

Neural Mechanisms of Binding in the Hippocampus and Neocortex: Insights from Computational Models

Daniel M. Cer & Randall C. O'Reilly
 Department of Psychology
 University of Colorado Boulder
 345 UCB
 Boulder, CO 80309
 oreilly@psych.colorado.edu

Abstract:

An account of the neurological mechanisms that underlie binding is given which is characterized by the decomposition of the binding problem into three distinct subproblems. Each subproblem is then supported by anatomically specialized brain regions. The posterior cortex employs coarse-coded distributed representations of low-order conjunctions to resolve binding ambiguities, while also supporting systematic generalization to novel stimuli and situations. These representations are slowly acquired over experience. The hippocampus can more rapidly bind higher-order conjunctions of information such as episodes or locations. Finally, the prefrontal cortex supports transient, actively maintained bindings that are used in the service of working memory. We argue that this approach to the binding problem compares favorably with those based on temporal synchrony binding.

Introduction

Nearly all cognitive phenomena explicitly or implicitly entail some degree of binding. For instance, visual perception involves correctly binding features such as shape, color, and location of the objects currently being perceived. Similarly, auditory perception implies binding temporally extended acoustic information to facilitate the interpretation of such acoustic sequences by downstream systems as particular sounds or phones. Finally, higher level cognitive processes such as abstract reasoning and planing seem to require flexible variable/value binding whereby various operations are defined over variable like entities and are then flexibly applied to any valid values a given variable can take on.

Accordingly, the development of accurate models of the neural mechanisms underlying binding represents a critical step in the understanding of the

mechanisms that give rise to most cognitive processes. In this chapter, we discuss two distinct approaches to the binding problem. This first, one that enjoys significant popularity, is *temporal synchrony* (e.g., von der Malsburg, 1981; Gray, Engel, Konig, & Singer, 1992; Engel, Konig, Kreiter, Schillen, & Singer, 1992; Zemel, Williams, & Mozer, 1995; Hummel & Biederman, 1992). Abstractly, theories that fall under this approach solve the binding problem by proposing that when neurons that represent various features fire together, the given features are bound together. The representation of multiple sets of bindings (e.g. to represent two distinct objects) is supported by the system alternating between representing each of the appropriate sets of bindings. Accordingly, the representations for the different sets can be said to fire out of phase with each other.

While temporal synchrony does have many attractive properties, such as being relatively easy to understand, it also has several drawbacks that motivate exploring a different approach to the binding problem. The alternative approach that we will present is based on a theoretical framework that pos-

To appear in: H.D. Zimmer, A. Mecklinger, and U. Lindenberger (Eds) *Binding in Memory*, Oxford: Oxford University Press. Supported by ONR grants N00014-00-1-0246 and N00014-03-1-0428, and NIH grants MH61316, MH069597 and MH64445.

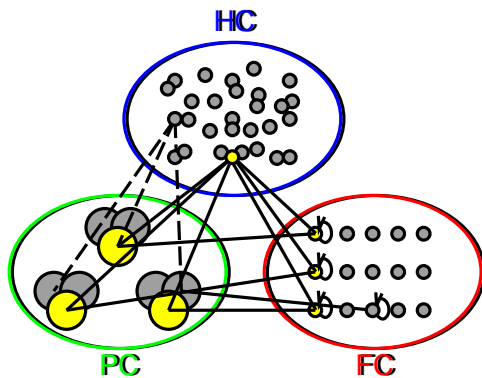


Figure 1: The components of the Specialized Neural Regions for Global Efficiency (SNRGE) framework, including the Posterior Cortex (PC), Hippocampus (HC) and Prefrontal Cortex (PFC) (motor frontal cortex constitutes a blend between PFC and PC specializations, and is included conceptually in PC). Large overlapping circles in PC represent overlapping distributed representations used to encode semantic and perceptual information. Small separated circles in HC represent sparse, pattern-separated representations used to rapidly encode (“bind”) entire patterns of information across cortex while minimizing interference. Isolated, self-connected representations in FC represent isolated stripes (columns) of neurons capable of sustained firing (i.e., active maintenance or working memory).

ulates that different regions of neural tissue are specialized to provide solutions to particular types of computational problems (O’Reilly, Busby, & Soto, 2003; O’Reilly & Norman, 2002; O’Reilly & Munakata, 2000; O’Reilly, Braver, & Cohen, 1999). We refer to this framework as the Specialized Neural Regions for Global Efficiency (SNRGE, pronounced “synergy”) framework, illustrated in Figure 1. The specializations associated with different brain areas represent computational trade-offs that are inherent in the neurobiological implementation of cognitive processes. That is, the trade-offs are a direct consequence of what computational processes can be easily implemented in the underlying biology. The specializations correspond anatomically to the hippocampus (HC), the prefrontal cortex (PFC), and all of neocortex that is posterior to prefrontal cortex (posterior cortex, PC). An overview of the computational properties and phenomena that can be associated with each of these three areas is presented next, followed by a more in-depth assess-

ment of the temporal synchrony approach. Then, we explore in more detail each of the three binding mechanisms involved in the SNRGE approach.

Posterior Cortex

Posterior cortex is heavily involved in any given cognitive task, contributing everything from sensory processing up through higher level semantic and associative processes. Indeed, it is often striking how much of cognition remains intact with frontal and hippocampal lesions. Essentially, prefrontal cortex and the hippocampus appear to serve as memory areas that dynamically and interactively support the computation that is being performed by posterior brain areas. To be able perform such computation, posterior cortex requires flexible representations that both encode semantic information about the world and facilitate efficient processing of novel information in the context of such previously learned material. These representations should also be relatively robust to both noise in the system and traumatic insult. One such class of representations that fit this criteria, and the one that will be examined here, is coarse-coded distributed representations (CCDR) (Hinton, McClelland, & Rumelhart, 1986; Wickelgren, 1969; Seidenberg & McClelland, 1989; St John & McClelland, 1990; Mozer, 1991; Mel & Fiser, 2000).

In a CCDR, neural units each encode in a graded manner a particular set of low-order conjunctions of features. For example, a unit in the visual system could be active in the presence of any one of the following: something that’s round and blue, something that’s squared shaped and red, or something that’s triangle shaped and green. Given that the conjunctions are low-order and each unit can code for multiple conjunctions, a large number of such units will generally be used to represent any given object. In other words, bindings of features to objects are represented by distributed representations over a population of neural units.

The intuition behind why CCDR are desirable is two fold. First, CCDR can allow for both efficient information processing by downstream neurons and so that the formation of the representations themselves from the input features performs a variety of useful computations. That is, in neural network

models of cognitive processes, the intermediate representations that are developed within the network's internal layers (i.e. its hidden layers), are not just an efficiently encoding of the network's inputs, but also a re-representation of such inputs that is computationally useful for the task the network was being trained on. Second, CCDR support a flexible, but compact, manner of encoding information. For instance, as will be demonstrated later in this section, CCDR allow for representations that can economically support a large number of possible binding relationships. Further, modeling has demonstrated the flexibility of CCDR in that such representations can support a great variety of computational tasks. Notably, for the purposes of this chapter, they support a very compact way of encoding binding information. However, CCDR representations are limited in that they take a substantial amount of training experience to form, and thus can not be used to rapidly encode novel information. Further, CCDR representations are driven by the current input to the system and thus can not actively maintain task relevant information unless such information is readily cued by some aspect of the environment. These limitations are directly addressed by the specializations seen in the two other regions described below.

