Available online at www.sciencedirect.com
SCIENCE DIRECT®
Developmental Review 24 (2004) 133–153
DEVELOPMENTAL REVIEW
www.elsevier.com/locate/dr

Computational cognitive neuroscience of early memory development☆

Yuko Munakata

Department of Psychology, University of Colorado Boulder, 345 UCB, Boulder, CO 80309, USA

Received 28 August 2003

Abstract

Numerous brain areas work in concert to subserve memory, with distinct memory functions relying differentially on distinct brain areas. For example, semantic memory relies heavily on posterior cortical regions, episodic memory on hippocampal regions, and working memory on prefrontal cortical regions. This article reviews relevant findings from computational cognitive neuroscience on why different neural regions might be specialized for different types of memory, and how this might impact early memory development. These findings demonstrate computational trade-offs among different memory functions, such that a single system cannot specialize on more than one function. Instead, the anatomical and physiological specializations of posterior cortical, hippocampal, and prefrontal cortical regions support their associated functions. This computational framework provides a mechanistic way of understanding memory distinctions described at the conceptual level. The developmental relevance of this framework is discussed—in the context of specific models, where available—for category learning, infantile amnesia and developmental amnesics, and the development of flexible behavior.

Keywords: Computational trade-offs; Hippocampus; Prefrontal cortex; Semantic; Episodic; Working memory development

Different brain areas make distinct contributions to cognition. There is broad consensus on this view, although debates on the localization of function have had a long history. Arguments for localization of function have come from a broad range of

☆ Manuscript for special issue of *Developmental Review* on The Nature and Consequences of Very Early Memory Development, Mark Howe and Mary Courage (Eds.).

E-mail address: munakata@psych.colorado.edu.

0273-2297/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.dr.2003.09.008

perspectives, including phrenologists almost two centuries ago who felt bumps on the skull in an attempt to measure underlying brain areas, researchers over a century ago working with patients with brain damage, who noted that certain behaviors and abilities seemed to depend on specific regions of cortex, and current-day neuroimagers, who record images of differential brain activity in various tasks. In contrast, other researchers have argued that the brain works according to a principle of mass action, whereby all brain areas contribute to all functions, and the effects of brain damage depend on how much (rather than which part) of the brain is damaged (e.g., Lashley, 1929). Although these debates have been largely resolved in favor of some localization of function, with numerous specialized brain areas working in concert in the service of cognition and behavior, many related debates are still ongoing. For example, how much of the neural specialization that we see in the adult reflects modular systems versus highly interactive ones (e.g., Farah, 1994; Fodor, 1983)? How much of this specialization is innately specified versus learned through experience (e.g., Hermer & Spelke, 1996; Karmiloff-Smith, 1992)?

Nonetheless, converging evidence from patients with brain damage (e.g., Farah, 1990; Scoville & Milner, 1957; Stuss & Benson, 1984), neuroimaging studies (e.g., Braver et al., 1997; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Thompson-Schill, Aguirre, & Farah, 1999), and single-cell recording data (e.g., Miller, Erickson, & Desimone, 1996; Tanaka, 1996) has led to fairly general agreement about three specializations for memory functions: posterior cortical regions for semantic memory (e.g., remembering what kind of an object a cup is, or the typical spatial arrangement of parts of a clock), hippocampal regions for episodic memory (e.g., remembering a conversation about email filing systems with certain friends in a Toronto subway), and prefrontal cortical regions for working memory (e.g., for mentally multiplying 42 times 17). This is not simply a one-to-one mapping, as other brain regions contribute to these memory functions, and other functions are subserved by these brain regions. However, these specializations appear to be important ones, which can be understood in a modeling framework in terms of computational trade-offs. Such trade-offs, and their developmental implications, are the focus of this review.

Computational framework for understanding neural specializations

A computational perspective can provide insight into how and why neural regions are specialized for different functions (reviewed in O'Reilly & Munakata, 2000). In particular, such specializations can be understood in terms of computational trade-offs, whereby two objectives cannot be achieved simultaneously. As a system specializes on its ability to achieve one objective, it must relinquish its ability to achieve another objective. For example, there is a computational trade-off between fast learning and slow learning; a system that specializes in learning rapidly is not well-suited to learning gradually and vice versa. Thus, if there are demands on a system for both fast and slow learning, these functions are likely to depend on distinct neural regions with unique specializations. Similarly, there is a computational trade-

off between representations that are highly overlapping and representations with little overlap, so that if both are desired, they too are likely to rely on specialized neural systems. These kinds of computational trade-offs, between distinct types of learning and representations, can provide insight into the specializations of posterior cortical, hippocampal, and prefrontal cortical regions. This computational approach can provide a mechanistic way of understanding distinctions between memory systems that have been characterized at a conceptual level (e.g., in terms of declarative/episodic vs. procedural memory, and implicit vs. explicit memory).

Posterior cortical areas

A standard style of neural network model may best capture the specializations of posterior cortical areas; variations on this standard model may be necessary to capture the specializations of hippocampal and prefrontal regions. Neural network models come in many flavors (Arbib, 2002), but certain features are common to many models (Fig. 1):

- Basic architectural elements: Individual processing units arranged in layers send projections to and receive projections from other processing units, allowing the communication of unit activity.
- Distributed, interactive representations: Information is represented as distributed patterns of activity across multiple processing units. These representations are highly interactive, such that activation spreads easily throughout the network.
- Slow learning: Learning occurs gradually with experience.

These features allow such models to simulate many important aspects of cognition (as emphasized in the original Parallel Distributed Processing volumes—McClelland,

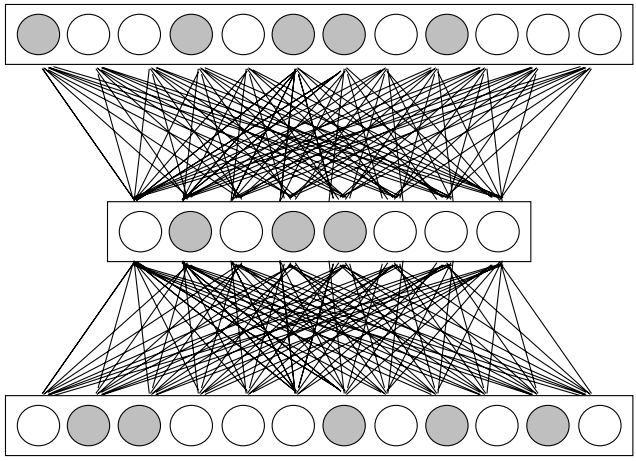

Fig. 1. Elements of one standard style of neural network model: individual processing units arranged in layers, and distributed, interactive representations. An additional aspect of standard models, slow learning, is not shown.

