Journal of Experimental Psychology: General

Incorporating Rapid Neocortical Learning of New Schema-Consistent Information Into Complementary Learning Systems Theory

James L. McClelland

Online First Publication, August 26, 2013. doi: 10.1037/a0033812

CITATION

McClelland, J. L. (2013, August 26). Incorporating Rapid Neocortical Learning of New Schema-Consistent Information Into Complementary Learning Systems Theory. *Journal of Experimental Psychology: General*. Advance online publication. doi: 10.1037/a0033812

Journal of Experimental Psychology: General

© 2013 American Psychological Association
0096-3445/13/$12.00 DOI: 10.1037/a0033812

THEORETICAL REVIEW

Incorporating Rapid Neocortical Learning of New Schema-Consistent Information Into Complementary Learning Systems Theory

James L. McClelland
Stanford University

The complementary learning systems theory of the roles of hippocampus and neocortex (McClelland, McNaughton, & O'Reilly, 1995) holds that the rapid integration of arbitrary new information into neocortical structures is avoided to prevent catastrophic interference with structured knowledge representations stored in synaptic connections among neocortical neurons. Recent studies (Tse et al., 2007, 2011) showed that neocortical circuits can rapidly acquire new associations that are consistent with prior knowledge. The findings challenge the complementary learning systems theory as previously presented. However, new simulations extending those reported in McClelland et al. (1995) show that new information that is *consistent* with knowledge previously acquired by a putatively cortexlike artificial neural network can be learned rapidly and without interfering with existing knowledge; it is when *inconsistent* new knowledge is acquired quickly that catastrophic interference ensues. Several important features of the findings of Tse et al. (2007, 2011) are captured in these simulations, indicating that the neural network model used in McClelland et al. has characteristics in common with neocortical learning mechanisms. An additional simulation generalizes beyond the network model previously used, showing how the rate of change of cortical connections can depend on prior knowledge in an arguably more biologically plausible network architecture. In sum, the findings of Tse et al. are fully consistent with the idea that hippocampus and neocortex are complementary learning systems. Taken together, these findings and the simulations reported here advance our knowledge by bringing out the role of consistency of new experience with existing knowledge and demonstrating that the rate of change of connections in real and artificial neural networks can be strongly prior-knowledge dependent.

Keywords: learning, memory, schemas, consolidation, hippocampus

It has long been known that it is relatively easy to learn new things that are consistent with prior knowledge (Bartlett, 1932; Bransford, 1979). In two recent articles, Tse et al. (2007, 2011) explored the neural basis of this fact. These studies showed that new associations are incorporated into neocortical structures quickly if they are consistent with prior knowledge. The findings challenge the complementary learning systems theory (CLST) of the roles of hippocampus and neocortex in learning and memory (McClelland, McNaughton, & O'Reilly, 1995). However, new simulations reported in this article using the model considered by McClelland et al. (1995) capture key aspects of the findings of Tse et al. and indicate that new information consistent with existing knowledge structures can be assimilated quickly and without interference into putatively neocortexlike neural networks. In our 1995 study, my colleagues and I did not emphasize the role of consistency with prior knowledge. The present article corrects this deficiency, showing how the simple model of neocortical learning we used in the 1995 study can address some of the key findings from the Tse et al. studies. The present article also considers limitations of the model used in McClelland et al. and points toward the possibility that more biologically plausible models of cortical learning may also capture the key effects.

According to CLST, structured knowledge representations depend on a learning system in the neocortex. In McClelland et al. (1995), we used an artificial neural network first introduced by Rumelhart (1990; Rumelhart & Todd, 1993) to illustrate how acquisition of such knowledge may proceed. Although the model abstracts from biological details and is far too simple to do justice to all of the cognitive and neurobiological characteristics of neocortical learning, subsequent work based on this simple network established its usefulness for capturing many findings in the literature on the emergence of knowledge over the first years of human life and on the disintegration of performance on semantic tasks in diseases affecting semantic knowledge. Specifically, simulations using this simple network (hereafter called the *Rumelhart network*)

James L. McClelland, Department of Psychology and Center for Mind, Brain, and Computation, Stanford University.

This material is based in part on research sponsored by the Air Force Research Laboratory, under Agreement FA9550-07-1-0537. I would like to thank Dharshan Kumaran, Ken Norman, and members of the Stanford PDP research group for helpful suggestions and comments on earlier versions of the manuscript.

Correspondence concerning this article should be addressed to James L. McClelland, Stanford University, Department of Psychology, 450 Serra Mall, Building 420, Stanford, CA 94305. E-mail: mcclelland@stanford.edu

reported in Rogers and McClelland (2004) capture many patterns of learning and generalization seen in development, including the gradual differentiation of conceptual knowledge in development (Keil, 1979; Mandler & Bauer, 1988), the reorganization of conceptual knowledge seen in some domains (Carey, 1985), U-shaped developmental patterns of overgeneralization (Mervis & Crisafi, 1982), and differential extension of properties of objects to other objects as a function of object category (Macario, 1991) or property type (Gelman & Markman, 1986). Other simulations using this network (Rogers & McClelland, 2004) and related architectures (Rogers et al., 2004) also captured the effects of damage to anterior temporal neocortex on object naming and on attribution of properties to objects, including loss of item-specific knowledge and preservation of category-general information, together with overgeneralization of frequent names and frequent object properties (Hodges, Graham, & Patterson, 1995; Patterson et al., 2006). In spite of its simplicity, then, the Rumelhart network and related architectures appear useful for characterizing many aspects of knowledge acquisition in neocortical networks and of disintegration of such knowledge when these networks are affected by disease. These models do have limitations that I will begin to address, but their usefulness as detailed above makes it seem worthwhile to understand better how they respond to new information that is or is not consistent with existing knowledge representations.