Hippocampus

The hippocampus is known to play a critical role in the formation of episodic memories as well as the rapid encoding of novel information. Accordingly, this entails an underlying mechanism that can quickly form persistent representations that bind a large number of arbitrary pieces of information into a collective whole. Additionally, the hippocampus seems to operate as a sort of content addressable memory system. That is, a chunk of stored information is retrieved by giving the system, as a retrieval cue, some subset of the information in the chunk this is to be retrieved. For example, passing a grocery store on the way home may serve as a cue for a memory formed earlier that day of running out of milk. Computationally, this sort of retrieval behavior can be described as pattern completion. Further, a neural mechanism that accounts for the behavior of this memory system is one in which neural units with high learning

rates, i.e. the connection strength between units can change rapidly, and sparse activation across a layer are used to form large scale conjunctive representations of stimuli (e.g., O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001; O'Reilly & Norman, 2002; Marr, 1971).

As will be explored later in this chapter, the sparse conjunctive nature of hippocampal representations has the desired properties of being able to rapidly encode new memories and retrieve such memories via pattern completion. Additionally, it also maintains existing memories in a way that is highly robust to interference from the encoding of new ones. This latter property is due to the low degree of representational overlap between any two memories in the system. However, this low degree of overlap significantly limits the amount of arbitrary computation that can be done by this system since any such computation would exhibit little to no generalization. Further, the hippocampus, like posterior cortex, is largely driven by input from other systems. As such, it can not actively maintain a representation, i.e. provide as output some memory, unless the retrieval cue for the memory is continuously provided as input. Of course, the computational limitations of the hippocampus are not a problem since the CCDR of posterior cortex facilitate general information processing in the brain. Also, as will be seen below, prefrontal cortex facilitates the active maintenance of information that is not readily available/computable from the immediate information.

Prefrontal Cortex

Prefrontal cortex has long been thought to support working memory in the form of active maintenance of task relevant information (e.g., Fuster & Alexander, 1971; Kubota & Niki, 1971; Goldman-Rakic, 1987). This ability to actively maintain task relevant information is critical for the rapid adaptation to novel situations and tasks. As demonstrated by patients with PFC damage, the lack of an intact PFC leads to perseveration when such patients are trained to do one task and then are subsequently required to perform another similar but not identical task (e.g., Stuss, Levine, Alexander, Hong, Palumbo, Hamer, Murphy, & Izukawa, 2000; Wein-

berger, Berman, & Daniel, 1991; Milner, 1963). That is, these patients take far longer than normals to learn the behavior that is appropriate for the second task. A critical observation for these experiments is found in that the patients and the normal participants take a more comparable amount of time to learn the first task. Accordingly, the patients' deficit is by in large not accounted for by a learning deficit, but rather an inability to flexibly adjust one's behavior.

Computationally, a model of the PFC must not only account for the active maintenance of relevant information but also the rapid updating of this information as circumstances change (e.g., O'Reilly et al., 1999). Further, the model must account for the interaction between the PFC and other cortical areas such that the PFC can strongly bias the processing that occurs in such areas (e.g., Miller & Cohen, 2001). We suggest that rapid and transient binding of task relevant information can emerge from the biological mechanisms that support these PFC functions. Unlike the hippocampus, information is only transiently stored in the PFC. Of course, since the PFC's transient storage can be actively maintained and hippocampus' long term storage can not, these two regions serve as complementary memory systems.

Summary of Binding in the SNRGE Model

As outlined above, the SNRGE approach entails partitioning the binding problem into three distinct subproblems. The first involves how binding occurs in long term semantic memory and how such a binding mechanism can facilitate processing of novel stimuli in the context of existing knowledge. The second involves how separate aspects of an experience are bound in order to form a single episodic memory. This second subproblem also includes how novel information is rapidly learned. Such learning necessarily requires binding the individual components of the learned information together. Finally, the third subproblem involves how task relevant information is bound with such bindings being actively maintained by the system. The motivation behind this decomposition is found both in the empirically observed functional specialization of the corresponding three brain areas, and in

the theoretical observation that computational specialization can alleviate tensions that would exist in a mechanism that tries to 'do it all'. This latter observation is particularly critical when the medium used to implement the computational system poses significant constraints on the solution space.

Temporal Synchrony and its Limitations

As described in the introduction, temporal synchrony is a popular way of accounting for how the brain flexibly performs binding. To review, the temporal synchrony account of binding is that when populations of neurons that represent various features fire together, those features are considered bound together. If the system needs to simultaneously represent multiple distinct sets of bindings, it alternates between representing each set of bindings. For instance, take the case where the system is asked to represent three objects. To do this, first all of the neurons that represent features of the first object would simultaneously, or nearly simultaneously, fire. Subsequently, all the neurons representing the features of the second object would fire. Then, the same would happen for the third. Finally, after representing the third object, the system would loop and represent the first object again. Accordingly, the simultaneous binding of the features for each of the three objects would be represented in the system by the neural representation for each the distinct set of bindings firing out of phase with the other representations.

A more concrete example is shown in Figure 2. Here, there are two objects, a blue square and a red triangle. Further, the observer has four neuronal units. For our purposes, it doesn't matter if these units represent individual neurons or populations of neurons. Perception of the two objects would be represented by oscillation between the red and the triangle unit firing together and then the square and the blue unit firing together. As such, the time course of firing serves to disambiguate what feature gets bound to what object.

Part of the appeal of temporal synchrony is that it appears to trivially solve the binding problem. Additionally, it appears to offer a general process that can account for all instances of binding during any

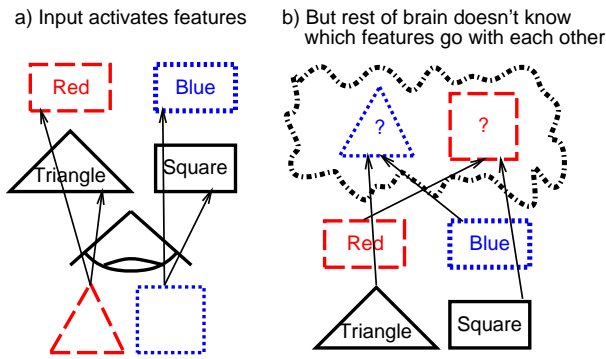


Figure 2: Illustration of the binding problem. a) Visual inputs (red triangle, blue square) activate separate representations of color and shape properties. b) However, just the mere activation of these features does not distinguish for the rest of the brain the alternative scenario of a blue triangle and a red square. Red is indicated by dashed outline and blue by a dotted outline.

given cognitive process. Accordingly, as a unitary mechanism, it also initially seems like a very parsimonious account of binding. Further, we do not have any issue with admitting the simultaneous firing of populations of neurons is important for facilitating binding. However, there are significant issues inherent in the proposition that simultaneously representing different sets of bindings is primarily done via oscillation between the representations for each set of bindings. Specifically, three primary criticisms of this mechanism addressed below are the transience of the binding relationships, the difficulties inherent in downstream systems decoding a set of objects encoded via out of phase firing, and the apparent fragility of this mechanism.