Rumelhart, & PDP Research Group , 1986; Rumelhart, McClelland, & PDP Research Group , 1986). For example, highly interactive representations, with numerous connections among units, allow a system to activate related representations from partial inputs. This would allow the system to go from incomplete information (e.g., a noisy image of a cat) to activate related information (e.g., what cats generally look like, what they like to eat, how they behave, and so on). Such spreading activation among interactive representations can support schemas, inferences, and semantic knowledge more generally. As we will see in the Prefrontal cortical regions section, there is a computational trade-off between this ability to represent semantic information and the ability to maintain information in working memory.

Distributed representations allow networks to encode similarity relationships among different patterns as a function of the number of units in common. This allows networks to respond appropriately to novel inputs in terms of their similarity to familiar inputs, supporting generalization (so that, e.g., one can make inferences about a cat that has never been seen before). And, distributed representations make the system more robust to damage, because there is some redundancy in the representation.

Slow learning allows networks to gradually extract statistical regularities in the environment over time. Infants, adults, and other species are quite skilled at picking up on statistical structure in their environments (Fiser & Aslin, 2001; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996). For example, after exposure to continuous speech sounds that followed regular patterns, 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).

In networks, changes to connection weights accumulate across experience, based on patterns of activity in response to environmental input. As a result, connection weights come to reflect the statistical structure in the environment. This can be demonstrated with a simple network in an environment that consists of only two stimuli (Fig. 2, O'Reilly & Munakata, 2000). All connection weights to the hidden unit begin at an uninformative .5. Stimulus 1 is present on 80% of trials; Stimulus 2 is present on the other 20% of trials. The network learns to abstract this statistical structure of its environment. The network learns according to a Hebbian learning rule, whereby "units that fire together wire together." Because Stimulus 1 is presented four times as often as Stimulus 2, the input units for Stimulus 1 have four times the opportunity to strengthen their connections to the hidden unit. As a result, the connection weights come to represent the statistical regularity of these inputs. With each epoch of training (sweep through 100 training patterns with 80 presentations of Stimulus 1 and 20 presentations of Stimulus 2), the connections from units for Stimulus 1 to the hidden unit increase until they reach .8, and the connections from units for Stimulus 2 to the hidden unit decrease until they reach .2 (Fig. 3). As we will see in the Hippocampal

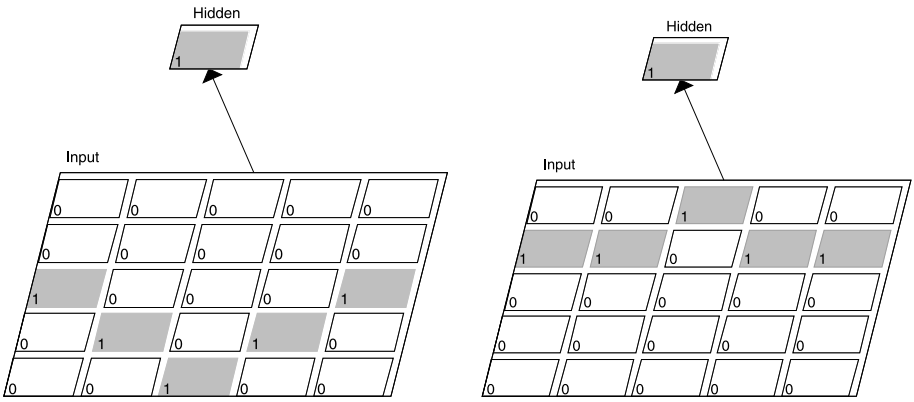

Fig. 2. Simple network in an environment with two stimuli. The stimulus on the left is present on 80% of trials; the stimulus on the right is present on the other 20% of trials.

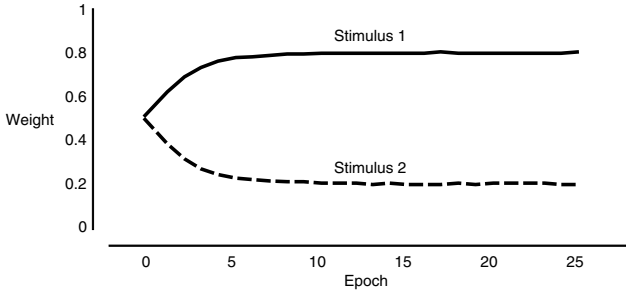

Fig. 3. Connection weights gradually learn to represent statistical structure in the environment: All weights begin at an uninformative .5. With repeated presentations of the two patterns in the environment, weights from input units representing Stimulus 1 (present 80% of the time) go to .8, and weights from input units representing Stimulus 2 (present 20% of the time) go to .2. Adapted from O'Reilly and Munakata (2000, Fig. 4.11, p. 131). Copyright 2000 MIT Press. Reprinted with permission of MIT Press Books.

regions section, there is a computational trade-off between this ability to represent statistical regularities across experiences and the ability to remember information about particular episodes.

Networks of this slow learning, distributed, and interactive representation style have been applied to a variety of developmental domains, to investigate how the gradual learning of environmental regularities could support the behaviors and representations observed. For example, such networks have demonstrated how the slow learning of statistical regularities in the visual environment can lead to the development of units that serve as edge detectors (Olshausen & Field, 1996; O'Reilly & Munakata, 2000), of the sort observed in primary visual cortex. These networks were presented with natural visual inputs. Statistical learning led the networks to represent reliable correlations in the environment (like the simple network shown in Figs.

2 and 3). Across natural images, the reliable correlations were edges (oriented transitions between dark and light). Thus, these models provide a sense of why neurons in primary visual cortex serve as edge detectors, and how they come to do so.

