# Discrete Representations in Working Memory: A Hypothesis and Computational Investigations


Randall C. O'Reilly
Department of Psychology
University of Colorado
oreilly@psych.colorado.edu

Michael Mozer
Computer Science Department
University of Colorado
mozer@cs.colorado.edu

Yuko Munakata
Department of Psychology
University of Denver
munakata@kore.psy.du.edu

Akira Miyake
Department of Psychology
University of Colorado
miyake@psych.colorado.edu


We present a novel hypothesis concerning the nature and development of working memory representations and some initial computational investigations of this hypothesis. Working memory refers to the active maintenance of information in the service of complex cognition, such as language comprehension, spatial thinking, and problem solving (Miyake & Shah, 1999). We propose that the unique demands placed on the working memory system shape its representations over learning and development, affecting the use of working memory by the cognitive system as a whole. Our primary source of insight into this process comes from a computational analysis, which is used to integrate and explore relevant findings from neurobiology as well as developmental and adult cognition.

Our specific hypothesis is that to maintain information in an active state over delays and in the face of interference (e.g., from incoming stimuli, ongoing processing, and noise), working memory representations should be discrete in nature. A discrete representation admits to only a finite set of possible states, rather than representing continuous states. For example, the integers from 1 to 100 form a discrete set, in contrast to the real numbers in this range. Discreteness imparts a measure of robustness to the representation because small amounts of noise can be overcome by interpreting an observed state as the nearest discrete state (Figure 1). Many different processing mechanisms could achieve this remapping of perturbed states, including attractor dynamics in neural networks (Hopfield, 1984; Smolensky, 1986), nearest-neighbor classifiers (Cover & Hart, 1967), winner-take-all networks, and rule-based systems.

We suggest that because information must be actively maintained over relatively long time periods in working memory to complete complex cognitive tasks (e.g., 10s of seconds to minutes), it is in a unique position to benefit from discreteness. This discreteness comes at a cost, however, because it limits the level of fine detail or graded information that can be encoded. Thus, the representations underlying other cognitive functions may utilize less discrete, graded representations because they do not require as much noise tolerance and can therefore encode finer detail and more graded information.

From the central property of discreteness, a number of other properties follow. For example, discrete representations should be: more categorical, more easily verbalizable and generally accessible to other parts of the cognitive system, better for perceiving or performing a sequence of steps, and more "symbolic" in some respects. All these properties have generally been attributed to working memory representations as well, and an important component of our research is to explore the idea that they all follow from the more basic property of discreteness.

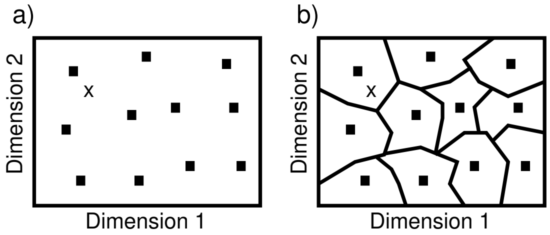


Figure 1: (a) Two-dimensional activity space with discrete representations where only a subset of states are meaningful, indicated by the small squares. The "X" denotes a corrupted version of one of the discrete states. (b) Discrete representations allow the space to be carved up into equivalence classes, and corrupted states (e.g., X) treated as equivalent within the boundaries of the discrete state.

## Existing Behavioral Data

A number of findings in the behavioral literature are consistent with the implications of our discreteness hypothesis, including:

- Continuous spatial location information is encoded in a categorical fashion in tasks where participants are asked to encode the location of a dot within a large circle (Huttenlocher, Hedges, & Duncan, 1991). Participants exhibit a systematic bias, shifting the dot closer to the center of the nearest quadrant of the circle. Our interpretation is that the categorical bias reflects the involvement of working memory. This interpretation is consistent with their finding that this bias effect increases in magnitude when a delay of 10 seconds is imposed between stimulus presentation and response selection, thereby necessitating working memory involvement.
- Categorization becomes more dimensionally focused

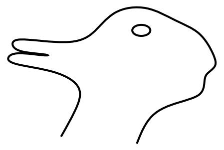

Figure 2: Ambiguous duck/rabbit figure.