Transience

Initially the transience of the bindings associated with temporal synchrony could be seen as a positive. That is, with this mechanism computation involving being various features together need not induce additional structure within the system. Since, in any given day, people are likely involved in a large number of cognitive processes that collectively require billions if not trillions of elementary binding operations, it would seem that transient bindings are a good thing as they significantly lower the storage requirements of the system. However, the flip side of this is that once a stimulus is removed there is

no memory of it. Therefore, returning to Figure 2, if the red square and blue triangle were removed from the subject's field of vision, he or she would have no memory of ever seeing the two objects. Of course, proponents of temporal synchrony have taken steps to explain how a lasting trace could be generated from within a temporal synchrony framework. These account for such long term traces by postulating a complementary memory system that can form persistent representations of sets of features previously bound together via synchronized firing.

For example, Hummel and Holyoak (1997) propose that such a memory system operates by forming a simple conjunctive representation of the features that are to be bound together. While this initially may seem like a workable solution, there is a significant body of empirical evidence that all experiences leave a lasting trace in the brain (cite people). Accordingly, following the approach proposed by Hummel and Holyoak (1997), a conjunctive representation would need to be formed for all items that were ever represented by the system. Even over the course of very short period of time, this could result in the system requiring a very large number of units to sort all of its conjunctive memories. Further, one of the most significant criticisms of systems that don't use temporal synchrony has been an intuition that such systems require an enormous number of units in order to arbitrarily bind any non-trivial set of features. As will be analytically demonstrated below, this criticism of such alternative models is in principle unfounded.

If a temporal synchrony model does use an efficient encoding system that can statically represent a large number of binding relationships, the question arises as to what is additionally to be gained by postulating that bindings are done through temporal synchrony. Proponents of temporal synchrony would point out that this modeling framework provides an unmatched level of systematicity in its representations (Hummel & Holyoak, 2003). Further, such systematicity is critical for generalization. Nonetheless, as will be discussed later, models using coarse-coded distributed representations do exhibit a promising degree of generalization. Temporal synchrony also complicates various computa-

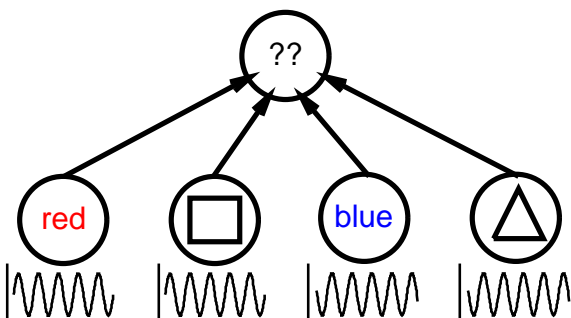


Figure 3: Decoding problem for temporal synchrony. Two sets of features are each firing in phase with each other, and out of phase with the other set (as indicated by the sine wave plots below the features). Without additional mechanisms, it is unclear how a downstream neuron can decode this information to determine what is actually present: it is being uniformly driven by synaptic input at all phases, and its activation would be the same for any combination of synchrony in the input features. One solution is to build in preferential weights for one set of features (e.g., “red square”) but this amounts to a conjunctive representation, which the temporal synchrony approach is designed to avoid in the first place.

tional processes, and it is not clear if the benefits added systematicity are worth the cost of this added complexity.

Decoding by Downstream Systems

Returning to Figure 2, imagine that the subject was involved in a task that required integrating information about both objects that he or she was presented with. For instance, if a red triangle and a blue square are presented, the subject must push a button on their left. However, if a blue triangle and a red square are presented the subject should push the button on their right. Finally, if any other pair of objects is presented (e.g. two red triangles), the subject should do nothing at all. Since success at this task entails both binding features to objects and then in some downstream process computing some function over multiple distinct bindings, it is not immediately obvious what is the best way to represent this process within a temporal synchrony framework.

Specifically, a system that simply combines the two representations so that all of the relevant features can be simultaneously presented to the downstream system will not work. In this case, the down-

stream system will not be able to distinguish between the case where it is presented with a red triangle with a blue square and the case where it is presented with a blue triangle and a red square (Figure 3). An alternative workable approach would be to intelligently integrate the individual representations of the two objects into a combined, non-temporally extended representation that maintains the appropriate binding information. But, once again, if the system must create such a representation in order to facilitate downstream processing, there is a question as to why temporal synchrony must be used in the first place. Why not just go directly to the unified representation? Of course, there are other possible solutions to facilitating downstream processing in a system based on temporal synchrony. But, such solutions involve sophisticated cognitive machinery. This machinery not only detracts from the apparent elegance of the temporal synchrony framework, but it also raises significant questions as to how learning would occur in such systems.

Fragility

The temporal synchrony approach to binding would require a precise timing mechanism that could coordinate the out of phase representation of different sets of bindings. Further, such precise timing would be dependant upon the the exact firing behavior of individual neurons. If this timing system were to fail, the results could be catastrophic. Features involved in different sets of bindings would be randomly mixed to create new sets of bindings that make little or no sense. Accordingly, it would seem that any system based on temporal synchrony would be necessarily fragile in that any small perturbation in timing would cause serious problems for the system. This runs counter to the observation that brain is both rather robust to insult and to interference with normal neurological processes by psychoactive agents such as alcohol. Specifically, under such conditions not only does performance degrade gracefully but there is also no selective early failure of the binding system. Finally, electrophysiological recordings strongly suggest that the brain is relatively noisy environment. Therefore, it seems like supporting temporal synchrony in such a con-

text would be necessarily very difficult. Further, the electrophysiological recordings that do support the notion of temporal synchrony only emerge after averaging over many trials. As such, they may just be artifacts of some other neurological process.

As will be illustrated in the following section the weakness of temporal synchrony are actually the strengths of the SNRGE approach. That is, through the CCDR of posterior cortex, and the large scale conjunctive representations of the hippocampus, the system naturally supports non-transient bindings and accounts for all experiences to leave some trace in the computational machinery. Also, unlike temporal synchrony, simultaneously representing multiple sets of bindings is assumed to be represented as a single unity representation. Accordingly, there is no difficulty in downstream systems decoding the set bindings for use in some cognitive task that depends on two or more sets of binding relationships. Additionally, since there is no need for a precise timing mechanism, the system intuitively would seem more robust than one based on temporal synchrony. This intuition is supported by evidence from computation simulations whereby systems based on the representations we are advocating degrade gracefully for injury or insult (e.g., O'Reilly & Munakata, 2000).

