Developmental and Computational Neuroscience Approaches to Cognition: The Case of Generalization

Yuko Munakata and Randall C. O'Reilly

Department of Psychology
University of Colorado Boulder
345 UCB
Boulder, CO 80309
munakata@psych.colorado.edu, oreilly@psych.colorado.edu

19th February 2003

Self-introductions: Yuko Munakata and Randall C. O'Reilly are Associate Professors of Cognitive Psychology at the University of Colorado, Boulder. They have co-authored a textbook on neural networks: "Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating Brain" (2000, MIT Press).

Munakata received an A.B. in Psychology and B.S. in Symbolic Systems from Stanford University in 1991, and an M.S. and Ph.D. in Psychology from Carnegie Mellon University in 1993 and 1996. Her research investigates memory development and the nature of representations leading to behavioral dissociations, through neural network modeling and behavioral testing of infants and children. She serves on a grant review panel of the National Institute of Health, and is a recipient of the Boyd McCandless Award for early contributions to developmental Psychology.

O'Reilly received an A.B. in Psychology from Harvard University in 1989, and an M.S. and Ph.D. in Psychology from Carnegie Mellon University in 1992 and 1996. His research focuses on the specialization of function in and interactions between the hippocampus, prefrontal cortex, and posterior neocortex in learning, memory, and controlled processing. He is an Associate Editor of "Cognitive Science" and a Reviewing Editor for "Hippocampus," and is a recipient of an Excellence in Teaching award.

Abstract

The ability to generalize — to abstract regularities from our experiences that can be applied to new experiences — is fundamental to human cognition and our abilities to flexibly adapt to changing situations. However, the generalization abilities of children and adults are far from perfect, with many clear demonstrations of failures to generalize in situations that would otherwise appear to lend themselves to generalization. It seems that people require extensive experience with a domain to demonstrate good generalization, and that their generalization abilities are best when dealing with relatively concrete, familiar situations. In this paper, we argue that people's successes and failures in generalization are well characterized by neural network models. Networks of neurons connected by synaptic weights are naturally predisposed to encode information in a highly specific fashion, which does not support generalization (as has been seized upon by critics of such models). However, with sufficient experience and appropriate architectural properties, such models can develop abstract representations that support good generalization. Implications for the neural bases and development of generalization abilities are discussed.

Figure 1: A grocery store sign in Wanaka, New Zealand.

A grocery store in Wanaka, New Zealand provides the store's hours on the front door (Figure 1). This sign amused us when we came upon it last year, because of course, these pieces of information could easily be compressed into something like: "Open daily 9:00 am – 8:00 pm." This kind of generalization would be much quicker to process, and could be more easily applied to know when the store would be open on any given day. Being able to generalize in this way is fundamental to human cognition and flexibility. (If the store example seems more amusing than compelling, consider a slightly more urgent case of needing to generalize across individual times when predators are at the watering hole.) Humans can be relatively skilled generalizers, though as the sign demonstrates, we are not 100% reliable.

Our abilities to generalize are evident in our daily lives. We can generalize based on our prior experiences to deal with novel instances of people we meet, stores we shop in, cars we rent, games we learn, and so on. Rather than treating each of these novel instances as completely foreign, we can abstract regularities from our prior experiences to expect certain things from the new instances — that we might exchange names in meeting the new person, use a key to start the rental car, and so on.

However, our generalization abilities also show limitations. For example, when we learn new information (e.g., in a classroom lecture), we tend to automatically encode the irrelevant background context (e.g., the shape and color of the room), such that our ability to recall this information is better if tested in the same context than in a different context (Godden & Baddeley, 1975). In general, people seem to require extensive experience with a given domain to exhibit good generalization, such that children early in development or adults in the early stages of learning a novel domain will tend to not generalize very well. Furthermore, generalization is limited to relatively concrete, familiar domains — we do not do particularly well at purely abstract reasoning.

These features of human generalization abilities are well captured by neural network models. Knowledge is represented in a neural network model by synaptic connection strengths (*weights*) between simulated neurons (*units*). In a simple toy model of learning new information in a classroom setting (Figure 2), the *input* units representing the external input to a person might encode things like the visual features of the environment, and individual words or concepts. To encode a new fact, an internal *hidden* unit would increase its weights to all of the input units that are active at a given point in time (this is known as *Hebbian* learning; Hebb, 1949). Thus, the irrelevant context gets bound up with the relevant content, and this produces the context effects mentioned above (knowledge is better when tested in the same context, because these context inputs support the activation of the hidden units that encoded the knowledge). More generally, knowledge in a neural network tends to be encoded in a highly specific manner, because it always involves making specific connections among specific neural units. Thus, the responses of a neural network tend to depend on the specifics of the input patterns, and therefore they tend to not generalize to novel input patterns.

So how does a neural network generalize? Going back to the classroom learning toy model, one way to make the knowledge less context dependent and specific is to learn about the same relevant content in a variety of different contexts. This will have the effect of continuing to strengthen the weights from the relevant content units, while decreasing the weights to the irrelevant context units (Figure 2b). Thus, as should be intuitively clear, as we repeatedly encounter information over many experiences, it becomes "decontextualized" and therefore more easily accessed in any context (i.e., it generalizes to more contexts). Interestingly, this transition from context-specific "episodic" memory to decontextualized "semantic" knowledge may also correspond to a difference in the relative importance of different brain