In networks like the Rumelhart network, rapid learning of new information inconsistent with prior knowledge can produce catastrophic interference (McClelland et al., 1995; McCloskey & Cohen, 1989). In McClelland et al. (1995), we showed that this problem could be avoided by *interleaved learning,* in which new information is repeatedly presented, interleaved with known information. Interleaving promotes gradual assimilation of the new information into connections among the network's neuronlike units with a minimum of interference. Drawing on ideas previously proposed by Marr (1971), we proposed that neocortex uses a slow learning system with features like those of the Rumelhart network, complemented by a fast learning system in hippocampus. According to the theory, the hippocampus quickly acquires new experiences. These can be reactivated to guide behavior and to support interleaved training of neocortex, allowing new knowledge to be integrated gradually into neocortical knowledge networks.

The recent studies of Tse et al. (2007, 2011) provided strong evidence that new information can be incorporated into neocortical structures rapidly if this information can be integrated into a previously acquired representation of a complex spatial environment. I begin my analysis by reviewing the key behavioral findings from Tse et al. (2007) that challenge the CLST.

In the Tse et al. (2007) study, rats were trained in a previously unfamiliar, complex spatial environment (see Figure 1) to associate flavor cues (presented in a start box) with locations in the environment. The environment consisted of a 1.6-m × 1.6-m arena and contained several distinctive landmarks. Transparent walls allowed animals to see out into the surrounding laboratory environment. The floor of the arena contained circular holes at six locations, at which food rewards could be hidden.

Rats received extensive training on flavor–location associations, interspersed with sessions using probe trials (see Figure 1A). In each training session, all trials started with the rat placed in one of four start boxes (one at the center of each side of the arena—starting location varied across sessions) where it was given a cue flavor, with the task of then foraging to find a larger sample buried at one of the six locations in the environment. Each flavor was associated with a unique location in the environment, and rats received one training trial with each flavor in each daily session. Rats gradually improved in their ability to find the correct location before visiting other locations (see Figure 1A and caption for details; a preliminary study established the importance of an intact hippocampus for learning in this task setting). At different points during training, rats were tested with unrewarded probe trials (see Figure 1B), revealing gradual learning over 16 training sessions and controlling for possible confounds.

Rats were then exposed to a single critical training session, in which four of the flavor–place associations were trained as usual but the remaining two old flavor–location pairs were replaced with two new ones, using new locations near the old replaced locations (see Figure 1C). Strikingly, at a probe test on the following day, rats showed strong evidence of learning, searching at the new cued location far more than other locations (also in Figure 1C). Another day later (48 hr after exposure to each of the new flavors), rats then received extensive hippocampal lesions or sham control lesions. After recovery, further probe tests (see Figure 1D) revealed that lesioned animals (as well as control animals) retained both the original and the new flavor–place associations, supporting the conclusion that new schema-consistent information can be assimilated rapidly into neocortical structures and that this learning does not interfere with previously acquired schema-consistent information. Only the controls, however, were able to learn another set of flavor–place associations. Additional control studies established that no performance improvement occurred even in intact animals after one exposure to two new flavor–place associations in a novel spatial environment. Improvement by control animals with new flavor–place associations in the new environment proceeded slowly, at the same slow pace as acquisition of the first set of associations in the original environment.

Taken together with additional controls, the findings of Tse et al. (2007) established that the hippocampus and related areas affected by their lesions are essential during the initial gradual acquisition of flavor–place associations in a previously unfamiliar environment and for the acquisition of new flavor–place associations once the environment has become familiar. Importantly, however, these structures need only be available to the rat for at most 48 hr after exposure to the new flavor–place associations in the familiar environment for the rat to exhibit retention subsequent to hippocampal lesion. Additional findings reported in Tse et al. (2011) provided striking evidence that learning new schema-consistent flavor–place associations results in the expression of genes associated with plasticity at several neocortical sites, to a greater extent than occurs with new flavor–location associations in a novel environment. These and other aspects of the results reported in the Tse et al. studies support the view that a structured body of knowledge or schema can be built up gradually over time in networks of neurons outside the hippocampus and that, once the schema has been acquired, new knowledge can be rapidly added to it.