Coarse-Coded Distributed Representations of Low-Order Conjunctions

One trivial solution to the binding problem is to use conjunctive representations to represent each binding that the system needs to perform. For example, returning to Figure 2, there would be a particular unit that codes for a blue square and another that codes for a red triangle. While it is intuitively easy to understand how such conjunctive representations solve the binding problem, they are intractable because they produce a combinatorial explosion in the number of units required to code for all possible bindings as the number of features to be bound increases. As an example, assume that all objects in the world can be described by 32 different dimensions (e.g., shape, size, color, etc), each of which contains 16 different feature values. To encode all possible bindings using the naive approach, 16^{32} ,

		RC GS						
obj1	obj2	R	G	B	S	C	T	BT
RS	GC	1	1	0	1	1	0	0
RC	GS	1	1	0	1	1	0	1
RS	GT	1	1	0	1	0	1	0
RT	GS	1	1	0	1	0	1	1
RS	BC	1	0	1	1	1	0	0
RC	BS	1	0	1	1	1	0	1
RS	BT	1	0	1	1	0	1	1
RT	BS	1	0	1	1	0	1	0
RC	GT	1	1	0	0	1	1	1
RT	GC	1	1	0	0	1	1	0
RC	BT	1	0	1	0	1	1	1
RT	BC	1	0	1	0	1	1	0
GS	BC	0	1	1	1	1	0	1
GC	BS	0	1	1	1	1	0	0
GS	BT	0	1	1	1	0	1	1
GT	BS	0	1	1	1	0	1	0
GC	BT	0	1	1	0	1	1	1
GT	BC	0	1	1	0	1	1	0

Table 1: Solution to the binding problem by using representations that encode combinations of input features (i.e., color and shape), but achieve greater efficiency by representing multiple such combinations. Obj1 and obj2 show the features of the two objects. The first six columns show the responses of a set of representations that encode the separate color and shape features: R = Red, G = Green, B = Blue, S = Square, C = Circle, T = Triangle. Using only these separate features causes the binding problem: observe that the two configurations in each pair are equivalent according to the separate feature representation. The final unit encodes a combination of the three different conjunctions shown at the top of the column, and this is enough to disambiguate the otherwise equivalent representations.

or 3.5×10^{38} units would be needed. If the system needed to bind features for 4 objects simultaneously, 4 times as many units would be needed. Of course, the brain binds many more types of features and does so with far less units. This combinatorial explosion problem for simple conjunctive representations is an important reason why they have been largely ignored as a solution to the binding problem.

However, there are far more efficient ways of implementing conjunctive representations that we show below have very modest lower bounds in terms of the number of units required to encode a large number of possible bindings. The effi-

cient conjunctive encoding we advocate is coarse coded distributed representations (CCDR) of low-order conjunctions. As described earlier, we believe this type of representation is used in posterior cortex to facilitate efficient binding in the immediate service of information processing. To review, in CCDR each unit can code in a graded fashion for multiple low level conjunctions, and thereby achieve much greater efficiency. Table 1 shows a simple example of this kind of representation, from O’Reilly and Munakata (2000). Here, localist units are used to either encode one of three shape features or one of three color features. Clearly, the use of these localist units alone does not allow the system to bind pairs of features together (e.g. blue binds with square, red binds with triangle, and circle binds with green). However, by only adding one additional unit that codes for 3 low level conjunctions, the system now has a unique representation for each possible binding of the feature types to three objects.

Analysis of Efficient Conjunctive Binding Representations

Here, we present some results that demonstrate the high level of representational efficiency that can in principle be obtained by coarse-coded distributed representations (CCDR). The key insight is that we can represent the efficiency of a distributed representation, which comes from representing a large number of possibilities using different combinations of a much smaller number of units, using an optimal binary encoding of bits. Thus, the number of bits required to encode all the different binding combinations gives an optimal lower-bound estimate for the number of binary thresholded units that would be required to distinguish the different binding cases. This lower bound does not account for whether a fixed set of neural weights could actually achieve the necessary pattern of firing required for such a maximally efficient representation. In addition, it does not take into account the kind of graded activations that are more consistent with the typical description of CCDR (which are typically much more efficient than binary units). Nevertheless, this analysis provides an easily calculated lower bound that makes it clear that the combinatorial explosion issue for conjunctive binding representations should not

pose a problem for coarse-coded distributed representations.

To parameterize the analysis, we consider D sets of mutually exclusive feature dimensions (e.g., shape, color, size, etc). Each feature dimension has a number of features F (e.g., for the shape dimension: square, triangle, circle, etc). One could easily consider different numbers of features per dimension, but this is not necessary for a basic analysis. Also, the feature set can contain a null element that represents no feature from the given set being bound to a given object. The system can represent (bind) N different items composed of these dimensions and features at a time.

Using this notation, the number of ways that features from each dimension can be bound to N objects is given by:

$$(F^D)^N \quad (1)$$

This is a very large number even for small values of F , D and N . However, taking the \log_2 of this quantity produces a much smaller number, which reflects the number of bits (i.e., binary thresholded units) required to represent each possible set of bindings. This simplifies to the following expression:

$$\text{min bits} = ND \log_2 F \quad (2)$$

Note that this expression is linear in the number of objects N and dimensions D , and even more efficient as the number of features per dimension F increases.

As an example of this efficiency, we return to the example given earlier regarding a system that must be able to represent all arbitrary bindings of 32 different dimensions (i.e., $D = 32$), each of which has 16 distinct features ($F_i = 16$), to 4 separate objects ($N = 4$). As noted earlier, a simple conjunctive encoding for such a system would require 1.36×10^9 units. However, using an optimal binary distributed representation, 512 units would be required. Again, the actual number of units required for an actual graded neural network encoding will likely be different, but should be roughly of the same order, and nowhere near as many as the simple conjunctive encoding.

Tuple Binding and Combinatorial Generalization

As a complement to the above analytical results, the remainder of this section focuses on empirical results for models that make use of CCDR. These results will further demonstrate both that CCDR can efficiently represent binding information, and that such representations can be learned by a model. Furthermore, we focus on the generalization performance of these models (i.e., their ability to process novel inputs in a systematic manner consistent with training), which has been raised as an important problem for CCDR networks as contrasted with temporal synchrony models (e.g., Hummel & Holyoak, 2003). Indeed, some would argue that generalization is a greater problem than that of capacity. In this respect, the arguments from temporal synchrony advocates strongly resemble those leveled at neural networks from the perspective of traditional symbolic cognitive models (e.g., Pinker & Prince, 1988; Fodor & Pylyshyn, 1988). This makes sense given that many extant temporal synchrony models can be characterized as essentially more elaborate implementations of these traditional symbolic architectures, particularly in their ability to leverage arbitrary symbol binding for producing systematic behavior. Furthermore, these temporal synchrony models suffer many of the same limitations as earlier symbolic models, particularly with respect to the difficulty of incorporating powerful learning mechanisms that can develop new knowledge and processing representations from initially undifferentiated neural tissue.

Therefore, to show that an approach based on CCDR offers a competitive alternative to temporal synchrony models, it is vital to demonstrate that they generalize sufficiently well. We discuss a number of generalization tests with relatively generic posterior cortex models employing CCDR below, which demonstrate that indeed these representations are capable of high levels of generalization in the context of tasks that have extensive binding demands. For other related results, see also Edelman and Intrator (2003). We then return to these issues in the section on prefrontal cortex, where we discuss recent results showing how rule-like representations in prefrontal cortex, learned in the context of

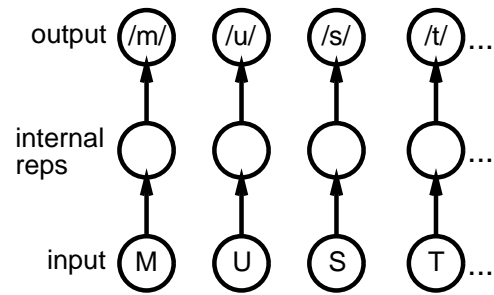


Figure 4: Illustration of the N-tuple or combinatorial generalization task, in the more naturalistic context of systematically pronouncing letter inputs. In the combinatorial extreme represented by this task, each letter position is mapped to a corresponding output in a way that does not depend on any of the other letters. Thus, the most efficient solution is to develop internal representations that encode each input/output tuple mapping separately. Clearly, this level of combinatoriality is too extreme in the case of letter pronunciation, but it nevertheless serves as a convenient benchmark task.

task performance, can promote even more systematic behavior in neural networks (Rougier, Noelle, Braver, Cohen, & O'Reilly, submitted).