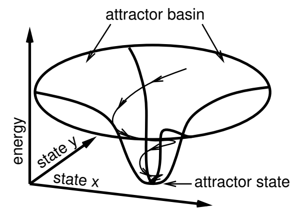


Figure 3: Illustration of an attractor state and its surrounding basin. Network updating causes the activation state (simplified by 2 variables, x & y) to converge on the attractor (by descending in energy) from anywhere within its surrounding basin.

and sensitive to exact identity matching with both increasing developmental age (Smith, 1989) and increasing amounts of processing time (Lamberts, 1995). In young toddlers and under speeded response conditions, categorization tends to be based on overall similarity regardless of differential similarities along different dimensions. However, older children and adults given sufficient processing time will place stronger weight on stimuli sharing the same value along a given dimension even when these stimuli are less similar along other dimensions (e.g., two red stimuli that are very different in size will be categorized together instead of an orange and a red stimulus of similar but not identical sizes). Thus, it appears that, under conditions when working memory is potentially engaged, categorization becomes much more discrete in both its dimensional focus and sensitivity to identity.

- Interpretation of ambiguous figures becomes more unambiguous (discrete) when they are held in working memory instead of being perceptually available. Chambers and Reisberg (1985) showed that participants were able to generate different interpretations of ambiguous figures (e.g., Figure 2) on direct viewing, but could typically generate only one interpretation when the figure was held in working memory. The inability to generate an alternative interpretation was not due to a lack of maintained detail in working memory, because participants could draw the figure from memory. Rather, we suggest the inability was due to the strong selection of one discrete interpretation in working memory.
- In tasks that require a comparison of alternatives that vary in fine-grained distinctions, such as faces or wines, verbalization impairs performance (e.g., Melcher & Schooler, 1996; Schooler & Engstler-Schooler, 1990). We suggest that requiring subjects to use discrete verbal representations for encoding will engage the use of the working memory to maintain the properties of one item while comparing it with another, instead of relying on more graded familiarity-like mechanisms that could be mediated directly by perceptual representations. Thus, we interpret these results as suggestive evidence that discrete working memory encoding has a deleterious effect on the ability to make fine-grained distinctions among stimuli.

## Computational Models

We have developed a series of computational models to explore working memory (including previous work by Munakata, McClelland, Johnson, & Siegler, 1997; Munakata, 1998). One line of research uses an abstract framework developed by Mozer (1998; Mathis & Mozer, 1996), and has demonstrated that discrete representations are more robust to noise, are more easily processed by other processing pathways, and are more influential over these other pathways. Further, we were able to characterize the circumstances under which discrete representations are likely to be important, as contrasted with other situations when they are less likely to be important. These results provide theoretical leverage in understanding what distinguishes working memory tasks from other tasks where working memory is not necessary.

Recently, we have used a more biologically-based framework (O'Reilly, 1998; O'Reilly, Braver, & Cohen, 1999) to explore some of the ways in which discrete representations can be manifested in biological systems such as the prefrontal cortex (PFC). These models take advantage of the *attractors* that form among recurrently-interconnected units to refresh and maintain active memories over time (Braver, Cohen, & Servan-Schreiber, 1995; Dehaene & Changeux, 1989; Zipser, Kehoe, Littlewort, & Fuster, 1993). Attractors (Figure 3) are so-named because they are states of activation that the network is drawn towards as activations are updated over time (settling). Attractors can maintain information by keeping the network stable in the attractor state over time.

The first series of three simulations reported here explore how different levels of discreteness can be manifested in terms of different patterns of interconnectivity among a set of neural units. These connectivity patterns result in different widths of the attractor *basins* (regions of activation space surrounding the attractor, from which the network will reliably settle into the attractor state, see Figure 3), and therefore the level of discreteness of the representational space. These attractor basins produce the equivalence classes shown in Figure 1.