Networks have also been applied to understanding how infants learn the phonemes of their language and why there appear to be critical periods in this process (McClelland, Thomas, McCandliss, & Fiez, 1999; see also Guenther & Gjaja, 1996). These models were presented with inputs corresponding to phonemes, which overlapped to varying degrees. The models learned to represent the regularities in these inputs, and then represented new inputs in terms of the regularities abstracted from the previous experience. For example, if a network was trained with distinct input representations for the phonemes /l/ and /r/ (as present in many languages), it learned these inputs as distinct phonemes in the environment; as a result, the network subsequently treated phonemes that were similar to /l/ as /l/s, and phonemes that were similar to /r/ as /r/s. In contrast, if a network was trained with a blended /lr/ sound (as in Japanese), it learned this regularity as a single phoneme in the environment. As a result, the network treated subsequently encountered /l/ and /r/ sounds as the blended /lr/ phoneme, failing to distinguish them (like many native speakers of Japanese). Thus, the ''older'' models (that had been exposed to phonemes of one language) had more difficulty learning the statistical regularities of a second language than ''younger'' models exposed to the same language as a first language.

Networks have also been applied to the development of categories and semantic knowledge (Mareschal & French, 2000; Quinn & Johnson, 1997; Rogers & McClelland, in press, in press). For example, models of semantic development (Rogers & McClelland, in press, in press; building on Rumelhart & Todd, 1993) have been used to explore how children learn to categorize semantically related items (e.g., animals) together, even when these items may be perceptually quite different (e.g., a bird vs. a dog). This work demonstrated how semantic categories can be formed through the learning of statistical structure in the environment, in particular, through the patterns of *coherent covariation* across different inputs. When items have similar representations and share many properties (all animals move on their own, make sound, etc.), the properties shared by these items will be coherent and will be a strong force driving learning, because they drive changes to connection weights in the same direction. In contrast, idiosyncratic properties (e.g., the fact that some animals can fly but not swim, and others do the reverse) drive weights in conflicting directions that tend to cancel each other out early in learning. Overall, this process leads coherent properties among categories to be learned earliest. These coherent properties need not be perceptually salient (e.g., the fact that an animal can grow); as long as they covary coherently, statistical learning mechanisms can use them to guide category learning. In this way, coherent covariation of properties can lead perceptually distinct items (such as birds and dogs) to be viewed as part of the same category.

In sum, the gradual learning of statistical regularities appears to be a fairly universal process, even within the first months of life. Standard neural network models are quite good at abstracting such regularities, and so have successfully simulated a variety of aspects of development.

Hippocampal regions

However, such standard models are not particularly good at fast learning, for example, single-trial learning of the sort that humans can do (McClelland, McNaughton, & O'Reilly, 1995; McCloskey & Cohen, 1989). Such models can be altered in various ways to support fast learning, but such alterations raise the issue of computational trade-offs (O'Reilly & Munakata, 2000). For example, increasing the learning rate, which governs how quickly connection weights change in response to experience, may improve a model's single-trial learning performance. However, increasing the learning rate may impair the model's ability to gradually extract regularities across experiences.

Consider the simple network described in the previous section (Figs. 2 and 3). This slow-learning model did a good job extracting the statistical regularities present in the environment. However, it did not show very good single-trial learning. In this case, single-trial learning would consist of the model representing which of the two possible stimuli had been present on the previous trial. For this model, once the previous stimulus is removed, there is no information in the connection weights to represent which stimulus was just present. Instead, the model only represents the overall regularities in the environment—that the previous stimulus had an 80% chance of being Stimulus 1 and a 20% chance of being Stimulus 2. This would be like a person representing that 80% of the time she parks her car in one parking lot and 20% of the time she parks her car in another parking lot, but not remembering on a given day where she happened to park.

We can manipulate the network's learning rate to influence its single-trial learning ability (Fig. 4, O'Reilly & Munakata, 2000). The larger the learning rate, the more a

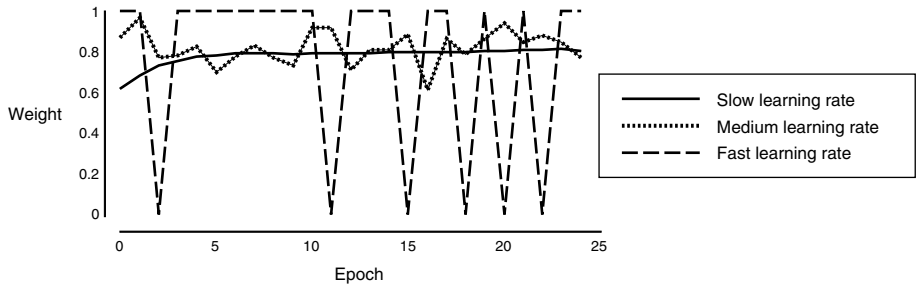

Fig. 4. Effects of learning rate on a network's ability to represent statistical regularities and perform single-trial learning: A slow learning rate produces weights from input units representing Stimulus 1 (present 80% of the time) of .8 (as shown earlier in Fig. 3). These weights capture the statistical regularities of the environment, but do not indicate whether Stimulus 1 was present on the previous trial. Increases to the learning rate produce weights that better support single-trial learning, that is, that more accurately represent which stimulus was present on the previous trial. However, there is a computational trade-off, such that as weights increasingly support single-trial learning, they represent less of the statistical regularity of the environment. The weights for Stimulus 2 show a complementary pattern, but are excluded for simplicity. Adapted from O'Reilly and Munakata (2000, Fig. 7.6, p. 215). Copyright 2000 MIT Press. Reprinted with permission of MIT Press Books.

network changes its weights in response to a single experience. The initial model had a slow learning rate that allowed the network to abstract statistical regularities across experiences. A medium learning rate produces weights that do not capture the statistical regularities as faithfully, but that better capture whether Stimulus 1 was just present (with weights above .8 suggesting the stimulus was just present and weights below .8 suggesting it was not). A fast learning rate produces weights that do not capture the statistical regularities at all at a given point in time (the weights are either the maximum of 1 or the minimum of 0), but that faithfully capture whether Stimulus 1 was just present (with weights of 1 indicating the stimulus was just present and weights of 0 indicating it was not). In this extreme, the weights update quickly based on the most recent experience, overwriting weight changes from previous experiences. This would be like a person remembering only where she parked her car today, but not representing that she tends to park in one lot more often than another. Thus, this simple simulation demonstrates a computational trade-off, between fast learning that can support single-trial learning and slow learning that can support the abstraction of statistical regularities in the environment.