Tuple Reordering Task

We have explored a variation of a widely explored test of generalization in neural networks called either the N-tuple combinatorial generalization task (Brousse & Smolensky, 1989; Phillips & Wiles, 1993; O'Reilly, 2001) (see Figure 4 for an illustration). Although the basic task does not require much in the way of binding, we were able to extend it to do so. This task has the form of a simple auto-associative network where the input is a tuple of N items and the target output is the same tuple of N items. Early work with this task suggested that feed-forward backpropagation neural networks could not adequately generalize on this task, supporting the need in cognitive modeling for alternative, more systematic, architectures (Brousse & Smolensky, 1989). However, subsequent work demonstrated that such neural networks could in fact generalize well on this task (Phillips & Wiles, 1993; O'Reilly, 2001). Successful networks learned to develop separate mapping pathways between input and output tuples, as illustrated in Figure 4.

We have developed a straightforward extension

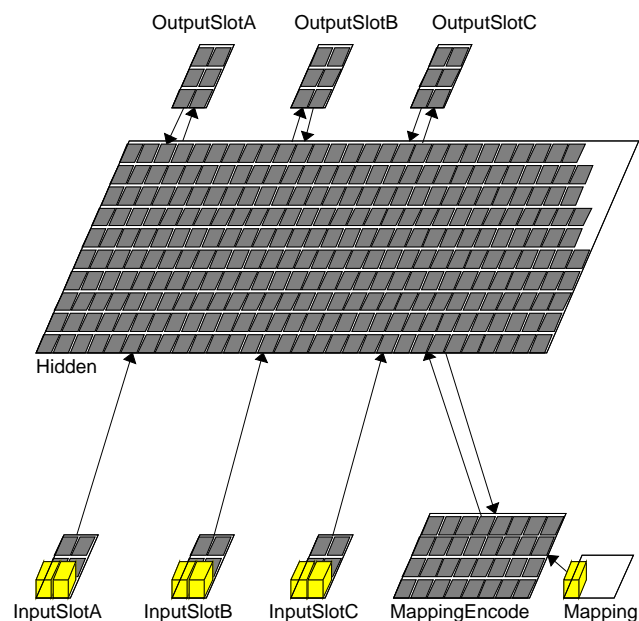


Figure 5: The tuple reordering network. Input patterns were presented across three slots, and the mapping input layer provided a transformation signal for reordering the presentation of these inputs across the three output slots. All possible

of the N tuple auto-associative task that requires a much more demanding solution, involving the binding and remapping of input features in slots (Cer & O’Reilly, in preparation) (Figure 5). Specifically, we introduced an additional input to the model that indicates how the items in the input tuple should be reordered when they are output by the model. That is, rather than mapping the i -th element of the input tuple directly to the i -th element of the output tuple, the mapping instruction given to the network will indicate the j -th element in the output that any given i -th element in the input should be mapped to. While superficially this may appear to be a trivial extension of the N tuple task, the current task represents a significant extension in terms of the computation that must be performed by the model. Specifically, note that the mapping operations that the model is instructed to perform are defined over the input and output slots. However, to perform the task the network must operate on the values represented within these slots. As such, success at this task would indicate that not only could a model bind values to variable like entities but also that the model can then systematically perform opera-

tions defined over the variables on the values held by those variables.

To explore how successful a neural network that performed binding via CCDR would be at this task, we constructed the model shown in Figure 5. Both input and output slots consist of 6 units, with the values represented within these slots by activating 2 of the 6 units. While this representation allows the slots to take on up to 15 different values, only 6 values were used in our experiment. The mapping instructions are presented using a localist representation. That is, each of the 6 units in the mapping layer corresponds to one of the 6 possible ways that three slots can be rearrange. Finally, the hidden layer consists of 300 hundred units.

The model is implemented in the Leabra framework (O’Reilly & Munakata, 2000; O’Reilly, 1998), which includes a biologically plausible form of error backpropagation, Hebbian learning, and inhibitory competition. The inhibitory competition and Hebbian learning in Leabra have been shown to produce improved generalization relative to plain backpropagation in a range of different tasks (O’Reilly, 2001; O’Reilly & Busby, 2002). Standard Leabra parameters were used in all parts of the network, except for in the mapping encode layer where the $kwta$ percentage was decreased from 0.25 to 0.20. Biologically, this roughly corresponds to increasing the degree of lateral inhibition in this layer.

The model’s training set consisted of 432 examples of how to perform the tuple mapping operation. This training set represents 33% of the total problem space. The test set consisted of 138 randomly selected tuple mappings that did not occur in the training set. After training for 25 epochs (where each epoch consisted of presenting every item in the training set once), the model was able to obtain perfect performance on the training set. Further, at this point, out of the entire test set the network only makes 2 errors. That is, it’s generalization performance as measured by the test set is 98.7%.

These results seem to indicate that CCDR can facilitate binding values to variable like entities and then perform operations defined over the variables on the appropriate values. Again, the good generalization performance of the model indicates that it

has not simply learned some degenerate associative mapping such as a holistic mapping of various patterns presented on the input layers to certain patterns of activity on the output layer. Rather, since the network is able to generalize well, its representations must have largely captured the abstract computational operation that was being asked of the network.

Spatial Relationship Binding Model

A more sophisticated binding task was explored by O'Reilly and Busby (2002), where by a network was trained to encode and report a number of relationships between items that were presented on its inputs. The model, shown in Figure 6, roughly represents a simplified model of the early visual system. During training the model is presented with a pair of input items in a simulated visual field and one of four corresponding questions. Two of the questions, "what" and "where", only required the network to report information pertaining to one of the two objects. For the what question, the *location* layer was used as an additional input to the network indicating the location of the input item it was being asked about. In response to this question, the model was trained to present in the *object* layer the item at the given location in the input field. In the case of the where question, the *object* layer acted as an additional input, and the network was trained to output in the *location* layer the input position corresponding item presented in the *object* layer. The two remaining questions require the network to identify relationships between the two items presented in the input layer. The relation-obj question is similar to the what question in that the *location* layer is used as an input that indicates the location in the input field of the object the network should output in the *object* layer. But, for this question the network must also output in the *relation* layer the relative relationship between the queried item and the other item presented in the input field. Similarly the relation-loc question is like the where question in that the *object* layer is used as an input that indicates the identity of the item in the input field whose location the network should output in the *location* layer. Like the relation-obj question, the network must also identify the relative location of the other item presented