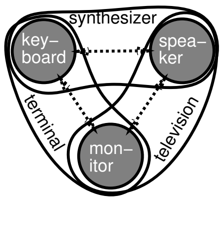


Figure 4: Distributed featural representations used for exploring active maintenance. Three objects (synthesizer, television, and terminal) can be represented using combinations of three features (keyboard, speaker, monitor).

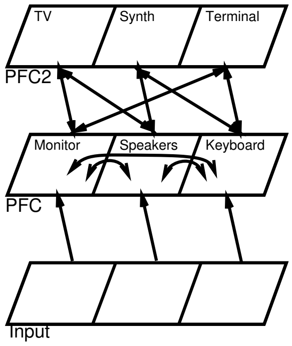


Figure 6: Network with higher-order representations encoding objects individually. This produces more discrete representations with wider attractor basins, and correspondingly better active maintenance.

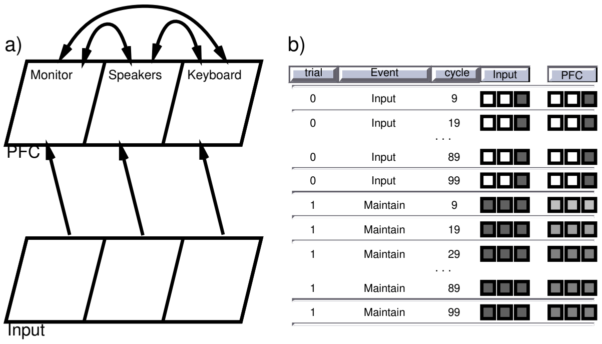


Figure 5: **a)** Network for simulation 1, with input providing activation to PFC (prefrontal cortex) units representing features as shown in previous figure. The feature units are all interconnected with each other to support the maintenance of activation. **b)** Activation spreads across the interconnections once the input is removed, as is shown by the PFC activation states plotted over time (lighter = more active).

The final simulation explores the use of a dopamine-based dynamic gating mechanism that is thought to actively regulate the strength of a subset of neuronal connections in the PFC (Cohen, Braver, & O'Reilly, 1996; Braver & Cohen, 1999). This gating mechanism imposes discreteness in the switching between maintenance and updating of working memory representations and is thus likely to contribute to the overall discreteness of working memory representations.

### Simulation 1: Spreading Activation in Continuous Distributed Representations

In this first simulation, we show that a working memory network with only pairwise (lateral) excitatory interconnections among distributed units (with global surround inhibition) exhibits virtually no ability to maintain information over time after the input is removed. Therefore, an alternative architecture is required, as discussed in Simulation 2. The simulation uses distributed representations as shown in Figure 4, and its goal is simply to maintain the representation of an object (e.g., "television" as encoded by the activity of the distributed features of "monitor" and "speakers") after the input pattern for that object is removed. The network is shown in Figure 5a. Although the PFC working memory units can encode a specific object using a distributed feature-based code while external inputs are present (e.g., when such an object is within view; shown in the top part of Figure 5b), once maintenance of this information is required without external input, the activation spreads across the interconnected distributed feature units and the object-specific encoding is lost (bottom part of Figure 5b). Thus, the network can no longer distinguish "television" from "terminal" or "synthesizer" based on the maintained information.

This result suggests that although overlapping, interconnected distributed representations can be very useful for perceptual processing (while inputs are present), they are not suitable for the active maintenance of information over time. These distributed representations produce a kind of representational continuum defined over the different combinations of unit activations within the space, and because all such combinations are supported by the distributed interconnections, the network is unable to lock onto and maintain only one of them. Thus, this example constitutes an extreme version of continuous (non-discrete) representations where active maintenance fails even in the absence of noise.

### Simulation 2: Wider, More Discrete, Attractor Basins

Next, we explore the effects of adding higher-order representations that encode the specific objects to be maintained (e.g., "television"). These representations amount to a second layer of units that connect to specific subsets of units in the first layer (Figure 6). With these representations, the network is able to maintain information over time, even with small amounts of noise, because there is a discrete attractor state corresponding to each object to be maintained. However, when the noise level is increased, the information is rapidly corrupted, because the distributed overlap in the feature layer makes the attractor basins relatively narrow. The level of noise robustness in the model can be predicted as a function of the distinctiveness of the working memory representations (i.e., the extent to which units are shared across multiple different representations). This suggests the idea explored in the next simulation.