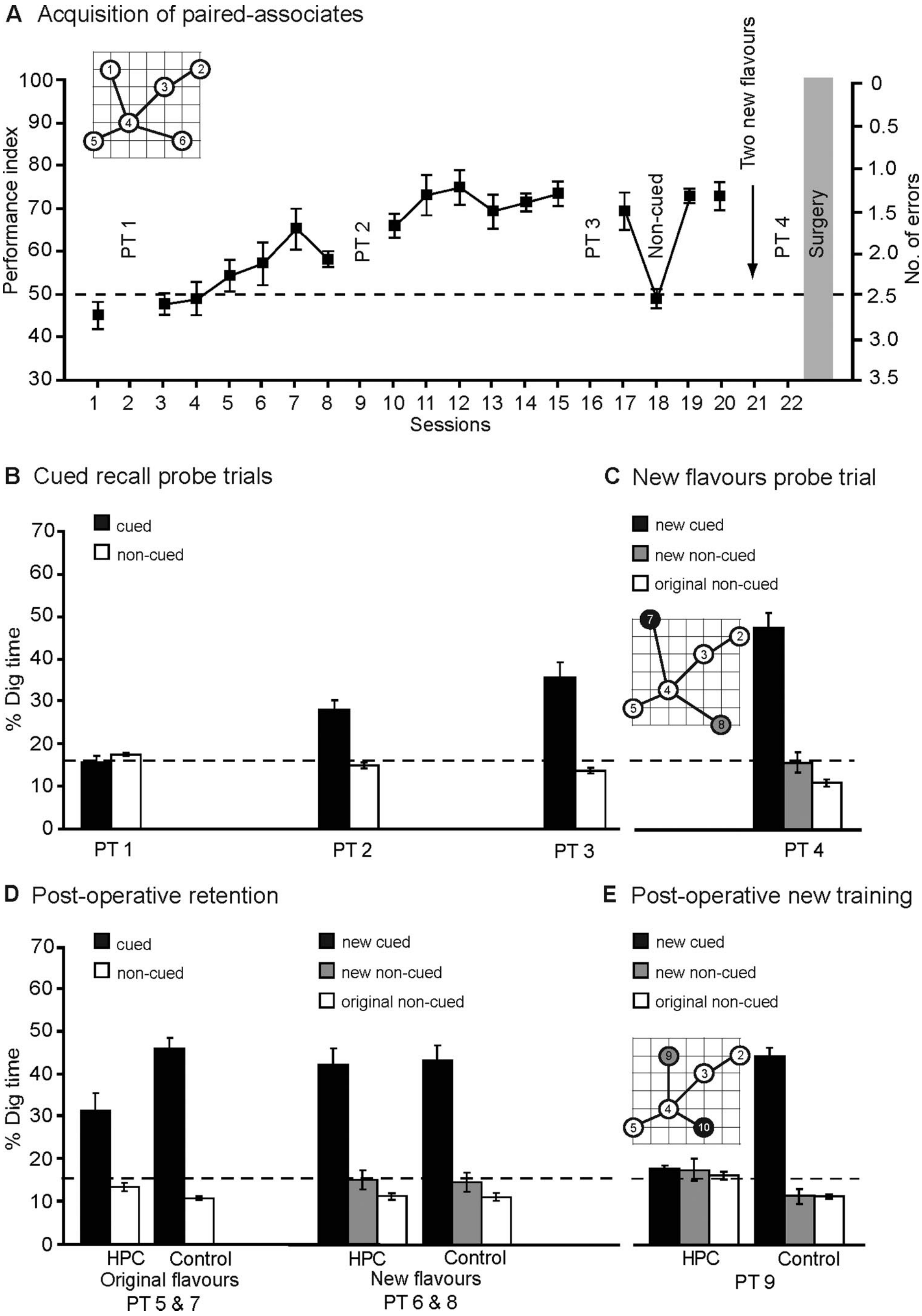

Figure 1. A: Training and testing protocol and performance during acquisition of six flavor–location associations in a previously unfamiliar environment. Inset shows locations associated with Flavors 1–6. The *x*-axis indicates daily training or testing sessions numbered from the start of the experiment. Performance index is based on the number of incorrect locations visited before visiting the correct location (see Tse et al., 2007, for details). B: Performance on original paired associates during unrewarded probe trials (PTs) on days indicated in Panel A. Chance level is indicated by the horizontal dashed line. C: Performance on new cued, new noncued, and original noncued PTs after exposure to two new flavor–location pairs (7 and 8; see inset) on Day 20. D: Performance of hippocampal lesioned (HPC) and sham-lesioned (Control) animals on new and original associations after recovery from surgery. E: Performance on two additional paired associates to which animals were first exposed after surgery (9 and 10; see inset). Error bars in all panels indicate the standard error of the mean. From “Schemas and Memory Consolidation,” by D. Tse et al., 2007, *Science, 316,* p. 78. Copyright 2007 by the American Association for the Advancement of Science. Reprinted with permission.

Implications for Complementary Learning Systems Theory

Many findings from Tse et al. (2007) are consistent with the CLST, in which the acquisition of a structured body of knowledge depends on a gradual learning process involving an interplay between hippocampus and neocortex, while rapid acquisition of new information depends on an intact hippocampal system. Notwithstanding these consistent aspects of the findings, Tse et al. noted that the rapid assimilation of new information into the neocortex that was demonstrated in their studies poses a challenge to the CLST. It was, however, a feature of the initial simulation reported in McClelland et al. (1995) that the new information used to illustrate catastrophic interference with existing knowledge in the putatively neocortexlike Rumelhart network was *inconsistent* with aspects of that existing knowledge. Though this inconsistency was mentioned in McClelland et al., it was not emphasized, and no explorations of the role of consistency were conducted.

The main simulations reported here address the challenge posed by the Tse et al. (2007, 2011) studies, focusing on the importance of consistency of new learning with prior knowledge in the Rumelhart network. The first simulation shows that new information that is consistent with prior knowledge is learned quickly in the model, just as it is in rodents, in contrast with inconsistent information, which is learned more slowly. The learning of new consistent information produces little interference with other information that is already known, so that it is not problematic for cortex to learn it quickly. The second simulation addresses the gene expression findings reported in Tse et al. (2011) and shows that consistent, but not inconsistent, new learning results in large changes in the critical connections supporting the addition of new consistent information into the network's learned schema.