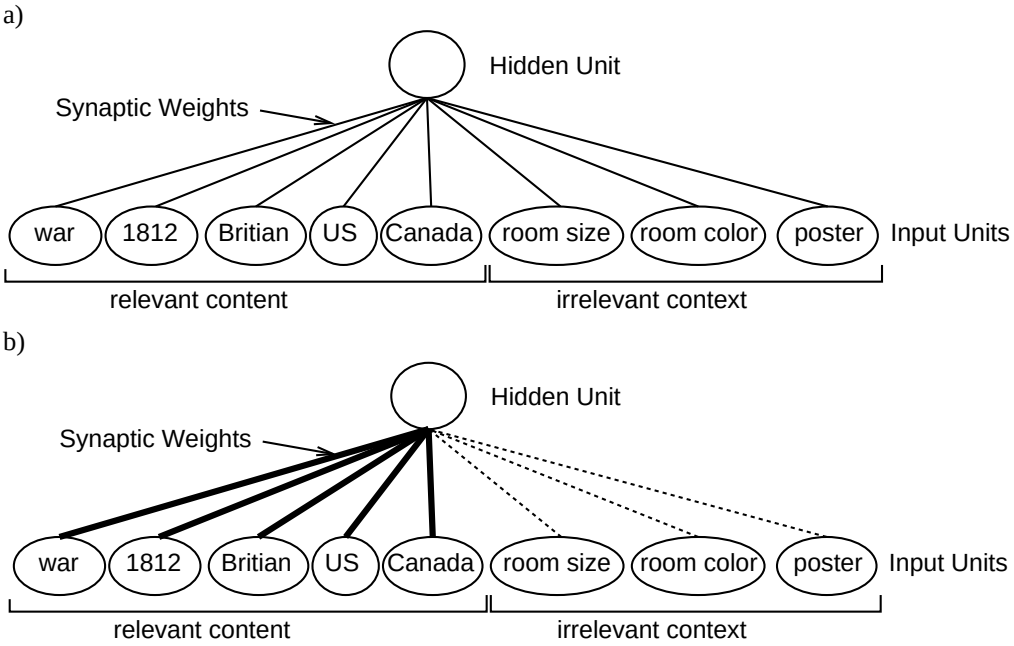

Figure 2: Toy example of how neural network learning can be highly context-sensitive and thus not support good generalization. **a)** A hidden unit encodes all of the information present during a classroom lecture on the war of 1812. This includes the irrelevant context of the room, etc. This predicts that knowledge will be better when tested in the same room, which is true for people. **b)** If the same knowledge is encountered in a variety of different contexts, the relevant content will continue to be strengthened, but the irrelevant context will be weakened.

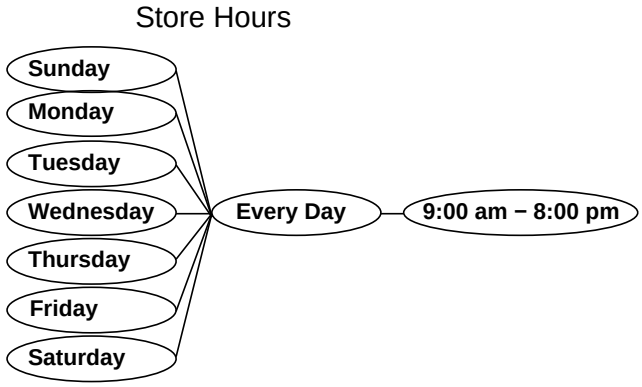

Figure 3: Introducing the abstract representation (category) of "every day" produces a more generalized encoding of the store hours.

areas supporting this knowledge (from the hippocampus to the cortex; e.g., McClelland, McNaughton, & O'Reilly, 1995).

Another way of thinking about generalization in a neural network is in terms of the development of *abstract* representations. In the Wanaka grocery store example (Figure 3), a suitable abstraction would be the notion of "every day," which groups together the individual days and allows them to be represented under a single representational unit (e.g., a *chunk* or *categorical* representation). The decontextualization of knowledge can also be considered a form of categorization or chunking — the collection of relevant content that repeats over time gets bound together as a reliable "chunk" of knowledge, and becomes dissociated from other specific details like the relevant content. Thus, the key to understanding human generalization from the neural network perspective is understanding the process of developing suitable abstract representations.

It is clear that the process of developing abstractions requires experience. In the process of decontextualization, one must experience the relevant content information across a variety of different contexts. Similarly, to develop an abstract category representation, one must integrate across a range of different experiences. For example, learning the category of "dog" requires experiencing a number of different dogs and learning that there is something in common between a Poodle and a German Shepard, despite their numerous obvious perceptual differences. After this category has been acquired, it can support generalizations (e.g., if you learn that a German Shepard eats dog food, you can surmise that this is true of Poodle's as well) — these generalizations might not be obvious given the raw perceptual differences among members of the dog category (e.g., you might otherwise be tempted to feed a Poodle cat food).

Therefore, the strong prediction from neural network models, which appears to be true of human cognition, is that people generalize best in domains that they have had considerable experience with. Novel situations that are difficult to map into existing abstract categorical representations (e.g., a particularly funny looking type of dog that one has never encountered before) will result in poorer generalization than novel situations that are more easily mapped onto existing knowledge. Developmentally, we expect generalization abilities to improve with the acquisition of relevant experience and, critically, with the development of suitable abstract representations (which can be assessed independent of generalization behavior).

Lest these predictions appear so obvious and intuitive as to be uninteresting, it is important to contrast the neural network models to symbolic models that represent knowledge in terms of abstract rules. Proponents of such models tout their immediate and perfect abilities to generalize, while criticizing neural network models' limited generalization abilities (e.g., Marcus, 1998; Pinker & Prince, 1988). However, we think that such models suffer an excess of generalization riches — they considerably overestimate people's actual generalization abilities. Furthermore, they do not provide natural mechanisms for understanding the progression of generalization abilities over development and over learning experiences.

In what follows, we briefly discuss generalization successes and failures in children and adults, and then in neural network models.

Generalization in Children and Adults

Generalization abilities have been documented systematically across domains and ages. As one example, as children learn words for objects, they generalize certain properties of words to novel instances. If they know that the plural form of "cat" is "cats," the plural form of "ball" is "balls," and so on, they can generalize this knowledge to know that more than one "wug" would be referred to as "wugs" (Berko58). Similarly, if children know that the shape of objects is critical for defining the words that name the objects, they can generalize this knowledge to expect that new words are defined through the shape of objects (e.g., Landau, Smith, & Jones, 1988).