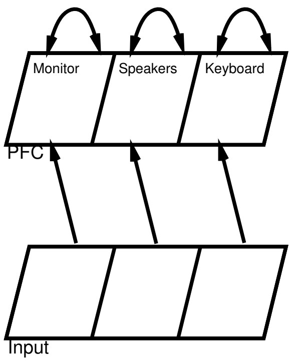


Figure 7: Network with completely isolated representations encoding features individually. This produces discrete representations with wide and robust attractor basins, and correspondingly better active maintenance.

| Trial | Input | Maint | Output |
|---|---|---|---|
| 1 | STORE-A | A | A |
| 2 | IGNORE-B | A | B |
| 3 | IGNORE-C | A | C |
| 4 | IGNORE-D | A | D |
| 5 | RECALL | A | A |

Table 1: A sequence of trials in the simple active maintenance task, showing the input (control cue and stimulus), what should be maintained in active memory, and what should be output.

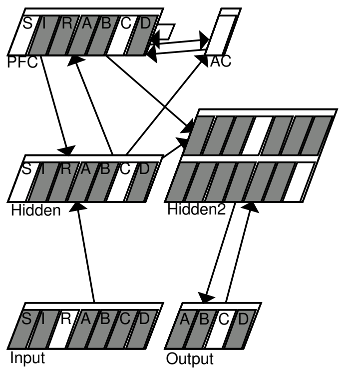


Figure 8: Dynamic updating and maintenance model. Input contains task control units (S=store, I=ignore, R=recall) and stimulus units (A-D). The output contains the stimulus units. The hidden (posterior cortex) and PFC have simple one-to-one representations of the input stimuli. The AC unit and the output hidden layer learn the significance of the cues for task performance, and how to produce the appropriate outputs.

### Simulation 3: Isolated Representations

This simulation explores the most robust configuration of the network, which is when the units are completely isolated from each other, and activation is maintained through excitatory self-connections (Figure 7). The complete isolation of the feature units from each other prevents any spread of activation between them, producing the robustness of active maintenance. When units are individually self-connected like this, their graded activations are transformed into discrete, binary activation states. Specifically, if the unit is active above a threshold (determined by a number of factors including strength of the self connection), then it activates itself strongly enough to maintain its activation over time, even in the absence of bottom-up input. This self-activation converges on a specific activation value, producing the "on" case of the two binary states. If the unit's activation is below the threshold, activation will dissipate when the input is removed, producing the "off" binary state. Thus, the width of the attractor basins in this network are enhanced by the discrete, binary character of the isolated units, which is consistent with our overall hypothesis that working memory benefits from the use of such discrete representations.

Interestingly, there is evidence that the PFC may have more isolated patterns of connectivity — neurons there appear to be interconnected within self-contained "stripe" patterns (Levitt, Lewis, Yoshioka, & Lund, 1993). Recent electrophysiological evidence further supports this notion, suggesting that the PFC is composed of small groups ("microcolumns") of iso-coding neurons, which are presumably tightly interconnected with each other (Rao, Williams, & Goldman-Rakic, 1999).

### Simulation 4: Dynamic Gating for Rapid Updating and Robust Maintenance

Even with the most robust isolated connectivity patterns, there remains the following fundamental problem: fixed levels of excitatory input weights into the simulated PFC working memory system cannot simultaneously allow information to be rapidly encoded while also protecting maintained information from the interfering effects of other inputs that should not be maintained. Rapid encoding requires relatively strong input weights, while protected maintenance requires weak ones. Thus, we suggest that a successful working memory system requires dynamic modulation of these input weights, and it appears that the dopamine neuromodulation of PFC could accomplish this (e.g., Williams & Goldman-Rakic, 1993). Furthermore, manipulations of frontal dopamine have been shown to affect working memory performance (e.g., Kimberg, D'Esposito, & Farah, 1997).