These ideas about computational trade-offs and specializations of hippocampal regions have been explored in more detailed simulations (Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). These simulations have shown how a system that learns rapidly with non-overlapping representations is crucial for the recollection of particular episodes (such as meeting a particular person—e.g., remembering where you met her, what her name is, who you were with, and so on). The rapid learning with non-overlapping representations allows the system to quickly encode the memory and keep it distinct from memories for similar episodes (e.g., the meeting of other people). In contrast, simulations that learn slowly using overlapping representations tend to collapse across the differences of individual episodes; as a result, these systems instead specialize on representing the underlying structure of the environment (e.g., a schema for what typically happens in meeting people). Both types of representations and learning are useful, but there is a computational trade-off between them; a single system cannot simultaneously specialize in both non-overlapping representations with fast learning and overlapping representations with slow learning. As a result, one neural system (the hippocampus) may specialize in the fast and non-overlapping functions, while another neural system (posterior cortical regions) may specialize in the slow and overlapping functions (McClelland et al., 1995; Norman & O'Reilly, 2003).

This computational approach is consistent with (and may help to make sense of) findings from neuroscience regarding the anatomy and physiology of the hippocampus (Squire, Shimamura, & Amaral, 1989). For example, areas of the hippocampus show very sparse levels of activity (Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990), which could contribute to relatively non-overlapping representations. This sparseness can also lead to conjunctive representations, which require a specific combination or conjunction of active units. Such conjunctive representations could be particularly relevant for episodic and spatial memories, which represent particular combinations of information (e.g., the circumstances associated with meeting a particular person, the spatial arrangement and relation of various visual cues) rather

than isolated pieces of information (e.g., just the name of the person you met or the size of one wall in a spatial layout without any other information). And, hippocampal cells appear to have very high rates of long-term potentiation (Monaghan & Cotman, 1989), a candidate mechanism for changes to connection weights in a neural network, which could contribute to relatively fast learning. Thus, a fundamental computational trade-off in memory suggests the need for two specialized systems that the hippocampus and cortex appear to satisfy.

Although the work to-date in this area has not been developmentally focused, it is developmentally relevant. Knowledge of the computational trade-offs inherent in neural specializations, together with the time course of these specializations, could aid in understanding both typical and atypical development. Two examples are considered below for the case of hippocampal and posterior cortical specializations; these await testing through implementation in simulations.

First, the computational account of hippocampal and posterior cortical specializations may provide an explanation of infantile amnesia (McClelland et al., 1995)—the finding that most people cannot remember any experiences they had before the age of about 2 (Howe & Courage, 1993). In this account for the mature system, the hippocampus quickly encodes individual episodes based on cortical representations; episodic memories can later be retrieved when these distinct hippocampal representations activate relevant cortical representations. Thus, the hippocampus is viewed as crucial for both encoding (through rapid learning via non-overlapping, conjunctive representations) and retrieval (through activation of hippocampal representations and associated cortical representations; see e.g., Nadel, Samsonovich, Ryan, & Moscovitch, 2000). In the developing system, the hippocampus would quickly encode individual episodes based on early cortical representations. However, as cortical representations changed over the first few years of life, they would no longer serve to activate the associated representations in the hippocampus. That is, because conjunctive representations require a specific combination of inputs, small changes to cortical inputs could lead to failure to access associated representations in the hippocampus. Thus, retrieval of early memories would be hindered.

Note that many theories attribute infantile amnesia to representations changing, and becoming incompatible, so that prior experiences can no longer be accessed. However, these theories face a difficulty: if the representations change so dramatically, so that prior experiences are effectively no longer represented, one would expect dramatic general deficits to be observed as representations change. All previous knowledge would effectively be overwritten, so that children should not be able to recognize familiar objects, music, words, and so on. But such dramatic general deficits are not observed. The computational trade-off story can make sense of this. Cortical changes are slow, and possibly even slight, such that general deficits are not observed. But these slow and possibly slight cortical changes can lead to large changes in the hippocampal representations that are activated. The large changes in hippocampal representations would lead to dramatic deficits, but only in episodic memory—hence infantile amnesia.

The computational approach may also be relevant for understanding the patterns observed in developmental amnesics (Vargha-Khadem et al., 1997). This population

had lesioned hippocampi from early in life. They show severe episodic memory deficits, as most theories of hippocampal functioning would predict. However, their semantic memory is remarkably intact. This profile is readily predicted by the computational trade-off framework (Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). Even without the ability to form episodic memories with a fast-learning system, the slow-learning posterior cortical system can gradually abstract statistical regularities to support semantic memory.

Thus, the specializations of posterior cortical and hippocampal regions, and the developmental implications, may be understood in terms of computational trade-offs between slow learning with distributed representations and fast learning with sparse representations.

Prefrontal cortical regions

Just as we saw that standard models are not particular by good at fast learning in the preceding section, we will see in this section that they are not particularly good at maintaining information in an active form across time. Such models can be altered in various ways to support better maintenance, but ultimately the issue of computational trade-offs arises again (O'Reilly, Mozer, Munakata, & Miyake, 1999b; O'Reilly & Munakata, 2000).

Consider the simple network in Fig. 5 (O'Reilly & Munakata, 2000). This network contains input and hidden units that represent a monitor, speakers, and keyboard.

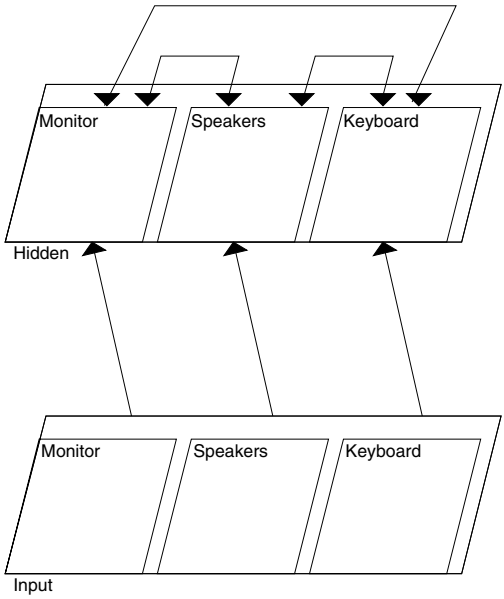

Fig. 5. Semantic network: Weights connect hidden units that represent semantically related information. Adapted from O'Reilly and Munakata (2000, Fig. 9.18, p. 301). Copyright 2000 MIT Press. Reprinted with permission of MIT Press Books.