However, limitations in generalization abilities can be particularly salient early in development. For example, learning in infants and children can be heavily tied to the particulars of their learning experiences, rather than being generalizable across different situations. Here we consider two such examples. First, infants can learn to kick to move a mobile hanging above them; the mobile is connected to the infant's leg with a string (Rovee-Collier, 1997). Rather than learning the general lesson that kicking moves mobiles, however, infants appear to learn about the specific testing situation. In many cases, they will not kick in the same way when the testing situation is changed, such as when alterations are made to objects in the mobile (e.g. Greco, Hayne, & Rovee-Collier, 1990) or even to incidental features of the environment such as the color of a bumper on the crib (Borovsky & Rovee-Collier, 1990). Similarly, older infants and toddlers will show deferred imitation of actions they have observed; after watching an adult model an action with an object, they will reproduce those actions when given the object at a later time (Meltzoff, 1985, 1988; Barr, Dowden, & Hayne, 1996). However, again the learning can be heavily tied to the particulars of the experience. In many cases, infants and toddlers will fail to imitate the observed actions when the testing situation is changed, such as when alterations are made to the test objects (e.g., a mouse puppet versus a rabbit puppet) or even to incidental aspects of the environment (e.g., whether they are testing in the home versus laboratory) (Barnat, Klein, & Meltzoff, 1996; Hayne, Boniface, & Barr, 2000).

As children get older, they generally show more generalizable learning, which is less tied to the particulars of the testing situation. For example, changes in the testing environment disrupted the deferred imitation performance of 6-month-olds, but not of 12- and 18-month-olds. Changes in the test object disrupted the performance of 6- and 12-month-olds, but not of 18-month-olds (Hayne et al., 2000). More dramatic changes to the test objects disrupted the

performance of 18-month-olds, but not of 21-month-olds (Hayne, MacDonald, & Barr, 1997).

Generalization in infants has also been studied in the statistical learning paradigm. When presented with only a couple minutes worth of stimuli that follow a regular pattern, infants, adults, and other species are able to abstract the statistical regularities of the input (Fiser & Aslin, 2001; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996). For example, after two minutes of exposure to continuous speech sounds (e.g., "ba ti lo gu la mu..."), which followed regular patterns (e.g., "ba" was always followed by "ti"), 8-month old infants distinguished speech streams that showed the same statistical regularities from speech streams that did not (Saffran et al., 1996). This behavior did not reflect a simple preference for familiar stimuli. When tested with two sets of equally familiar stimuli, with one set better matching the conditional probabilities present in the training stimuli, infants still preferred the set better matching the statistical regularities in the input (Aslin, Saffran, & Newport, 1998).

Stronger tests of generalization in the statistical learning paradigm have been conducted, by testing with novel phoneme stimuli that were not presented during training. When presented with continuous speech sounds (e.g., "ba ti ti go la la"), which followed a general pattern (e.g., a single phoneme was always followed by a double — an "ABB" pattern), 7-month-old infants abstracted this general statistical regularity and applied it to new stimuli (Marcus, Vijayan, & Vishton, 1999). These authors argued that this reflected a fully general abstraction of the "syntactic" rules of these phoneme sequences, and therefore that neural network models (which don't tend to do this) are insufficient to model human cognition. However, other work has demonstrated that, consistent with the broader human generalization literature, there are important limitations on the abstractions that infants form in these statistical learning paradigms (e.g., Gomez, Gerken, & Schvaneveldt, 2000; Gomez, 2002). For example, generalization in the Gomez (2002) study depended critically on the amount of variability infants were exposed to in an independently varying portion of a grammar — only with higher levels of variability could they generalize that this part was independent of its surrounding elements. This is strongly consistent with the decontextualization mechanisms characteristic of neural network models. The importance of varied experience in supporting increased generalization has also been documented in a variety of other paradigms (e.g., Gentner & Namy, 1999; Greco et al., 1990; Needham, Dueker, & Lockhead, 2002).

Adults also show limitations in their generalization abilities. For example, after studying lists of words, adults' memory for those words can be tied to the particulars of their learning experience. They have more difficulty remembering following changes to incidental aspects of the testing situation, such as the environment context (e.g., Godden & Baddeley, 1975). Thus, what is learned is not a completely abstract, generalizable kind of knowledge; rather, it is tied to the particulars of how the knowledge was acquired.

Similarly, adults can fail to abstract rules that would help them generalize to novel instances. For example, when we learn how to solve a puzzle (e.g., the widely studied tower-of-Hanoi problem involving moving different-sized disks stacked on three poles), we have difficulty transferring (generalizing) the learned strategy to a new puzzle with different *surface features* (e.g., using monsters and caves instead of disks and poles) that nevertheless has exactly the same solution (e.g., Newell & Simon, 1972). Similarly, consider a problem in which an army of a certain size is required to win a battle; however, the entire army cannot go together to the battle site because no single route to the site has the capacity for them. Adults can fail to see the solution to this problem (break the army up into separate groups and send them to the battle site using multiple routes), despite being presented with an analogous problem and its solution. For example, radiation of a certain dose is required to kill a tumor, however, the entire does of radiation cannot be directed to the tumor site because it will kill too much of the tissue in the path. The solution is to send multiple smaller doses of radiation through a number of paths. Generalizing this knowledge would support a solution to the analogical problem, but adults often fail to do so (Glick & Holyoak, 1987).

Finally, even when adults appear to successfully abstract knowledge that would allow them to generalize to novel instances, in many cases this abstracted knowledge is still sensitive to the particulars of testing, indicating that it is not fully generalizable. For example, in the Wason card sorting task, adults are presented with a set of individual cards, each with a number on one side and a letter on the other side. They are asked which cards need to be flipped over to determine whether a given rule is true of the cards (e.g., that all cards with a "P" on one side have a "4" on the other). The correct answer is that all cards with a "P" and all cards with a number other than "4" must be flipped over; the common error is to flip over cards with the number 4. Adults can solve real-world versions of this task (e.g., when enforcing the alcohol drinking age law "if drinking, must be over 21", people have no hesitation in checking if people who are not over 21 are drinking) more readily than the abstract version (Cheng & Holyoak, 1985; Johnson-Laird, Legrenzi, & Legrenzi, 1972), indicating that their knowledge of these logical rules is tied to particulars and not fully generalizable.