We have argued that the control of working memory updating via dopamine is synergistic with the role of dopamine in reinforcement-based learning (Cohen et al., 1996; O'Reilly et al., 1999). This learning mechanism provides a means of adaptively controlling the working memory system, which is essential to avoid the need for some kind of homunculus-like controlling mechanism. The simulation described here shows how this learning mechanism can learn to discretely update working memory based on a set of control signals that it initially knows nothing about.

The task we used to test the active maintenance mechanism involves storing a stimulus item in active memory in the face of a variable number of intervening distractor items, and then recalling the stored item. The network is provided with inputs that explicitly mark when a stimulus should be stored, ignored, or recalled (Table 1), but it does not initially know the meaning of these signals. The network (Figure 8) learns by trial-and-error — noise in the dopamine gating system enables it to store information randomly, and when it stores the stimulus identity on the store trial, it can produce the correct answer on the recall trial. This correct performance results in a reward signal, and the learning mechanism learns to associate this reward with the stimuli maintained in active memory. Because the store signal is one of the maintained stimuli, it becomes associated with reward, and thus it will tend to activate a prediction of future reward on subsequent trials. This reward-prediction activation, which is thought to correspond to a burst of dopamine, triggers the updating of working memory, and thus the storage of the stimulus information.

In summary, this simulation demonstrates the efficacy of the learned dynamic control mechanism. Because this mechanism works by discretely switching working memory between updating and maintenance, it is consistent with our overall hypothesis that working memory employs discrete representations. We plan to explore the implications of this discrete control mechanism for the development of working memory representations in future research.

## Conclusion

Working memory plays a central role in most accounts of complex cognitive function, because working memory is required in any task that involves multiple steps or a temporally extended focus of attention. It is essential to understand the nature of the representations in this system and to understand how people learn to use working memory in the service of complex cognition. Our computational investigations help advance our knowledge in this important area by exploring the consequences of various biologically-motivated factors (connectivity patterns, dopamine-modulated control) on the discreteness of working memory representations, and the resulting effectiveness of the working memory system for maintaining information in an active state over time.

## References


Braver, T. S., & Cohen, J. D. (1999). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell, & J. Driver (Eds.), *Attention and performance XVII*. Cambridge, MA: MIT Press.

Braver, T. S., Cohen, J. D., & Servan-Schreiber, D. (1995). A computational model of prefrontal cortex function. In D. S. Touretzky, G. Tesauro, & T. K. Leen (Eds.), *Advances in neural information processing systems* (pp. 141–148). Cambridge, MA: MIT Press.

Chambers, D., & Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 317–328.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society (London) B*, *351*, 1515–1527.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Dehaene, S., & Changeux, J. P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, *1*, 244–261.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*, 3088–3092.

Huttenlocher, J., Hedges, L., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*, 352–376.

Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport*, *8*, 3581–3585.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*, 161–180.

Levitt, J. B., Lewis, D. A., Yoshioka, T., & Lund, J. S. (1993). Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 & 46). *Journal of Comparative Neurology*, *338*, 360–376.

Mathis, D. A., & Mozer, M. C. (1996). Conscious and unconscious perception: A computational theory. In G. Cottrell (Ed.), *Proceedings of the 18th Annual*

*Conference of the Cognitive Science Society* (pp. 324–328). Mahwah, NJ: Lawrence Erlbaum.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*, 231–245.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Mozer, M. C. (1998). The temporal dynamics of information flow in cognition. *Paper presented at the Cognitive Neuroscience Society Annual Meeting, March 1998.*

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the $A\overline{B}$ task. *Developmental Science*, *1*, 161–184.

Munakata, Y., McClelland, J. L., Johnson, M. J., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control.* (pp. 375–411). New York: Cambridge University Press.

Rao, S. G., Williams, G. V., & Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: Evidence for microcolumnar organization in pfc. *Journal of Neurophysiology*, *81*, 1903.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, *96*, 125–144.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 5, pp. 282–317). Cambridge, MA: MIT Press.

Williams, M. S., & Goldman-Rakic, P. S. (1993). Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody. *Cerebral Cortex*, *3*, 199–222.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*, 3406–3420.