It is important to emphasize that the simulations are unlikely to do full justice to the cognitive, behavioral, and biological properties of natural learning systems in humans and animals, including the details of the spatial learning situation employed in the Tse et al. (2007, 2011) studies or the characteristics of the circuitry underlying learning and memory in the mammalian brain. Rather than attempt to model spatial learning in a task like that used by Tse et al. or to attempt a fully biologically realistic model of the learning system available in mammalian brains, I have stuck with Rumelhart's original semantic knowledge task, training set, and network architecture. I have done so because, as I shortly argue, the knowledge acquired in this task is schemalike and because I have found that this model's characteristics—without any modification, only extending it to the explicit examination of its acquisition of prior-knowledge-consistent information—provide a sufficient basis for capturing some of the key findings from the studies of Tse et al. and for elucidating how new learning may interact with structured knowledge in a system with several putative brainlike properties. The simulations may help elucidate certain features that the real learning system used in the brain may have—features that have not been laid fully bare in previous work, including work of my own. As already noted, however, the Rumelhart network and related models do have some limitations, which I point out after presenting the main simulations. An additional, very preliminary simulation considers an alternative, arguably more biologically plausible network architecture that captures aspects of prior-knowledge dependence as well. Overall, the simulations are small steps toward understanding the brain mechanisms underlying schema-consistent learning.

Simulation 1

For Simulation 1, I first trained the Rumelhart network with Rumelhart's original database of knowledge about eight living things using interleaved learning, as in previous work with this network (McClelland et al., 1995; Rogers & McClelland, 2004; Rumelhart & Todd, 1993). The knowledge acquired by the network is schemalike in the sense that the items in the database have a complex pattern of similarity relationships, such that they can be mapped to different points in a multidimensional similarity space, where they are arranged into two superordinate categories (*plants* and *animals*) and, within each superordinate category, into two subcategories (*trees* and *flowers, birds* and *fish).* This category structure captures patterns of intercorrelations among features of items. This category structure is similar in some ways to the spatial structure learned in the Tse et al. (2007) task, in that the locations in the environment have a complex pattern of similarity relationships owing to their locations in space and relations to intra- and extra-area cues in the environment. While the detailed structure of the similarities among concepts in a taxonomic space and locations in a physical space may be different, both kinds of structure can be captured within a Rumelhart-networklike architecture (Rogers & McClelland, 2008). Note, furthermore, that there is an element of arbitrariness in both the Tse et al. task and the task facing the Rumelhart network. In the Tse et al. task, flavors are mapped completely arbitrarily onto spatial locations in the environment; in the Rumelhart model, input units (one for each concept) are mapped arbitrarily onto concepts.[1] Paralleling findings from Tse et al., the Rumelhart network acquires knowledge of the properties of the eight items in the training set gradually.

The simulation then explored the fundamental challenge to the CLST posed by the Tse et al. (2007) findings—that new schema-consistent information can be rapidly integrated into an existing schema—by considering the learning of new schema-consistent information in the Rumelhart model. The simulation showed that in fact, just such rapid assimilation can occur for new schema-consistent items. Specifically, I compared learning of a new schema-consistent item (either a *trout,* a fish that swims like other fish, or a *cardinal,* a bird that flies like other birds) with learning of the (partially) schema-inconsistent *penguin* (a bird that swims but does not fly) previously used in McClelland et al. (1995). The same values of all parameters were used in each case. Focused learning about either consistent item produced very rapid learning and produced little interference with existing knowledge in the network. In contrast, as previously shown in McClelland et al., focused learning about the penguin required many more presentations and led to considerable interference (degradation of performance on related items); for this schema-inconsistent item, neocortical learning did not occur rapidly, and interleaving was necessary to avoid interference.

[1] Input units for similar concepts are located near each other (e.g., *robin* is next to *canary.*) However, the network has no access to this proximity information, and it is irrelevant to the learning that occurs in the network.

Method

Neural network and training materials. The network architecture used in all of the simulations is the one introduced in Rumelhart (1990; Rumelhart & Todd, 1993; see Figure 2). This architecture was implemented in the ***bp*** program in the MATLAB-based PDPTool network simulation system (McClelland, 2011). The base network included eight item input units (corresponding to *pine, oak, rose, daisy, robin, canary, salmon,* and *sunfish*) and four relation input units (corresponding to *isa, is, can,* and *has*), eight representation units (pool to the right of the item input units), 15 hidden units (next pool to the right), and 36 output units. Included among the output units were eight units corresponding to item-specific names (e.g., *pine, oak,* etc.). The network contained a complete set of connection weights linking each item unit to each representation unit, each representation and relation unit to each hidden unit, and each hidden unit to each output unit. The network also contained a modifiable bias weight for each representation and hidden unit (not shown in the figure). Following the usage in Rogers and McClelland (2004), the bias weights on the output units were fixed at −2, so that before learning occurred, output