In these ways, generalization abilities can improve with development, but still show limitations in adulthood. Thus, in the endeavor of exploring human cognition by simulating it, the target is *not* a system that can generalize perfectly. Failures to generalize in and of themselves do not challenge models of cognition. (On the flip side, perfect generalization ability would challenge a model of cognition). In the next section, we consider neural network models' abilities to generalize.

Neural Network Models of Generalization

Although relatively few of the generalization phenomena discussed above have been specifically modeled, we discuss here a selection of neural network models that demonstrate the basic principles governing human generalization abilities. We begin with two models that demonstrate generalization based on decontextualization of representations, and development of abstract representations, respectively. Then, we discuss other models that have been applied to more specific behavioral data.

Generalization Through Decontextualization

In the introduction, we noted that neural networks tend to form associations with all information present together at a given time, including irrelevant background context. We argued that over many learning experiences, networks can learn that the relevant stuff repeatedly occurs together, while the irrelevant context does not. This is the central principle behind Hebbian learning (Hebb, 1949): things that reliably co-occur should be associated together (and things that do not, should not). The importance of Hebbian learning for generalization was demonstrated in a recent model (O'Reilly, 2001). Specifically, this model showed that adding Hebbian learning (and inhibitory competition, which is needed for Hebbian learning to work effectively) to a network using a biologically-plausible version of error-driven backpropagation learning yielded substantially better generalization than the base network alone. Furthermore, analyses showed that this improved generalization arose because the network strengthened its weights to relevant, co-occurring information, and decreased its weights to irrelevant "background" information. In addition to demonstrating the importance of the decontextualizing principle for generalization in neural networks, this model shows that the choice of learning mechanisms used to train a network can have important implications for generalization results.

Before discussing the details of the O'Reilly (2001) model, some background on neural network learning mechanisms is needed. Most existing neural network models of generalization have used the backpropagation learning mechanism (Rumelhart, Hinton, & Williams, 1986). This mechanism adapts synaptic connection weights among units according to the derivative of the global error signal — in other words, it learns to correct its errors. However, the backpropagation mechanism has been widely criticized for being biologically implausible (Crick, 1989; Zipser & Andersen, 1988). Nevertheless, it is possible to perform backpropagation learning in a biologically-plausible manner using only locally-available activation states of the sending and receiving units (O'Reilly, 1996). This biologically-plausible form of backpropagation, called *GeneRec* (generalized *recirculation*; Hinton & McClelland, 1988) requires that the network have bidirectional connectivity among all the units, such that activation flows simultaneously in both directions through the network.

The starting point for the O'Reilly (2001) model is the finding that this GeneRec network exhibits very bad levels of generalization, relative to standard (biologically-implausible) backpropagation (Figure 5). The upshot is that it would appear impossible to satisfy both biological constraints (by using GeneRec instead of backpropagation) and cognitive constraints (by having reasonable levels of generalization) in one model. However, this unfortunate conclusion is avoided by adding Hebbian learning and inhibitory competition to a GeneRec network; the result is the *Leabra* algorithm, which is motivated by a number of additional biological and cognitive considerations (O'Reilly, 1996, 1998; O'Reilly & Munakata, 2000). Thus, the conclusion from the O'Reilly (2001) paper is that the Leabra model produces good generalization while also being biologically plausible.

The specific model explored in O'Reilly (2001) is shown in Figure 4. Decontextualization-based generalization is tested in this model by having 4 separate groups of input units (called *slots*) that each have *independent* patterns of activity across them. The task the network is trained on is to identify each of these 4 independent patterns separately across a corresponding set of 4 output slots. However, the network has no way of knowing in advance that there are 4 independent input patterns — it just looks like one big input pattern to the network (and similarly for the output pattern). Therefore, over repeated experience with the input patterns, the network must learn to decontextualize its representations of each of the 4 input/output slots, so that it can process each of them independently. This decontex-

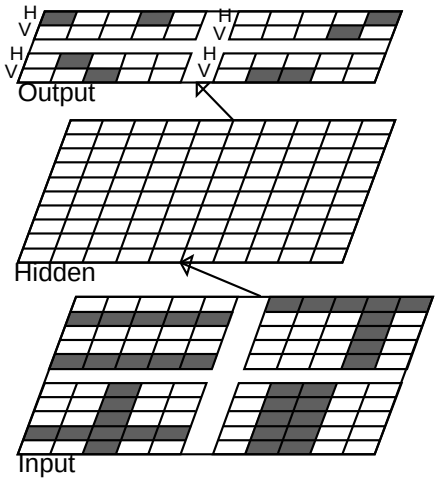

Figure 4: Architecture of the combinatorial generalization network. The input and output are composed of four slots, with the input pattern for each slot being one of 45 different possible combinations of 2 horizontal and/or vertical bars in the 5x5 slot grid, and the output pattern being a localist identification of each of the 2 lines (the first row of 5 output units for each slot representing the vertical lines, and the second row representing the horizontal lines). Each slot is independent, and thus the other slots constitute irrelevant background context for a given slot. The grouping of units according to the slots is only for display purposes — to the network, the input and output are large undifferentiated collections of units. Darkened units show an example input/output pattern.

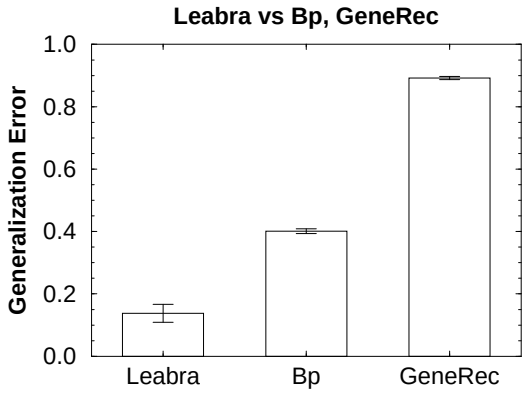

Figure 5: Generalization results comparing Leabra (GeneRec plus inhibitory competition and Hebbian learning) with standard feedforward backpropagation (Bp) and interactive GeneRec. The additional biases in Leabra result in much better generalization than even feedforward backpropagation.

tualization ability is tested in a generalization test where novel combinations of patterns across the slots are presented — if the network's representations are appropriately decontextualized, then it will treat each slot independently and perform correctly. However, if the representations are not decontextualized, and thus include significant inputs from multiple different slots, then the novel combinations of inputs across slots will confuse the network, and it will produce errors. Thus, the number of errors made during the generalization test with novel combinations of patterns across the slots is the primary measure of generalization.