Weights connect hidden units that represent semantically related information; in this case, each hidden unit is connected to the other two. Such interactive representations confer the semantic benefits described earlier, such as allowing a system to go from incomplete information to activate related information. However, such interactive representations also come with a price: loss of information when it is supposed to be maintained across delays. When this network is presented with a monitor and speakers, the network correctly activates the corresponding hidden units (top half of Fig. 6). However, when the input is removed, the activation spreads across the hidden units via the weights connecting all of the units (bottom half of Fig. 6). As a result, during the maintenance period, it is no longer clear what the network was initially presented with; the network has failed to cleanly maintain this information.

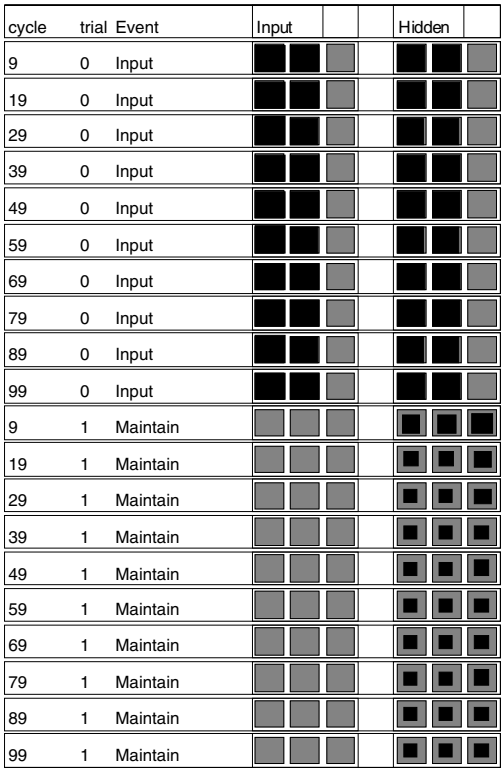

Fig. 6. Input and hidden unit activity as the network in Fig. 5 is presented with two inputs (top half of figure) and then those inputs are removed (bottom half of figure). Each row corresponds to one time step of processing. Each unit's activity level is represented by the size of the corresponding black square. The network activates the corresponding hidden units when the inputs are present, but fails to maintain this information when the input is removed, due to interactive representations. Adapted from O'Reilly and Munakata (2000, Fig. 9.19, p. 301). Copyright 2000 MIT Press. Reprinted with permission of MIT Press Books.

The simple network in Fig. 5 can be elaborated to improve its active maintenance abilities. For example, higher-order representations can be added, which connect with lower-level features that go together (Fig. 7). This improves the network's ability to maintain information after it is removed. Instead of the activation simply spreading from Monitor and Speakers to Keyboard, for example, the higher-order representation of TV is preferentially activated in the second hidden layer, and this preferentially activates Monitor and Speakers. However, because the system is still relatively interconnected (e.g., Monitor also connects to Terminal and Speakers also connects to Synth), this solution is not particularly robust. When a small amount of noise is introduced into the network processing (of the sort that our brains likely contend with on a regular basis), activation again spreads beyond the initial input, due to the connections with other units.

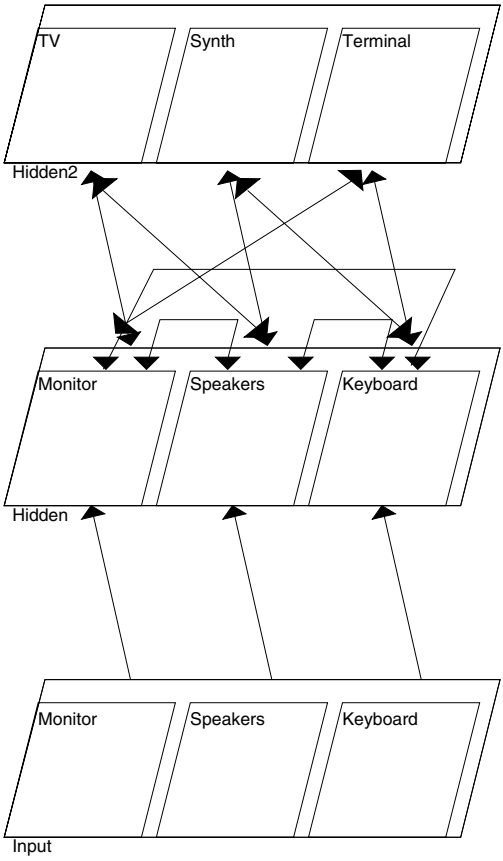

Fig. 7. Semantic network with higher order representations: Weights connect hidden units that represent semantically related information, with a layer of higher-order representations that connect with lower-level features. Adapted from O'Reilly and Munakata (2000, Fig. 9.20, p. 302). Copyright 2000 MIT Press. Permission pending.

More isolated representations may be required for systems to maintain representations over delays, in the absence of input, and in the face of noise (e.g., for working memory). An extreme form of such isolated representations is shown in Fig. 8. In this network, each input unit is connected to its corresponding hidden unit, and each hidden unit is connected only to itself, rather than to the other semantically-related hidden units. When this network is presented with a monitor and speakers, the network correctly activates the corresponding hidden units (top half of Fig. 9). And, when the input is removed, the activation is maintained in these units (bottom half of Fig. 9), because there is no way for the activation to spread from these units to any other units. As a result, this network successfully maintains the previously presented information during the maintenance period. This solution is robust to noise in the network processing.