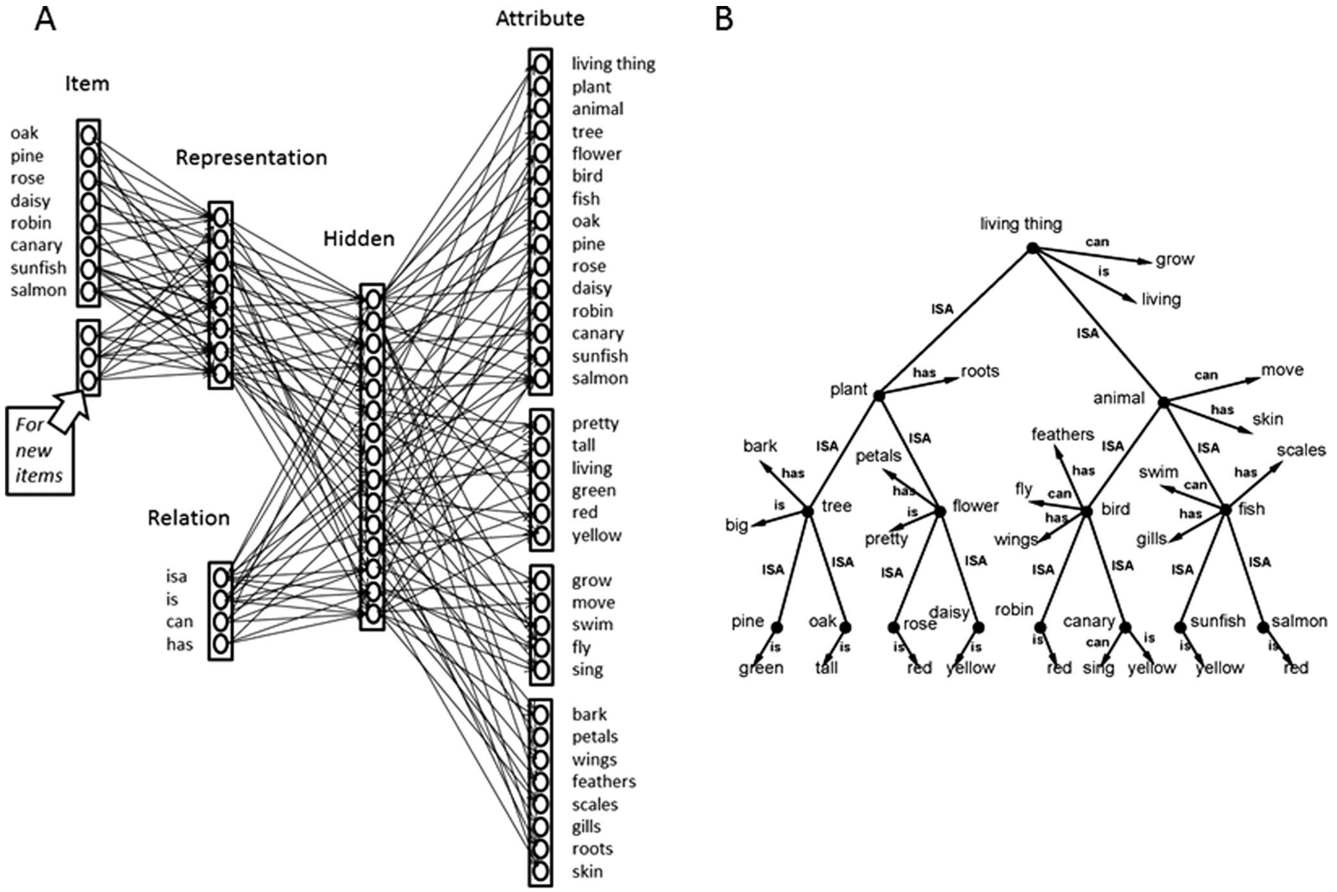

Figure 2. A: The artificial neural network model introduced by Rumelhart (1990) to explore learning of a structured body of knowledge about living things. Each oval represents a neuronlike processing unit. Connections between units are represented with arrows (only a subset of the connections is shown—each unit in a layer connects to all units in the next layer to its right). Testing occurs by activating an item (such as *canary*) and a relation (such as *can*), propagating activation rightward, and observing activations of output units. The network is trained on facts such as *canary can* {*grow, move, fly, sing*} by activating input units (left) corresponding to *canary* and *can,* propagating activation rightward, comparing the obtained output with the correct pattern of activation across the output units, then using back propagation (Rumelhart, Hinton, & Williams, 1986) to specify connection weight changes (see Simulation 1's Method section for details). B: The full set of facts used in training the network, organized into a hierarchy to emphasize the structure present in the training set (e.g., what a canary can do can be found at the end of an arrow labeled *can* at the node for canary or at any node that can be reached by following ISA links upward from *canary*). Panel A from *Semantic Cognition: A Parallel Distributed Processing Approach* (p. 56), by T. T. Rogers and J. L. McClelland, 2004, Cambridge, MA: MIT Press. Copyright 2004 by the MIT Press. Reprinted with permission. Panel B from *Semantic Cognition: A Parallel Distributed Processing Approach* (p. 6), by T. T. Rogers and J. L. McClelland, 2004, Cambridge, MA: MIT Press. Copyright 2004 by the MIT Press. Reprinted with permission.

units tended to have relatively low activation values (around .12), approximating the overall probability that an output unit would be active in a training example.

As in McClelland et al. (1995), a single extra input unit was added for the new to-be-learned animal, providing an extra input unit for the new item to be used after initial training (see Figure 2). The extra input unit was not used in the initial training of the network, but its connections to the representation units were initialized just as the connections from the other item input units were.

For the initial training, there were 32 training examples, each consisting of an input pattern and a target output pattern. The input patterns each specified an activation value of 1 for one of the eight item input units and for one of the four relations, and an activation of 0 for all other input units. The corresponding output pattern specified a target activation of 1 for correct values for the given item–relation combination, based on the information contained in Figure 2B, and a target activation of 0 for all other attributes. For example, for *canary can,* the correct values are {*move, grow, fly, sing*}.

Processing an item during training followed the standard procedure for back-propagation learning (Rumelhart, Hinton, & Williams, 1986). First, the input unit activations were set as specified in the current input pattern. Activation then propagated forward through the network via intermediate layers to the output layer (left to right through the network in Figure 2A). The obtained activation of the output units was then compared to the values specified in the current target output pattern. The resulting error signal was then propagated backward through the network, and finally, the connection weight adjustments were calculated and added into the connection weights.