Figure 5 shows the generalization performance of a standard backpropagation network (Bp), GeneRec, and the Leabra network. First, as noted earlier, the GeneRec network generalizes much worse than a standard backpropagation network. Second, the Leabra network, which is just the GeneRec network plus Hebbian learning and inhibitory competition, generalizes best of all. The reasons for both of these results can be seen in the patterns of weights that develop in the networks (Figure 6). The units in the GeneRec network (Figure 6a) are clearly not decontextualized — there are strong weights from all of the slots. In contrast, the Leabra units (Figure 6b) are almost completely decontextualized — they encode only one input/output slot, and almost completely ignore other slots. Thus, when a novel combination of patterns across the input slots is presented, the Leabra network will process each slot independently,

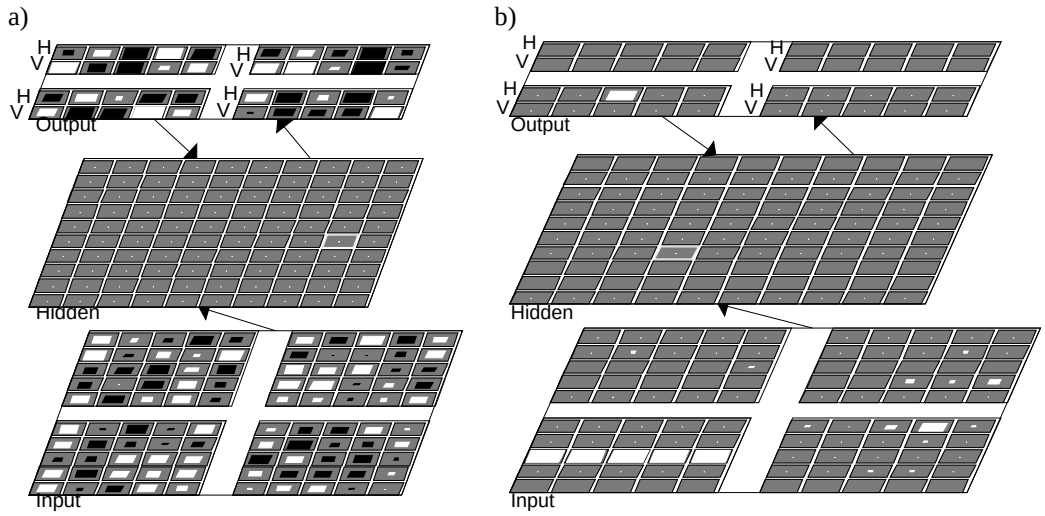

Figure 6: **a)** Weights for a typical hidden unit in the GeneRec network after training, with the size of the square indicating magnitude, and color indicating sign (white=positive, black=negative). Both feedforward weights from the input layer and feedback weights from the output layer are shown (the GeneRec weights are symmetric so these feedback weights also indicate the nature of the hidden-to-output weights; also note that there are no within-layer connections, as indicated by the zero-value grey squares). It is clear that the units are not dividing up the task into separable mappings for each slot — instead, each unit participates in multiple slot mappings. **b)** Weights for a typical hidden unit in the Leabra network after training. In contrast with the GeneRec network, it is clear that the units are dividing up the task into separable mappings for each slot — each unit participates in mapping a single line within a slot (in this case the middle horizontal line in the lower left hand slot) to its corresponding identification output.

and will not be confused by the novelty of the patterns across slots (enough patterns (100) are presented during training so that it can reliably process any of the patterns within a slot). In contrast, the units in the GeneRec network remain context sensitive, and the novel context across the other slots therefore interfere with its performance.

The one remaining question is why GeneRec performs so much worse than backpropagation, given that the backpropagation network's units look very similar to GeneRec's (not shown). The key difference is that, by virtue of its bidirectional connectivity (necessary for performing error-driven learning in a biologically-plausible manner), GeneRec is an *attractor* network, and this makes it much more sensitive to the irrelevant context information than the feedforward backpropagation network. Thus, the novel context sends the GeneRec network off into a random attractor state, while it more gently perturbs the backpropagation network. Readers familiar with the "butterfly-effect" sensitivity of dynamic systems (also known as "chaos") will recognize this hyper-sensitivity of the dynamic GeneRec network.

One additional important finding from the O'Reilly (2001) model is shown in Figure 7, which plots generalization performance as a function of training time (in epochs, where one epoch is one pass through all 100 of the training patterns). As you can see, the generalization performance of the Leabra network takes a long time to develop after it has mastered the task. During this time, Hebbian learning is slowly reshaping the network's internal representations, making them less and less context sensitive. This trajectory is consistent with gradual improvements in generalization over development (as reviewed above). Furthermore, the model predicts incubation-like effects in adults that slowly take place after initial competence is demonstrated — this might explain the kinds of improvements in cognitive flexibility that occur in the transition from novice to expert levels of experience.

Finally, it is important to quantify the generalization abilities of these networks in terms of how much generalization is produced for a given amount of training. The generalization results presented in O'Reilly (2001) demonstrate that neural networks are capable of substantial levels of generalization based on a relatively small sample of the environment (e.g., 100 training patterns out of 4.1 million possible (.0024%) resulting in an average of 3.5 million correct responses (85% correct)). However, this extremely high level of generalization depends in part on the simplicity of this particular task. Nevertheless, results should help to counter the persistent claims that neural networks are incapable of producing systematic behavior based on statistical learning of the environment (e.g., Marcus, 1998; Pinker & Prince, 1988).