In this way, these simple simulations demonstrate a computational trade-off, between interactive representations that can support semantic knowledge and isolated representations that can subserve active maintenance of information across delays, of the sort required for working memory. Again, both types of representations are useful, but there is a computational trade-off between them; a single system cannot simultaneously specialize on interconnected and isolated representations. As a result, one neural system (posterior cortex) may specialize on interconnected representations,

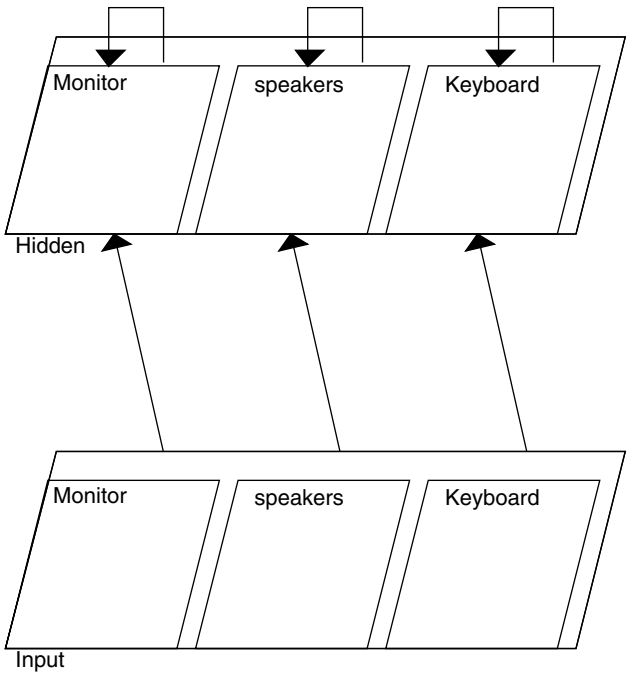

Fig. 8. Network with isolated representations: Each hidden unit connects to only itself, rather than to other semantically related units.

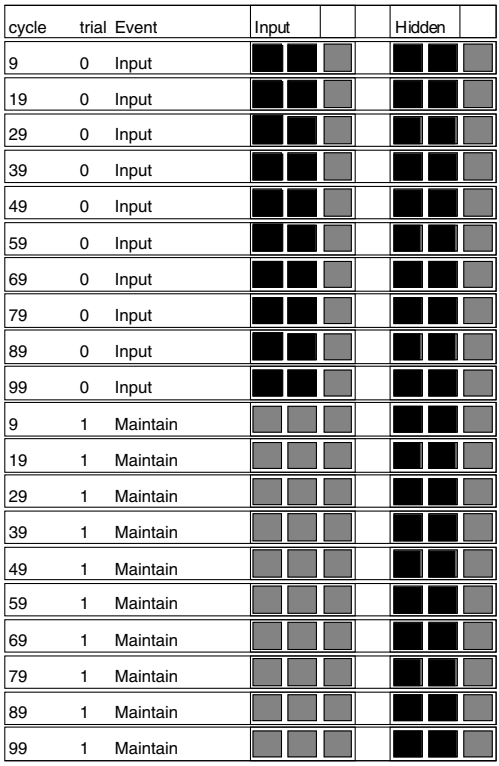

Fig. 9. Input and hidden unit activity as the network in Fig. 8 is presented with two inputs (top half of figure) and then those inputs are removed (bottom half of figure). The network activates the corresponding hidden units when the inputs are present, and maintains this information when the input is removed, due to isolated representations.

while another system (prefrontal cortex) may specialized on isolated representations. Again, this computational approach is consistent with (and may help to make sense of) findings from neuroscience regarding the anatomy (Levitt, Lewis, Yoshioka, & Lund, 1993) and physiology (Rao, Williams, & Goldman-Rakic, 1999) of prefrontal cortex, which may suggest more isolated representations in this region.

These ideas about computational trade-offs and specializations of the prefrontal cortex have been applied to understanding developmental changes in task-switching. Children show remarkable limitations and developments in their ability to switch from one task to another. For example, infants from 8 months or so can successfully retrieve a toy after seeing it hidden. However, infants tend to reach back to this hiding location even after watching as the toy is hidden in a new location (Diamond, 1985; Piaget, 1954). That is, they *perseverate*, repeating their previous behavior when it is no longer appropriate, instead of responding flexibly to the change in task. Similarly, 3-year-old children can sort cards according to their color or their shape, as first instructed by an experimenter. However, they tend to perseverate with this first rule, even after being instructed to switch to sort by the other rule (i.e., shape if they

started with color, or color if they started with shape) (Zelazo, Frye, & Rapus, 1996). Six-year-old children show a similar pattern in a speech interpretation task (Morton & Trehub, 2001). In this task, children hear utterances with conflicting emotional cues (e.g., "My dog ran away from home" spoken in a happy tone of voice). They can judge the speaker's feelings from the content of the utterances, but when instructed to switch to judge the speaker's feelings from the tone of voice, they perseverate and continue to base their judgments on content (Morton & Munakata, 2002b; Morton, Trehub, & Zelazo, in press).

Neural network simulations have been used to investigate the basis for such limitations (and subsequent developments) in flexibility (Munakata, Morton, & Stedron, 2003). Developments in prefrontal cortex are thought to play a critical role in ultimately supporting flexible behavior in such tasks (Diamond, 2002, Chap. 22; Miller & Cohen, 2001; Miyake & Shah, 1999; O'Reilly, Braver, & Cohen, 1999a; Roberts & Pennington, 1996; Stuss & Benson, 1984). In neural networks, changes corresponding to developments in prefrontal cortex can be instantiated via connection weights supporting isolated representations (as in Fig. 8). Increases to such connection weights increase one aspect of working memory, improving networks' abilities to maintain currently relevant information. As networks get better at maintaining such information, they are better able to respond based on, for example, where a toy was recently hidden or what rule is currently relevant for sorting cards or interpreting spoken utterances, rather than falling back on previously relevant information and behaviors (instantiated via increased connections in posterior cortical areas).

Further, increases in working memory in such systems are gradual rather than all-or-none. A weak ability to maintain currently relevant information may suffice for some tasks but not others. As a result, these networks also simulate the phenomenon of behavioral dissociations often observed in tasks of flexibility. In such dissociations, children pass certain measures of flexibility while failing others. For example, even as infants reach perseveratively to a previous hiding location for a toy, they occasionally gaze at the correct hiding location (Diamond, 1985; Hofstadter & Reznick, 1996; Piaget, 1954). And, even as children sort perseveratively according to a previous rule, they can correctly answer questions about the new rule, such as where trucks should go in the shape game (Zelazo et al., 1996), or what aspect of a speaker's voice they should be listening to (Morton & Munakata, 2002b; Morton et al., in press). Such dissociations arise naturally in neural network models based on gradual increases in working memory. For example, a weak representation of a toy's current hiding location allows a network to respond correctly through a system that updates frequently (such as gazing), but not through a system that updates less frequently (such as reaching) (Munakata, 1998a). Similarly, a weak representation of a currently relevant rule allows a network to respond correctly to tasks that do not involve conflict (such as questions about where trucks should go in the shape game), but not to tasks that require conflict to be resolved (such as sorting a card that has currently relevant information about shape and previously relevant information about color) (Morton & Munakata, 2002a). This neural network approach thus predicted that children's behavioral dissociations should disappear when such tasks are

equated for conflict (e.g., by asking where *red* trucks go in the shape game, and requiring a red truck to be sorted). This prediction has been confirmed (Morton & Munakata, 2002b; Munakata & Yerys, 2001).