Schema acquisition training. The network was initialized with random connection weights in each of the projections and for the bias weights to the representation and hidden units. These values were distributed uniformly in the range [−.45, .45]. The network was then trained for 1,000 epochs. All 32 training examples were presented for processing once in a new permuted order in each epoch. Connection weight adjustments were made after processing each example, using a learning rate of .1, no weight decay, and no momentum.

Learning new consistent and inconsistent information. Two training examples were made for each of three new items called *penguin, trout,* and *cardinal*. The one added input unit mentioned above was used as the item input unit in all three cases since each new item was used in a separate run of the simulation. One training example combined the new item with the *isa* relation, and the other combined it with the *can* relation, as in McClelland et al. (1995). The target output for *penguin-isa* and *cardinal-isa* was {*living-thing, animal, bird*}; for *trout-isa,* the output was {*living-thing, animal, fish*}. No target value was specified for the eight item-specific name output units. The output for *penguin-can* and *trout-can* was {*grow, move, swim*}; for *cardinal-can,* the output was {*grow, move, fly*}. Note that the *cardinal-can* and *cardinal-isa* patterns are consistent with previously learned birds and that the *trout-can* and *trout-isa* patterns are consistent with previously learned fish. The *penguin* is partially inconsistent with both, since its *isa* properties are those of a bird and its *can* properties are those of a fish.

Six different runs of the simulation were carried out, three using focused learning and three using interleaved learning. For each run, connection weights were initialized to the values obtained at the end of initial learning (see above). For the three focused learning runs, the network was trained only with the two training examples for one of the new items—the *penguin,* the *cardinal,* or the *trout*. Training proceeded as in the initial training described above, using the same parameters. After each epoch (one presentation of each of the two training examples), the network's performance was tested on the two new input patterns (e.g., *trout-isa* and *trout-can*) and on the *isa* and *can* examples for each or the four animals in the initial training set (hereafter called interference items). No connection adjustment occurred during testing.

For the three interleaved learning runs, the network was trained with the two training items from one of the new animals together with all 32 initial training items. Testing occurred after each epoch, as in the focused runs, and the performance measures were also the same.

Results and Discussion

The results of Simulation 1 demonstrate the importance of schema consistency in new learning within the Rumelhart network. As shown in the top panels of Figure 3, focused learning of schema-consistent information (*trout* or *cardinal,* solid curves) can occur rapidly in the network (left) while producing little interference with prior knowledge of other concepts (right). In comparison, replicating simulations in McClelland et al. (1995), focused learning of schema-inconsistent information (*penguin,* dashed curves) occurs more slowly and interferes with existing knowledge. The performance measure (*y*-axes in Figure 3) is the average over the relevant examples (e.g., *trout-isa, trout-can*) of the sum over the critical output units of the absolute value of the error. The error is the difference between the activation of the unit produced in processing the example and the correct or target value for the unit. The critical output units are those corresponding to *fish, bird, fly,* and *swim* (little error occurs at other output units). The performance values in the runs using the *trout* and the *cardinal* were similar to each other and were averaged to facilitate comparison with the values for the training run using the *penguin*.[2] The horizontal line near the top of each panel represents the average absolute error over the same units for known *animal-can* and known *animal-isa* items at the end of schema acquisition training on the original Rumelhart training items. The known animals are *canary, robin, salmon,* and *sunfish*.

The results just presented are the critical ones for making the point that new schema-consistent information can be assimilated into neocortical networks quickly and without interfering

[2] The Rumelhart network is a deterministic feed-forward network; the only elements of randomness in it are in the initial values of the starting weights and in the order of pattern presentations during training. Thus, while the details of the time course vary slightly from run to run of the network, the pattern of performance is very similar from run to run. The learned representations come progressively to capture the principal dimensions of variation in the input–output covariance structure present in the training examples, only varying slightly from run to run in the timing of learning each dimension (Rogers & McClelland, 2004; Saxe, McClelland, & Ganguli, 2013).