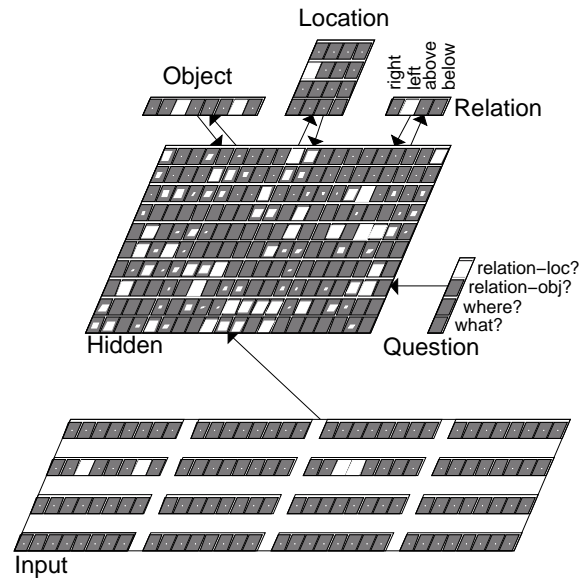


Figure 6: Spatial relationship binding model of O'Reilly & Busby (2002). Objects are represented by distributed patterns of activation over 8 feature values in each location, with the input containing a 4x4 array of object locations. Input patterns contain two different objects, arranged either vertically or horizontally. The network answers different questions about the inputs based on the activation of the Question input layer. For the "what?" question, the location of one of the objects is activated as an input in the Location layer, and the network must produce the correct object features for the object in that location. For the "where?" question, the object features for one of the objects are activated in the Object layer, and the network must produce the correct location activation for that object. For the "relation-obj?" question, the object features for one object are activated, and the network must activate the relationship between this object and the other object, in addition to activating the location for this object. For the "relation-loc?" question, the location of one of the objects is activated, and the network must activate the relationship between this object and the other object, in addition to activating the object features for this object (this is the example shown in the network, responding that the target object is to the left of the other object). Thus, the hidden layer must have bound object, location, and relationship information in its encoding of the input.

in the input field relative to the queried item and report this information in the *relation* layer.

Like the model described in the last section, the current model was implemented as a recurrent neural network. In addition to a Leabra im-

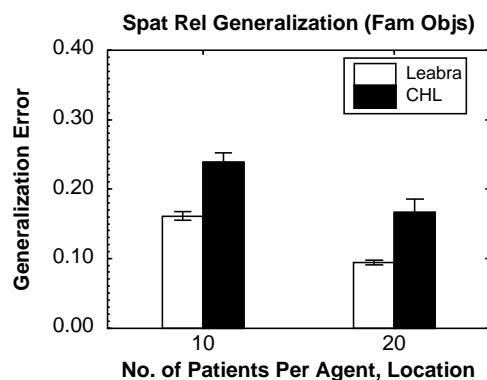


Figure 7: Generalization results for different algorithms on the spatial relationship binding task (testing on familiar objects in novel locations; similar results hold for novel objects as well). Only the 400 Agent, Location \times 10 or 20 Patient, Location cases are shown. It is clear that Leabra performed roughly twice as well as the CHL algorithm, consistent with earlier results on other tasks (O’Reilly, 2001).

plementation, a model using only contrastive Hebbian (CHL) error-driven learning, and another using the Almedia-Pineda recurrent backpropagation algorithm, were also run. Of these, it was found that Almedia-Pineda was not able to learn to successfully perform the task. While both the Leabra and CHL networks were able to learn, the additional constraints in Leabra (Hebbian learning and inhibitory competition) produced nearly twice as good generalization as CHL (Figure 7).

These experiments demonstrate both that coarse coded distributed representations can systematically perform binding relationships, and that not all mechanisms for developing such relationships are equivalent. Specifically, by incorporating additional, biologically motivated constraints on the development of internal representations in the network, the Leabra model is able to achieve more systematicity in its representations, which subsequently give rise to better generalization performance.

Language Surface Form Transformations

The models presented above strongly suggest that coarse-coded distributed representations (CCDR) can not only bind various features together, but also that CCDR also facilitates systematic oper-

ations over these bindings. However, the tasks given so far were designed explicitly to be rather pure investigations of binding. As such, the possibility remains that such tasks biased the learning that occurred in the network in such a way that the binding performance of the system was exaggerated over and above what can be typically expected of similar networks applied to more realistic cognitive tasks. That is, it could be possible that a more complex task would make it harder for the network to identify the abstract computational process that it is being asked to perform and thus bias the system toward finding degenerate solutions.

In order to further explore this issue, we constructed a task involving sentence surface form transformations (Cer & O’Reilly, in preparation). Specifically, the task involves giving a network a sentence one word at a time during the encoding part of the task. Then, during the decoding part of the task, the network is either asked to repeat back the same sentence it was given during encoding, or some transformation of it. The transformations we selected were turning an active sentence to a passive, or transforming a passive to an active. Further, when asked to perform a transformation, the model is not told explicitly what sort of transformation it should perform (i.e., active to passive vs. passive to active). Rather, during decoding, it is just told whether or not it should transform the sentence. This complicates the task, because the network must condition its transformation on the type of the current sentence.

The linguistic environment used for this task was a simple English like grammar that supports the two constructions given below:

- Active construction: [Det] [Noun] [Verb] [Det] [Noun]
- Passive construction: [Det] [Noun] was [Verb] by [Det] [Noun]

For the experiment reported here, this simple language included 32 nouns, 8 verbs, and 2 determiners. All verbs had the same form in the active and passive constructions. Additionally, when such sentences are given to the model, they are wrapped in begin and end of sentence markers.

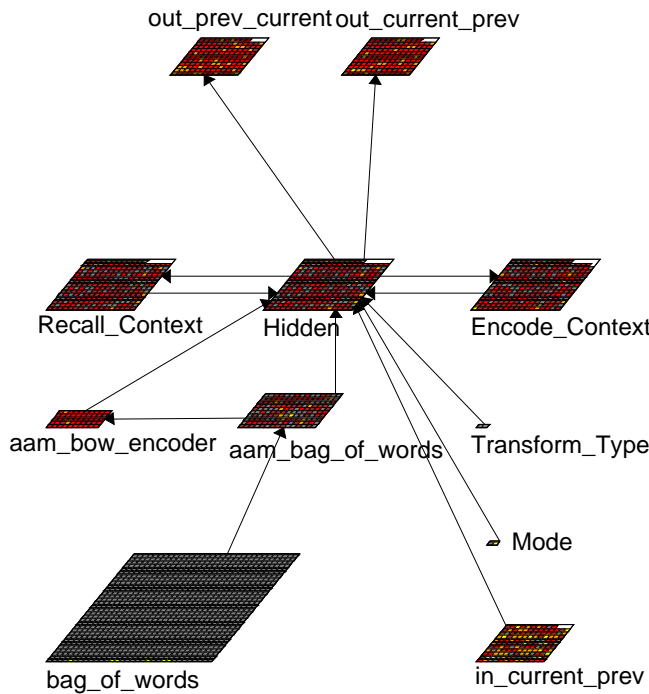


Figure 8: Language transformation network. See text for details on the role of each layer and the nature of the input/output patterns used.

Notably, this task requires binding words into some structure that represents their relative positions in the sentence. Additionally, the network must be able to flexibly extract information from this structure such that both the active and the passive form of a sentence can be reconstructed. Further, since the task is more computationally complex than those described above, the added complexity should make it so that it is not trivially easy to identify that a good solution to the task involves such binding.