In many of the simulations investigating computational trade-offs between posterior and prefrontal cortical regions, the connections to the prefrontal regions of the models were set to learn at a slower rate than other connections in the models. This manipulation helped the prefrontal systems to be less prone to repeating prior behaviors, allowing them to guide other regions to overcome biases toward previously relevant information and behaviors. The need for this manipulation suggests that learning rate might be another factor distinguishing the computational specializations of prefrontal and posterior cortical regions.

This idea might seem counterintuitive, given that the prefrontal cortex is viewed as critical for "higher level cognition"—executive function, controlled processing, flexible behavior, and so on. How could a brain region that was part of such a system be slow learning? One possibility is that the quickness of prefrontal systems comes not through learning rate (i.e., fast changes to connection weights), but through the updating of information held in mind. That is, the flexibility of this system comes through the ability to rapidly update and manipulate activity states when necessary, rather than through the ability to rapidly learn new representations. The relationship between such rapid updating and the maintenance of representations has been investigated in more detailed simulations of prefrontal systems (Frank, Loughry, & O'Reilly, 2001; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2003; Rougier & O'Reilly, 2002). These simulations demonstrated how interactions with another specialized system, the basal ganglia, may serve as a gating mechanism for whether representations in prefrontal cortex are maintained or updated. Further, the simulations showed how such interactions may play an important role in adults' abilities to dynamically switch from one task to another. This work represents an important step in addressing aspects of prefrontal function other than maintenance. Additional work is needed to assess the relevance of this framework for understanding developmental changes in such processes.

In sum, the specializations of prefrontal and posterior cortical regions, and the developmental implications, may be understood in terms of computational trade-offs between isolated and interactive representations, and possibly between different learning and updating rates.

Discussion

In summary (Table 1), there are computational trade-offs inherent across different types of representations and learning. A single system cannot specialize on both fast and slow learning, both interactive and isolated representations, and so on. This computational framework can help make clear why different neural regions are required for distinct types of memory functions, and can help make sense of physiological and anatomical findings about different brain areas. This framework also has broad developmental relevance, and has already shown promise in a variety of

Table 1
Summary of computational framework for considering neural specializations for memory functions and their developmental relevance

Region	Representations	Learning	Memory specialization	Developmental relevance
Posterior cortex	Distributed, interactive	Slow	Semantic	Category learning
				Statistical learning
Hippocampal regions	Sparse, conjunctive	Fast	Episodic	Childhood amnesia
			Spatial	Developmental amnesics
Prefrontal cortex	Isolated	Slowest?	Working	Perseveration, flexibility

domains, including category learning, developmental amnesia, and the development of flexibility.

Understanding neural specializations in terms of computational trade-offs could also help to make sense of developmental changes in patterns of brain activity and effects of brain damage. For example, an area that subserves a computational function crucial for learning a skill might be unimportant for executing the skill once it is learned. Understanding these kinds of computational dependencies could help to explain why certain kinds of brain damage (or activity) are linked to the learning of skills, such as reading, but not to the expert execution of those skills (Stiles, Bates, Thal, Trauner, & Reilly, 2002).

There is another way in which computational trade-offs may prove to be relevant for understanding development, and development may prove to be relevant for understanding computational trade-offs. In particular, the way in which neural systems specialize and accommodate computational trade-offs may change during the course of development. Such developmental changes might suggest that the neural specializations observed in the mature system may best be understood by considering the developmental processes that led to them. And, fully understanding the relevance of computational trade-offs for development may require us to go beyond taking the adult model and applying it to the developing system; a full understanding may require a consideration of interactions between computational trade-offs and developmental processes (see Karmiloff-Smith, 1998 for similar arguments for understanding developmental disorders).

This presentation has focused on the computational contrasts among brain regions, as a way of understanding their functional specializations. However, it is important to note that these contrasts are likely to be more graded in nature rather than all-or-none. For example, conjunctive representations exist outside of hippocampal regions, as in posterior visual cortical areas that respond to conjunctions of visual cues (Tanaka, 1996), parietal cortical areas that respond to conjunctions of perceptual information and eye movement plans (Colby, Duhamel, & Goldberg, 1996), and prefrontal cortical areas that respond to conjunctions of cues and responses (Asaad, Rainer, & Miller, 1998). Similarly, topographic organization of related information exists outside of

prefrontal regions, as in the columns of cells responding to similar information observed in posterior cortical regions (Tanaka, 1996). Thus, the specializations discussed in this paper (and summarized in Table 1) are likely to reflect *relative* specializations. Graded differences across neural systems lead to computational advantages and disadvantages, which subserve relative functional specializations across areas.

Note that this computational approach focuses on the mechanisms underlying different types of memory specializations—the nature of representations and learning required to subserve different types of memory functions. As such, this approach may shed light on existing memory distinctions cast in terms of declarative versus procedural, episodic versus semantic, explicit versus implicit, and so on (Farah & McClelland, 1991; Munakata, 1998b; O'Reilly & Rudy, 2001). In some cases, the computational characterizations may provide a mechanistic account that maps directly onto existing distinctions made at the verbal level. The computational accounts would thus serve as another level of understanding existing verbal distinctions. In other cases, the computational characterizations may offer a re-casting of existing distinctions made at the verbal level.