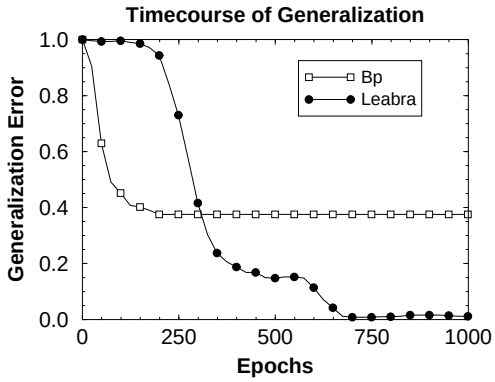

Figure 7: Timecourse of generalization in Leabra (best generalizing network) and Bp, showing that generalization performance in Leabra only starts to improve well after the task has been learned by error-driven learning (which occurs within the first 10 epochs). This identifies Hebbian learning as specifically important. In contrast, generalization in Bp begins much more rapidly, and more closely parallels the timecourse of learning (Bp takes roughly 50 epochs to learn the task).

Generalization Through Abstraction

The critical mechanism in neural network models for achieving more powerful forms of generalization beyond decontextualization is developing abstract representations (e.g., categorical representations). This was illustrated in Figure 3 in the case of generalizing the Wanaka grocery store hours. The development of abstract representations to support generalization across multiple different tasks was explored in a recent neural network model (Rougier, Noelle, Cohen, Braver, & O'Reilly, in preparation). This model incorporates a simulated prefrontal cortex, which we hypothesize to be important for developing abstract representations. The prefrontal cortex in this model is capable of maintaining information over time in an active form (i.e., activation-based working memory; Miller & Cohen, 2001; Munakata, 1998; Rougier & O'Reilly, 2002; O'Reilly, Braver, & Cohen, 1999). This working memory function can foster the development of abstract representations as described below.

Rougier et al. (in preparation) trained a range of different models on a varying number of related tasks operating on simple visual stimuli (e.g., *name* a "feature" of the stimulus along a given "dimension" such as its color, shape, or size; *match* two stimuli along one of these dimensions; *compare* the relative size of two stimuli). To test for generalization, we only trained a given task on a small percentage (e.g., 30%) of all the stimuli, and then tested that task on stimuli that were trained in other tasks.

As shown in Figure 8, the model with the full prefrontal working memory mechanisms achieved significantly higher levels of generalization than otherwise comparable models that lacked these specialized mechanisms. Furthermore, this benefit of the prefrontal mechanisms interacted with the breadth of experience the network had across a range of different tasks. Thus, the model exhibited an interesting interaction between nature (the specialized prefrontal mechanisms) and nurture (the breadth of experience): both were required to achieve high levels of generalization. We consider the protracted period of development of the prefrontal cortex (up through puberty) as the timeframe over which prefrontal representations are shaped, and the huge breadth of experience during that time then leads to what systematic reasoning abilities we have as adults.

The main reason why the prefrontal mechanisms led to such good generalization in our simple task domain is that they enabled the network to develop clean, abstract representations of the stimulus dimensions. Specifically, the network was trained such that a given stimulus dimension was relevant across a series of individual training trials. Furthermore, in some cases, the network had to "guess" what this relevant dimension was, and maintain it in the absence of direct environmental feedback over a sequence of trials. Critically, the robust activation-based working memory functions of the prefrontal model enabled the network to maintaining the same representation over time, and therefore these representations learned to abstract the dimensional information that was common across trials, while filtering out the irrelevant information that varied across these trials.

Other comparison networks that could maintain information over time, but lacked a dynamically gated working memory system (e.g., a simple recurrent network or SRN; Elman, 1990), ended up using a variable set of representations over a given dimension, and thus did not develop the appropriate abstractions. We think this pattern of results

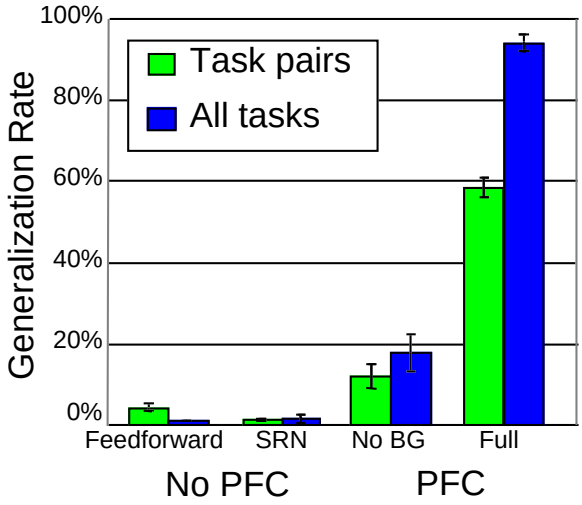

Figure 8: Results of the Rougier et al. (in preparation) cross-task generalization model on novel test patterns that were only experienced on tasks other than the testing task. As expected, generalization performance (number of correctly processed test patterns) increases when more tasks are trained (i.e., broader experience across different tasks improves the systematicity of the learned representations). The key result, however, is that this breadth of experience across tasks interacts with the presence of specialized prefrontal cortex active maintenance mechanisms, such that the full model having these specialized mechanisms generalizes much better than comparable networks without them. Thus, cognitive flexibility (generalization) depends on the interaction between experience and biological predispositions (nature *and* nurture).

reflect a general principle for why the prefrontal cortex should develop more abstract representations than posterior cortex, and thus facilitate flexible generalization to novel environments: abstraction derives from the maintenance of stable representations over time, interacting with learning mechanisms that extract commonalities over varying inputs. Supporting this view are data showing that damage to prefrontal cortex impairs abstraction abilities (e.g., Dominey & Georgieff, 1997), and that prefrontal cortex in monkeys develops more abstract category representations than those in posterior cortex (Wallis, Anderson, & Miller, 2001; Freedman, Riesenhuber, Poggio, & Miller, 2002; Nieder, Freedman, & Miller, 2002).

Generalization in Other Models

Language and Statistical Learning

As noted earlier, standard neural networks have been criticized for not developing fully abstract rules, and therefore of being incapable of simulating generalization performance in the statistical learning procedure (Marcus et al., 1999). However, a number of successful models of this data have been developed (Altmann & Dienes, 1999; Christiansen & Curtin, 1999; Gasser & Colunga, 1999; Seidenberg & Elman, 1999; Hanson & Negishi, 2002; Altmann, 2002). Some of these models demonstrate the importance of prior experience in establishing a set of more abstract representations that support subsequent generalization (Altmann, 2002; Hanson & Negishi, 2002). This is sensible — infants come into these studies with 7 months worth of experience hearing richly structured phonetic sequences, and it is likely that they have developed some more abstract representations of these sequences during this time.