The network that was trained to perform this task is illustrated in Figure 8. This network was originally developed to examine some psycholinguistic phenomena, although it was easily adapted for the task at hand. During encoding, words are presented one at a time in the layer labeled *in_current_prev*. For each word that is presented as input, the network is trained to produce the same word in the *out_current_prev* layer. Additionally, the network must reproduce the word that was presented immediately before the current word in the *out_prev_current* layer. The representations of

words used within these three layers correspond to distributed representations developed during training using FGREP (Forming Global Representations with Extended back- Propagation; Miikkulainen, 1993). After the presentation of each word, the activation values for the units in the hidden layer are copied to the units in the *encode_context* layer. Also, after each word is presented to the network, the unit in the “bag of words” layer that corresponds to that word is activated (if it is not already active from a previous presentation of the same word in the current sentence). Once activated, the units in the bag of words layer remain active for the rest of the encoding and decoding process for the current sentence. Finally, the fact that a sentence is being presented for encoding is cued by the activation of the first unit in the mode layer.

During decoding, the second unit in the mode layer is activated. Additionally, if the network should reproduce the exact sentence it was given during encoding, the first unit in the transformation layer is activated. If the network should transform the encoded sentence, the second unit in the transformation layer is activated. The first item presented in the *in_current_prev* layer during decoding is the beginning of sentence marker. In response to this, the network must reproduce the beginning of sentence marker in the *out_current_prev* layer, and the first word from the appropriate form of the sentence in the *out_prev_current* layer. Similarly, during the next time step, the network is given the first word in the sentence as input in the *in_current_prev* layer and must produce the second word in the sentence in the *out_prev_current* layer. Note that the type of sentence production scheme used here is similar to the constrained production paradigm given in Rohde (2002). Between each time step, the activations of the units in the hidden layer are copied over to the units in the *decode_context* layer. Also, note that during decoding the *encode_context* layer is frozen to what ever the last pattern of activation was in the hidden layer at the end of the encoding process.

The hidden layer and the two context layers each have 250 units. The three FGREP layers each have 140 units. The bag of words layer has 1024 units, although only 46 of this are used for the given task. The AAM bad of words layer has 150 units.

Notably the connections from the *bag_of_words* input layer and the *AAM_bag_of_words* layer are pre-trained in an autoencoder network over all representations that bag of words layer can take on in the training set. These connections are then fixed during the training of the larger network. This pre-training allows for a stable CCDR distributed representation of each pattern that is presented in the bag of words layer (which would otherwise use localist representations all words). The *bag_of_words_encoder* layer has 50 units. Finally, model was implemented as a standard backpropagation network. A learning rate of 0.1 and momentum of 0.9 were used.

The network was trained on 4000 transformations, this representing 6.1% of the total problem space. Testing was done using 100 randomly selected transformations of sentences that did not occur in the training set. Network performance was evaluated by scoring the representations produced by the network during decoding. Accordingly, scoring was restricted to the network's ability to construct/re-construct the appropriate surface form of the previously presented sentence. The representations were scored by identifying the word whose representation most closely matched the representation produced by the network as measure by the Euclidean distance between the two.

After training, the network was able to obtain 84.2% generalization performance over the test set. While this is not close to the perfect generalization performance of the previously presented models, it is still relatively good performance given the dramatically more difficult task. Accordingly, these results suggest the network was able to form coarse-coded distributed representations that overall served to perform the binding necessary to encode the sequential order of the words in the sentence, and then systematically transform them during decoding.

As demonstrated in the three models we've presented and in the analytical results, CCDR's represent an efficient means of encoding binding relationships. However, such CCDR representations both take a substantial amount of time to develop and are always driven by inputs from other systems. Accordingly, they do not account for how people rapidly form episodic memories or learn new material such that a large number of arbitrary features are

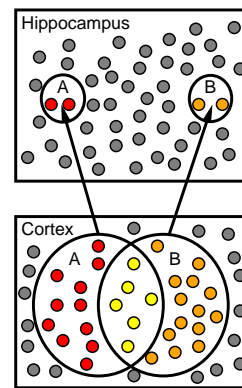


Figure 9: Sparse representations in the hippocampus relative to cortex leads to pattern separation (less probability of activation patterns overlapping, as is evident in the figure), and to units in hippocampus representing larger conjunctions of features in the cortex. This means that hippocampus performs higher-order conjunctive binding relative to cortex.

durably bound together. CCDR also do not account for how people actively maintain bindings that are not directly driven by the immediate environment. As will be shown below, these two variants of the binding problem are addressed by the hippocampus and the prefrontal cortex, respectively.

Hippocampal Conjunctive Binding

The role of the hippocampus in binding can be contrasted with the posterior cortex models just discussed along several dimensions. First, the hippocampus has sparser activity levels than the posterior cortex (roughly 5% to less than 1% in different regions of the hippocampus, compared to roughly 15-25% for cortex). These sparse hippocampal representations cause units to only respond to specific patterns of activity across the cortex (illustrated in Figure 9). Therefore the hippocampal representations encode more specific *high-order* conjunctions of many features, which contrasts with the relatively *low-order* conjunctions (i.e., conjoining relatively few features) in the posterior cortical representations.

Thus, the hippocampal units more uniquely encode specific events, while the cortical units encode smaller, recurring subsets of events. Therefore, the cortical representations support similarity-

based generalization to novel situations, whereas the hippocampal representations are better able to avoid interference between similar events, especially when rapid learning is required to encode fleeting episodes. The details of the hippocampal models have been published in a number of papers, and so are not repeated here (Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001; O'Reilly & Munakata, 2000). These models have demonstrated the ability to explain a wide range of data from animal and human experiments.

In one example, experiments with rats and corresponding models showed that the hippocampus is essential for rapidly binding together the stimulus features that define an environment (Rudy, Barrientos, & O'Reilly, 2002). In the experiments, the rats were repeatedly transported in a distinctive black ice bucket to a *preexposure* environment. Then, the rats were brought in this bucket into a different *conditioning* environment, where they were shocked. One day later, the rats were transported in a distinct cage to either the original preexposure environment or the conditioning environment. We found that rats expressed fear conditioning (freezing behavior) to the preexposure environment, and not the conditioning environment.

We interpreted this result as reflecting the rapid binding of the preexposure environment features, together with the bucket, into a conjunctive hippocampal representation. This representation was re-activated by the bucket cue just prior to the conditioning, causing the rat to associate the shock with a memory of the preexposure environment, instead of the actual environment in which it was shocked. This interpretation was supported by a computational model, and confirmed by hippocampal lesions in the rats, which abolished the fear responding to the preexposure environment.

In summary, the results reviewed here, and many more like them, suggest that the hippocampus is specialized for rapidly binding together the features or elements of episodes and environments. The resulting conjunctive representations are distinctive from those in cortex by virtue of being highly specific (i.e., higher-order), in contrast to the low-order conjunctive representations found in cortex. Both types of representations have their costs and bene-

fits, and thus the SNRGE model suggests that different brain areas are specialized for each of these functions.

Prefrontal Cortex

As noted earlier, the prefrontal cortex is thought to be specialized for the active maintenance and updating of information, commonly referred to as working memory. This specialization has several implications for binding. Generally speaking, actively maintained information can be used to perform transient forms of binding needed only for a short time during the performance of a given task. This is in contrast with the relatively long-lasting forms of binding represented by both the low-order and high-order conjunctive representations associated with cortex and hippocampus, respectively. For example, the *phonological loop* is a working memory system that can actively maintain a short chunk of phonological (verbal) information (e.g., Baddeley, 1986; Baddeley, Gathercole, & Papagno, 1998; Burgess & Hitch, 1999; Emerson & Miyake, 2003). This actively maintained verbal information is often used by people to maintain bindings necessary for solving a given task.

