their behavioral predictions and the range of potential implementations. It might be nice if one could declare one approach correct and the other incorrect. Another possible, and probably more likely outcome, however, is that both reflect some insight into how working memory maintenance behaves in different circumstances. One means for reconciliation is to note that the connectionist view of working memory maintenance as something that retains goal states in prefrontal cortex does not overlap particularly strongly with the view of a gradually-changing representation of recent experience in extra-hippocampal medial temporal lobe cortical areas. Perhaps both are accurate descriptions of related but distinct phenomena. In this case, bistable maintenance of goal states would not typically have much to do with how we remember recent events, whereas a gradually-changing state of temporal context would not be much use in performing the Stroop task.

Computational models of memory can be roughly divided into three classes that differ in their level of realism and commitment to explaining cognitive *vs* neural data: formal or mathematical models, connectionist models and biophysical models. Although all of these approaches are valid in their own right, the fact that recent models are attempting to bridge these levels is a promising development. Computational models of working memory at all three levels are reviewed, but it should be kept in mind that similar developments are taking place in the study of many types of memory and many different brain regions.

Suggested further reading

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, p. 89-105). New York: Academic Press.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624-52.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3-42.

Durstewitz, D., & Seamans, J. K. (2006). Beyond bistability: biophysics and temporal dynamics of working memory. *Neuroscience*, *139*(1), 119-33.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*, 1184-91.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145-154.

Fransén, E., Tahvildari, B., Egorov, A. V., Hasselmo, M. E., & Alonso, A. A. (2006). Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron*, *49*(5), 735-46.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75-116.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299.

Jensen, O., & Lisman, J. E. (1998). An oscillatory short-term memory buffer model can account for data on the Sternberg task. *Journal of Neuroscience*, *18*, 10688-10699.

Jensen, O., & Lisman, J. E. (2005). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in Neuroscience*, *28*(2), 67-72.

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

Miller, P., & Wang, X. J. (2006). Power-law neuronal fluctuations in a recurrent network model of parametric working memory. *Journal of Neurophysiology*, *95*(2), 1099-114.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611-46.

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.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Science, USA*, *102*(20), 7338-43.

Wang, X. J., Tegnér, J., Constantinidis, C., & Goldman-Rakic, P. S. (2004). Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proceedings of the National Academy of Science, USA*, *101*(5), 1368-73.