Generalization has also been modeled in adult language performance. For example, several models have explored the ability to pronounce nonwords according to statistical regularities learned during training on the entire set of English monosyllabic words (Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996; O'Reilly & Munakata, 2000). When an appropriate balance between contextualized and decontextualized encoding of letter and phoneme sequences is achieved, the models generalized to the nonwords in the same way that people do.

There has been some debate about the sufficiency of neural network models to account for generalization at the level of syntactic production. For example, Marcus (1998) argued that a syntactic model developed by Elman (1991) could not solve the following generalization test. First, the model was trained with sentences such as "A rose is a rose is a rose" and "A tulip is a tulip is a tulip", and then the model was tested on its ability to correctly complete a novel sentence, such as "A blicket is a blicket is a ?" It failed to pass this test, which obviously people can easily pass.

The generalization failure of the model may have been, at least in part, an artifact of the way in which novel stimuli were presented — as the activation of units that have never before been activated. This instantiation of novelty assumes that new stimuli (e.g., the word "blicket") activate neurons that have never fired before. No evidence supports this assumption. Instead, neural patterns of firing reflect the similarity of inputs (Desimone & Ungerleider, 1989; Tanaka, 1996), suggesting that generalization to new instances would occur through the overlap between resulting patterns of firing and patterns of firing in previous experiences.

In particular, the novel word "blicket" is likely represented as a pattern of phonemes, each of which are quite familiar, even though the overall combination is novel. Thus, a more appropriate network model would be given extensive training on sequences of phoneme groups (words) appearing in different orders and combinations. When it is then tested on its ability to replicate a particular such sequence with a novel combination of phonemes, it should have little difficulty because it can rely on learned abstractions, as amply demonstrated by existing models described above.

Object permanence

Another model criticized for its generalization failures focused on infants' understanding of the permanence of objects (Munakata, McClelland, Johnson, & Siegler, 1997, see also Mareschal, Plunkett, & Harris, 1999; Schlesinger & Barto, 1999). The model was presented with objects moving back and forth in a simple environment, in which one object occasionally moved in front of and occluded another object. The model's goal was to predict on each time step what it would see on the following time step. With repeated experiences with objects becoming occluded and then reappearing, the model learned to maintain representations of the objects while they were occluded — developing some sense of "object permanence." The model's representations of hidden objects were graded in nature, becoming increasingly strong over a protracted course of development. Weak representations early in development were sufficient for the model to succeed at some tasks testing its sensitivity to hidden objects, while stronger representations were required for the model to succeed on other tasks. Thus, the model demonstrated how graded representations could contribute to behavioral dissociations of the sort often observed in infants and children (see also Morton & Munakata, 2002; Munakata, 1998, 2001). Further, the model showed some ability to generalize its knowledge of objects. When presented with novel objects (represented as novel combinations of visual features), the model demonstrated generalization in representing such novel objects as continuing to exist when occluded and predicting that they would reappear.

However, the model was criticized as a model of object permanence understanding, due to its failures to generalize its knowledge more broadly (Marcus, 2001). When the model was presented with events that were quite different from those it had been trained on, it was unable to form correct predictions about what it would see next. For example, if the model had been trained with an occluder that always moved, or that always moved in a particular manner, the model was unable to form accurate predictions about the occluder if it remained still at test, or if it moved in a novel manner. Infants' knowledge of objects and their permanence is viewed as more generalizable than this. A richer experience base might allow networks to generalize their knowledge better; on the flip side, a more restricted range of experiences might limit infants' abilities to generalize (Munakata et al., 1997). However, these issues of generalization have not been systematically tested with this network.

Structured Representations

Critics of neural network models often claim that these models cannot efficiently and systematically represent the kinds of *structured* knowledge that is needed for higher-level cognitive functions, such as analogical reasoning (e.g., Holyoak & Hummel, 2000). The key feature of structured knowledge is that it encodes relationships among different elements. For example, an analogy applies the same abstract relationship across different elements (e.g., the relationship of orbital motion of planets around the sun and orbital motion of electrons around the atomic nucleus). To test these claims, O'Reilly and Busby (2002) explored the ability of distributed representations in a neural network to efficiently encode relational information and generalize to novel cases (Figure 9). This network was capable of representing spatial relationships (above, below, right, left) among visual objects, while at the same time identifying these objects and their locations. Thus, the internal representation layer had to bind together all the relevant information about a given input "scene". It did this using coarse-coded distributed representations that encoded low-order conjunctions among all the elements of the scene. Furthermore, the model demonstrated generalization rates of 95% accuracy after only being trained on 25% of the total possible input patterns. As with other generalization tests (O'Reilly, 2001), this generalization performance was about twice as good as a comparison network using only error-driven learning.

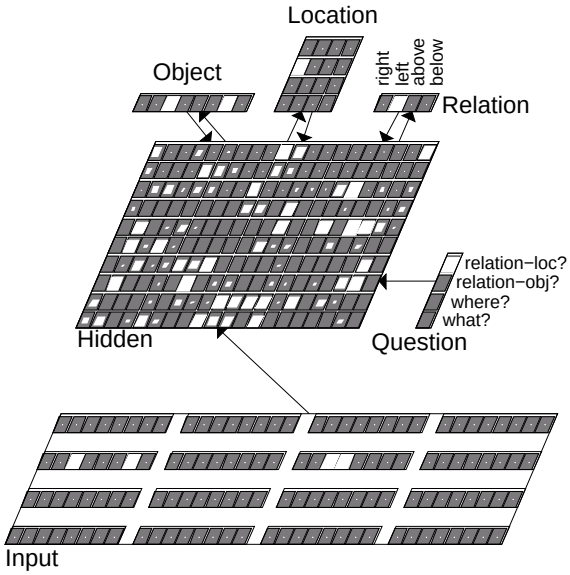

Figure 9: Relational binding model (O'Reilly & Busby, 2002). This model demonstrated the ability of distributed representations to efficiently encode spatial relationships (above, below, right, left) among input objects, and to generalize this knowledge to novel input configurations and objects. Such relational information has typically been assumed to require complex temporal synchrony binding mechanisms — instead the distributed representations in this model performed binding via coarse-coded low-order conjunctive representations.