An example of this form of transient, phonologically-dependent binding comes from a task studied by Miyake and Soto (in preparation). In this task, participants saw sequentially-presented colored letters one at a time on a computer display, and had to respond to *targets* of a red X or a green Y, but not to any other color-letter combination (e.g., green X's and red Y's, which were also presented). After an initial series of trials with this set of targets, the targets were switched to be a green X and a red Y. Thus, the task clearly requires binding of color and letter information, and updating of these bindings after the switch condition. Miyake and Soto (in preparation) found that if they simply had participants repeat the word "the" over and over during the task (i.e., *articulatory suppression*), it interfered significantly with performance. In contrast, performing a similar repeated motor response that did not involve the phonological system (repeated foot tapping) did not interfere (but this task did interfere at the same level as articulatory

suppression in a control visual search task, so one cannot argue that the interference was simply a matter of differential task difficulty). Miyake and Soto (in preparation) interpret this pattern of results as showing that the phonological loop supports the binding of stimulus features (e.g., participants repeatedly say to themselves “red X, green Y...”, which is supported by debriefing reports), and that the use of this phonological system for unrelated information during articulatory suppression leads to the observed performance deficits.

This transient binding function of prefrontal cortex (PFC) was simulated in a model that simulates (in a simplified fashion) several of the biological specializations associated with the PFC (O’Reilly & Soto, 2002). Specifically, this model included separate “stripes” (columns) in the PFC, with each stripe receiving a separate updating signal from a simulated basal ganglia system (Frank, Loughry, & O’Reilly, 2001; O’Reilly & Munakata, 2000). These features enabled the PFC to use dedicated stripes for each sequential position in a stream of phonemes. This is possible because phonemes are a small, closed class, and thus each stripe can have representations for all possible phonemes. Taken together, these specializations enable this phonological loop model to maintain arbitrary phonological sequences in active memory. Furthermore, the network exhibited high levels of generalization after training on a small subset (10%) of possible phonological sequences. Thus, this model suggests how the PFC can maintain arbitrary bindings in a phonological code; these phonological sequences will then impact semantically associated representations throughout cortex to support task-appropriate processing (Miller & Cohen, 2001; Cohen, Dunbar, & McClelland, 1990).

Another critical contribution of the PFC for supporting task-relevant processing in a flexible, generalizable manner comes from abstract, rule-like representations that can develop through an interaction between biological specializations of the PFC and broad experience across different tasks. Rougier et al. (submitted) developed a model that addresses this fundamental question: How is information represented in PFC and, critically, how does this develop? We showed that PFC-specific mechanisms

interact with the breadth of training experience to produce abstract, rule-like representations that support generalization of performance in novel task circumstances. We also showed that these rule-like representations support patterns of performance characteristic of neurologically intact and frontally-damaged people on benchmark tasks of cognitive control (Stroop and WCST). Although the rule-like representations that developed in the model of PFC support flexible cognitive control, they did so in a way that is fundamentally different from symbolic representations characteristic of more traditional unified theories of cognition. Therefore, these results bear on both the organization and development of PFC at the neurobiological level, as well as debates regarding the nature of cognitive flexibility and rule-like behavior at the psychological level. Specifically, this model demonstrates that systematic generalization of the sort emphasized by symbolic approaches (e.g., Pinker & Prince, 1988; Fodor & Pylyshyn, 1988; Hummel & Holyoak, 2003) can emerge from biological specializations in neural network models.

Conclusions

In summary, we have presented a range of computational models based on the biological specializations associated with different brain areas, that support a range of different contributions to binding. The posterior cortex can learn coarse-coded distributed representations of low-order conjunctions, which can efficiently and systematically bind information in the service of many different forms of cortical information processing. However, these representations are learned slowly over experience; in contrast, the hippocampus is specialized for rapidly binding novel information into high-order conjunctive representations (e.g., of episodes or locations). Finally, the prefrontal cortex can actively maintain dynamic bindings in working memory, and, through more abstract rule-like representations, support more flexible generalization of behavior across novel task contexts. Taken together, we believe this overall biologically-based cognitive architecture represents a more plausible framework for understanding binding than that provided by

temporal synchrony approaches.

References

- Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*, 158.
- Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.
- Brousse, O., & Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. *Proceedings of the Eleventh Annual Cognitive Science Society Conference* (pp. 26–33). Cognitive Science Society, Hillsdale, NJ: Erlbaum.
- Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.
- Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.
- Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science*, *27*, 73–109.
- Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*, 148–168.
- Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences*, *15*(6), 218–226.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.
- Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience*, *1*, 137–160.
- Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*, 652–654.
- Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handbook of Physiology — The Nervous System*, *5*, 373–417.
- Gray, C. M., Engel, A. K., Konig, P., & Singer, W. (1992). Synchronization of oscillatory neuronal responses in cat striate cortex — temporal properties. *Visual Neuroscience*, *8*, 337–347.
- Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 3, pp. 77–109). Cambridge, MA: MIT Press.
- Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.
- Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*, 337–347.
- Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, *262*, 23–81.
- Mel, B. A., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation*, *12*, 731–762.
- Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. Cambridge, MA: MIT Press.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

- Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology*, *9*, 90–100.
- Miyake, A., & Soto, R. (in preparation). The role of the phonological loop in executive control.
- Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach*. Cambridge, MA: MIT Press.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*, 611–646.
- 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. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199–1242.
- 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.
- O'Reilly, R. C., & Busby, R. S. (2002). Generalizable relational binding from coarse-coded distributed representations. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., Busby, R. S., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation* (pp. 168–192). Oxford: Oxford University Press.
- O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *6*, 505–510.
- O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.
- O'Reilly, R. C., & Soto, R. (2002). A model of the phonological loop: Generalization and binding. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.
- Phillips, S., & Wiles, J. (1993). Exponential generalizations from a polynomial number of examples in a combinatorial domain. *Proceedings of the IJCNN*. IJCNN93.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.
- Rohde, D. L. T. (2002). *A connectionist model of sentence comprehension and production*. PhD thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.
- Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (submitted). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols.
- Rudy, J. W., Barrientos, R. M., & O'Reilly, R. C. (2002). Hippocampal formation supports conditioning to memory of a context. *Behavioral Neuroscience*, *116*, 530–538.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
- St John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sen-

- tence comprehension. *Artificial Intelligence*, 46, 217–257.
- Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., Murphy, K. J., & Izukawa, D. (2000). Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, 38, 388–402.
- von der Malsburg, C. (1981). The correlation theory of brain function. MPI Biophysical Chemistry, Internal Report 81-2. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks, II (1994)*. Berlin: Springer.
- Weinberger, D. R., Berman, K. F., & Daniel, D. G. (1991). Prefrontal cortex dysfunction in schizophrenia. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 276–285). New York: Oxford University Press.
- Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76, 1–15.
- Zemel, R. S., Williams, C. K., & Mozer, M. C. (1995). Lending direction to neural networks. *Neural Networks*, 8, 503.