In conclusion, interest and research activity has been increasing in both cognitive neuroscience approaches to development (see Johnson, 1997; Johnson, Munakata, & Gilmore, 2002; Nelson & Luciana, 2001, for reviews) and computational approaches to development (see Elman et al., 1996; Mareschal & Shultz, 1996; Munakata & Stedron, 2001; Simon & Halford, 1995; Thelen & Smith, 1994, for reviews). This article has focused on a growing intersection of these approaches—simulations that may serve as a bridge in the study of brain–behavior relations. These simulations investigate how computational trade-offs can lead to specializations of neural regions for different functions, and how an understanding of these computational trade-offs can inform the study of development. Much of this work is in relatively early stages, with great potential for additional simulation work to implement and test the mechanisms proposed. Such work will likely lead to elaborations of the specific ideas reviewed here regarding semantic, episodic, and working memory, and the implications for early memory development. Whatever the ultimate conclusions, computational models should continue to serve as a useful tool for investigating neural specializations and their developmental relevance.

References

Arbib, M. A. (Ed.). (2002). *The handbook of brain theory and neutral networks* (2nd ed.). Cambridge, MA: MIT Press.

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron, 21*, 1399–1407.

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.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research, 83*, 287–300.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of frontal cortex involvement in human working memory. *NeuroImage, 5*, 49–62.

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1996). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of Neurophysiology, 76*, 2841.

Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants' performance on$A\bar{B}$. *Child Development, 56*, 868–883.

Diamond, A. (2002). A model system for studying the role of dopamine in prefrontal cortex during early development in humans. In M. H. Johnson, Y. Munakata, & R. O. Gilmore (Eds.), *Brain development and cognition: A reader* (pp. 441–493). Oxford: Blackwell.

Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience, 3*(11), 1149–1152.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Farah, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press.

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences, 17*, 43–104.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120*, 339–357.

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

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT/Bradford Press.

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.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100*, 1111–1121.

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.

Hermer, L., & Spelke, E. (1996). Modularity and development: the case of spatial reorientation. *Cognition, 61*, 195–232.

Hofstadter, M. C., & Reznick, J. S. (1996). Response modality affects human infant delayed-response performance. *Child Development, 67*, 646–658.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin, 113*, 305–326.

Johnson, M. H. (1997). *Developmental cognitive neuroscience: An introduction*. Blackwell Publishers.

Johnson, M. H., Munakata, Y., & Gilmore, R. O. (Eds.). (2002). *Brain development and cognition: A reader* (2nd ed.). Oxford: Blackwell.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences, 2*, 389–398.

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.

Lashley, K. S. (1929). *Brain mechanisms and intelligence*. Chicago: University of Chicago Press.

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

Mareschal, D., & French, R. (2000). Mechanisms of categorization in infancy. *Infancy, 1*, 59–76.

Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development, 11*, 571–603.

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.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group (Eds.). (1986). *Parallel distributed processing. Volume 2: Psychological and biological models*. Cambridge, MA: MIT Press.

McClelland, J. L., Thomas, A., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. In J. Reggia, E. Ruppin, & D. Glanzman (Eds.), *Brain, behavioral, and cognitive disorders: The neurocomputational perspective* (pp. 75–80). Oxford: Elsevier.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–164). San Diego, CA: Academic Press.

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

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefontal cortex of the macaque. *Journal of Neuroscience, 16*, 5154–5167.

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

Monaghan, D. T., & Cotman, C. W. (1989). Regional variations in NMDA receptor properties. In J. C. Watkins & G. L. Collingridge (Eds.), *The NMDA receptor* (pp. 53–64). Oxford, UK: Oxford University Press.

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

Morton, J. B., & Munakata, Y. (2002b). Are you listening? Exploring a knowledge action dissociation in a speech interpretation task. *Developmental Science, 5*, 435–440.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72*(3), 834–843.

Morton, J. B., Trehub, S. E., & Zelazo, P. D. (in press). Representational inflexibility in 6-year-olds' understanding of emotion in speech. Child Development.

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

Munakata, Y. (1998b). Infant perseveration: Rethinking data, theory, and the role of modelling. *Developmental Science, 1*, 205–212.

Munakata, Y., Morton, J. B., & Stedron, J. M. (2003). The role of prefrontal cortex in perseveration: Developmental and computational explorations. In P. Quinlan (Ed.), *Connectionist models of development*. East Sussex: Psychology Press.

Munakata, Y., & Stedron, J. M. (2001). Neural network models of cognitive development. In C. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 159–171). Cambridge, MA: MIT Press.

Munakata, Y., & Yerys, B. E. (2001). All together now: When dissociations between knowledge and action disappear? *Psychological Science, 12*(4), 335–337.

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus, 10*, 352–368.

Nelson, C. A.& Luciana, M. (Eds.). (2001). *Handbook of developmental cognitive neuroscience*. Cambridge: 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.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999a). 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., Mozer, M., Munakata, Y., & Miyake, A. (1999b). Discrete representations in working memory: A hypothesis and computational investigations. *The second international conference on cognitive science* (pp. 183–188). Tokyo: Japanese Cognitive Science Society.

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

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.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. *Journal of Experimental child Psychology, 66*, 236–263.

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

Roberts, R. J., & Pennington, B. F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology, 12*(1), 105–126.

Rogers, T., & McClelland, J. (in press). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (in press). A parallel distributed processing approach to semantic cognition: Applications to conceptual development. In Rakison, D., Gershkoff-Stowe, L., (Eds.), *Building object categories in developmental time: Proceedings of the carnegie symposium on cognition* (Vol. 32). Hillsdale, NJ: Erlbaum.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2003). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. Manuscript submitted for publication.

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.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (Eds.). (1986). *Parallel distributed processing. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

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

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20*, 11–21.

Simon, T. J., & Halford, G. S. (1995). *Developing cognitive competence: New approaches to process modeling*. Erlbaum.

Squire, L. R., Shimamura, A. P., & Amaral, D. G. (1989). Memory and the hippocampus. In J. H. Byrne & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches*. San Diego, CA: Academic Press.

Stiles, J., Bates, E. A., Thal, D., Trauner, D., & Reilly, J. (2002). Linguistic and spatial cognitive development in children with pre- and perinatal focal brain injury: A ten-year overview from the San Diego longitudinal project. In M. H. Johnson, R. O. Gilmore, & Y. Munakata (Eds.), *Brain development and cognition: A reader* (2nd ed., pp. 272–291). Oxford: Blackwell.

Stuss, D., & Benson, D. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin, 95*, 3–28.

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

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thompson-Schill, S. L., Aguirre, G. K., & Farah, M. J. (1999). A neural basis for category and modality specificity of semantic knowledge. *Neuropsychologia, 37*, 671–676.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277*, 376.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*, 37–63.