Summary and Implications of the Models

The neural network models described above demonstrate several key points. First, they show that neural networks can in fact generalize at high levels, when given sufficient amounts of training. Although it is difficult to quantitatively compare these training and generalization levels with human data, at a qualitative level this is certainly the observed pattern: people do not generalize well in unfamiliar, novel domains, and do generalize well in familiar, overtrained domains. Second, the networks clearly elucidate the specific neural mechanisms that underlie good generalization, namely decontextualization and development of abstract representations. Furthermore, they point to the importance of specific learning mechanisms (e.g., Hebbian learning), and brain areas (e.g., prefrontal cortex) in producing good generalization. Third, they elucidate the kinds of training regimes that lead to good generalization, for example including multiple different but related tasks led to better generalization in the Rougier et al. (in preparation) model than did training on a single task. Fourth, models have demonstrated good generalization in more complex domains that require structured knowledge based on representing relationships among items. Finally, models have been applied to successfully simulate some specific human generalization performance. Clearly, more work is needed to address a broader range of empirical data to more thoroughly test the basic mechanisms.

Conclusions

As in many domains, the broad conclusion about people's ability to generalize is that it lies "somewhere in the middle" — people are not perfect nor terrible generalizers. In this paper, we have tried to convey how neural network models may contribute to an understanding of the neural mechanisms leading to successes and failures in generalization performance. The principles underlying these successes and failures are quite basic and so should be general across a range of domains, but future work remains to test their scope.

References

Altmann, G. T. (2002). Learning and development in neural networks - the importance of prior experience. *Cognition, 85*, B43–B50.

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Barnat, S., Klein, P., & Meltzoff, A. N. (1996). Deferred imitation across changes in context and object: memory and generalization. *Infant behavior and development, 19*, 241–252.

Barr, R., Dowden, A., & Hayne, H. (1996). Developmental changes in deferred imitation by 6- to 24-month-old infants. *Infant behavior and development, 19*, 159–170.

Borovsky, D., & Rovee-Collier, C. (1990). Contextual constraints on memory retrieval at six months. *Child Development, 61*, 1569–1583.

Cheng, P., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

Christiansen, M. H., & Curtin, S. (1999). Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Sciences, 3*, 289–290.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature, 337*, 129–132.

Desimone, R., & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeys. In F. Boller, & J. Grafman (Eds.), *Handbook of neuropsychology*, Vol. 2 (Chap. 14, pp. 267–299). New York: Elsevier Science Publishers B. V.

Dominey, P. F., & Georgieff, N. (1997). Schizophrenics learn surface but not abstract structure in a serial reaction time task. *Neuroreport, 8*, 2877.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 194–220.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*, 499–504.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2002). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology, 88*, 929–941.

Gasser, M., & Colunga, E. (1999). Babies, variables, and connectionist networks. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (p. 794). Mahwah, NJ: Lawrence Erlbaum.

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development, 14*, 487–513.

Glick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier, & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications*. Orlando, FL: Academic Press.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology, 66*, 325–331.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431–436.

Gomez, R. L., Gerken, L., & Schvaneveldt, R. W. (2000). The basis of tranfer in artificial grammar learning. *Memory and Cognition, 28*, 253–263.

Greco, C., Hayne, H., & Rovee-Collier, C. (1990). Roles of function, reminding, and variability in categorization by 3-month-old infants. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 617–633.

Hanson, S. J., & Negishi, M. (2002). On the emergence of rules in neural networks. *Neural Computation, 14*, 2245–2268.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton top tamarins. *Cognition, 78*, B54–B64.

Hayne, H., Boniface, J., & Barr, R. (2000). The development of declarative memory in human infants: Age-related changes in deferred imitation. *Behavioral Neuroscience, 114*, 77.

Hayne, H., MacDonald, S., & Barr, R. (1997). Developmental changes in the specificity of memory over the second year of life. *Infant Behavior & Development, 20*, 233.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems, 1987* (pp. 358–366). New York: American Institute of Physics.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich, & A. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines*. Mahwah, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, *63*, 395–400.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, 243.

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, G. F., Vijayan, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77.

Mareschal, D., Plunkett, K., & Harris, P. (1999). A computational and neuropsychological account of object-oriented behaviours in infancy. *Developmental Science*, *2*, 306–317.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

Meltzoff, A. N. (1985). Immediate and deferred imitation in 14- and 24-month-old infants. *Child Development*, *56*, 62–72.

Meltzoff, A. S. (1988). Infant imitation and memory: Nine-month-olds in immediate and deferred tests. *Child Development*, *59*, 217–225.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *24*, 167–202.

Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: A neural network model of perseveration and dissociation in early childhood. *Developmental Psychobiology*, *40*, 255–265.

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. (2001). Graded representations in behavioral dissociations. *Trends in Cognitive Sciences*, *5*(7), 309–315.

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

Needham, A., Dueker, G., & Lockhead, G. (2002). Category experience facilitates young infants' object segregation. Manuscript submitted for publication.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *298*, 1708–1711.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

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., & Munakata, Y. (2000). *Computational explorations in cogntive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Rougier, N. P., Noelle, D., Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (in preparation). The prefrontal cortex and flexible control of behavior: Cross-task generalization from systematic representations.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*, 503–520.

Rovee-Collier, C. (1997). Dissociations in infant memory: Rethinking the development of implicit and explicit memory. *Psychological Review*, *104*, 467.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 8, pp. 318–362). Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old. *Science*, *274*, 19.

Schlesinger, M., & Barto, A. (1999). Optimal control methods for simulation the perception of causality in young infants. *Proceedings of the 21st Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Seidenberg, M. S., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*, 435–436.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

Wallis, J. D., Anderson, Kathleen, C., & Miller, E. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*, 953–956.

Zipser, D., & Andersen, R. A. (1988). A back propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.