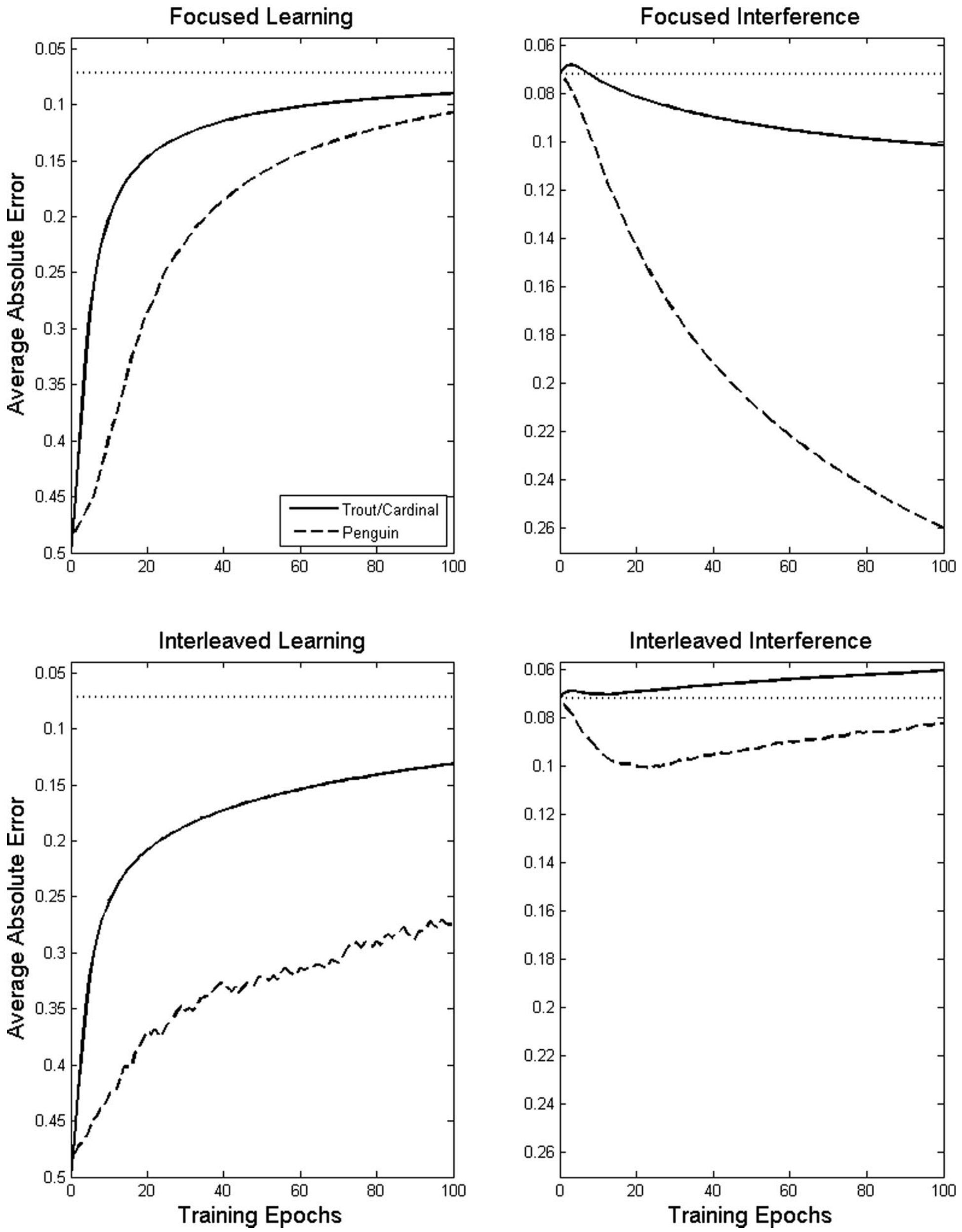

Figure 3. Simulated focused and interleaved results for learning new items (left panels) and the consequent interference with existing knowledge of other items (right panels). The *x*-axes show the number of training epochs. Each epoch involves one presentation of a new *animal-isa* and *animal-can* pattern either without (focused learning, top panels) or with (interleaved learning, bottom panels) ongoing exposure to the original training examples. The *y*-axes show the average across the relevant items (left: new *animal-isa* and new *animal-can*; right, known *animal-isa* and known *animal-can*) of the absolute per item error. Smaller error corresponds to better performance. The horizontal line near top of each panel represents the average absolute error for known *animal-can* and known *animal-isa* items at the end of network pretraining.

with existing cortical knowledge structures. For completeness, I also report results from interleaved learning (see lower panels of Figure 3). In this case, each training epoch included the two training examples for the single new item (*trout, cardinal,* or *penguin*) interleaved with the full set of 32 training items used in establishing a structured knowledge representation during pretraining. Interleaving retards learning new facts, even facts that are consistent with what the network already knows (as in the case of the *trout* and the *cardinal*; the reasons for this are discussed below in the section entitled Further Exploration of Changes in the Input-to-Representation Connections). The benefit is reduced interference (compare top and bottom right panels), but the interference is very low to begin with for schema-consistent information.

In summary, Simulation 1 demonstrates that new information consistent with a previously acquired schema can be assimilated rapidly into new cortical structures without interfering the network's representation of a previously acquired schema. If the first few training trials with either the *trout* or the *cardinal* were considered simulation analogues of initial exposure to Tse et al.'s (2007) new schema-consistent flavor–location trials in familiar environment (perhaps including a hippocampally mediated replay shortly after initial exposure, including replay during the test 24 hr after initial learning), the simulation might therefore be seen as capturing the rapid assimilation of new schema-consistent information into neocortical networks, as demonstrated in Tse et al. Correspondingly, the relatively slower learning of the (partially) schema-inconsistent *penguin* might be seen as a simulation analogue of the much slower acquisition of flavor–place associations in a novel environment. A fuller exploration of this issue was undertaken in Simulation 2.

Simulation 2

Building on Simulation 1, I conducted a second simulation designed to parallel the key experiment in Tse et al. (2011). That study examined the effects of new consistent and inconsistent learning on expression of *immediate-early genes* (IEGs) thought to be associated with synaptic plasticity in neocortex. Training with two new flavor–location pairs in a familiar environment led to both rapid learning and extensive gene expression, while training with six new pairs in an unfamiliar environment led to little or no learning and far less expression, further supporting the idea that new schema-consistent information is rapidly assimilated into neocortical structures, while new schema-inconsistent information is not.

The study included three conditions: a *new map* (NM) condition, in which six novel flavor–location pairs were presented in a previously unfamiliar environment; a *new paired associate* (NPA) condition, in which two new flavor–location associations were presented along with four known schema-consistent flavor–location items within the now-familiar environment; and an *old paired associate* (OPA) condition, in which the six already-known schema-consistent flavor–location pairs were presented, again in the familiar environment.

Animals were first trained for 6 weeks in the original Tse et al. (2007) environment, then divided into groups (n = 7 each) assigned to the three conditions described above. The key manipulation occurred in a single critical session. At the start of the session, four of the associations were presented over a 30-min period, then after 180 min in the home cage, two more associations were presented (see Figure 4 for details). Eighty min later, animals received an unrewarded probe test. They were sacrificed 5 min later, and tissue in several brain areas was assessed for IEG expression. Seven additional caged control animals who received no maze exposure served as a control for baseline levels of gene expression. As expected, animals in both the OPA and NPA groups performed well on old associations, animals in the NPA group exhibited learning of the two new associations in the recall test, and animals in the NM condition showed little or no learning in the novel environment.

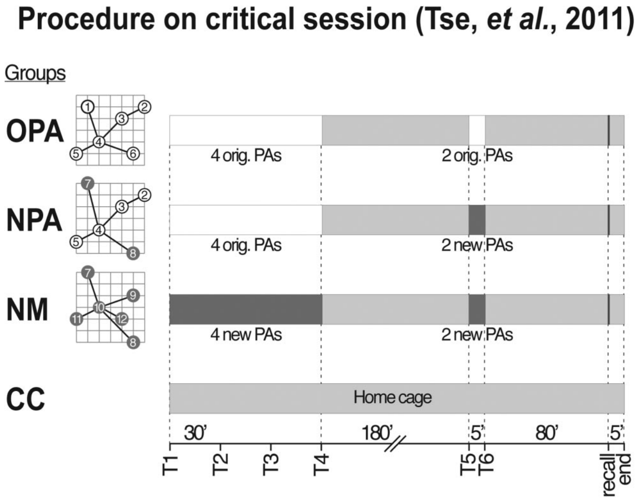

Figure 4. Experimental paradigm used in the experiment of Tse et al. (2011). Light grey indicates time spent in home cage; dark grey indicates time in the (new or old) experimental environment during which exposure to new flavor–location pairs occurred; white indicates time in the old experimental environment during which exposure to old paired associates occurred. Grids indicate locations of buried rewards associated with old (1–6) and new (7–12) flavors. orig. PA = original paired associate; OPA = old paired associate condition; NPA = new paired associate condition; NM = new map condition; CC = caged control condition; T = time. Reprinted from "Schema-Dependent Gene Activation and Memory Encoding in Neocortex," by D. Tse et al., 2011, *Science, 333,* p. 892. Copyright 2011 by the American Association for the Advancement of Science. Reprinted with permission.

The key finding is that the NPA training resulted in far greater expression of IEGs at several neocortical sites than did the NM condition (results from one site are shown in Figure 5), in spite of the fact that far more new information was involved in the NM condition (six new pairs in a new environment) compared to the NPA condition (two new pairs) or compared with the OPA condition. The findings suggest that large changes occur at neocortical synapses after one exposure to new schema-consistent associations but much smaller changes occur after reexposure to already-known information or after one exposure to new associations in a setting inconsistent with the previously acquired schema. Simulation 2 examined whether a similar pattern would be observed in the Rumelhart model. That is, Simulation 2 examined whether exposure to new schema-consistent information produces larger changes in the strengths of connections in the Rumelhart network than either reexposure to already-known information or exposure to new, schema-inconsistent information.

Method

Pretraining and test materials. I created analogues of Tse et al.'s (2011) three conditions as follows. First, I pretrained a new instance of the Rumelhart network with three added item input units following the same procedures as in Simulation 1. Only the first eight item input units were used during initial training. The resulting connection weights were saved for use in the new simulations. For the simulation analogue of each of Tse et al.'s conditions, I used a total of six training examples including an

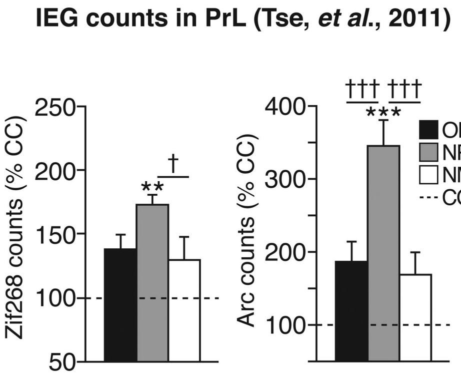

Figure 5. Gene expression results from Tse et al. (2011) for two different markers of induced synaptic plasticity for the prelimbic (PrL) region of neocortex. Results are shown separately for each of three different learning conditions: known consistent items (black), new consistent items (grey) and new inconsistent items (white). Error bars indicate the standard error of the mean. The *y*-axes represent the counts of the indicated marker relative to counts obtained from caged control (CC) animals (dashed horizontal line). The new consistent condition produced significantly higher counts compared both to baseline and to either of the other conditions for Arc and compared to baseline and the NM condition for Zif268. Significance levels relative to CC indicated by *s; significance relative to other conditions indicated by †s (one symbol, $p < .05$; two symbols, $p < .01$; three symbols, $p < .001$). IEG = immediate-early genes; OPA = old paired associate condition; NPA = new paired associate condition; NM = new map condition. Reprinted from "Schema-Dependent Gene Activation and Memory Encoding in Neocortex," by D. Tse et al., 2011, *Science, 333,* p. 892. Copyright 2011 by the American Association for the Advancement of Science. Reprinted with permission.

item-isa and an *item-can* example for each of three items. For the NPA condition, two of the examples came from a new schema-consistent item, and four came from two different already-known items. There were two variants of this simulation: In one variant, the network was trained with the *trout-isa* and *trout-can* items described above together with the *isa* and *can* items for two of the animals from the initial training set (*canary* and *salmon*). In the other variant, the *cardinal-isa* and *cardinal-can* examples replaced the *trout* examples. For both variants, the first of the three added units was used as the input unit for the new animal. Two runs were conducted with each variant. For the OPA condition, the network was trained on the *isa* and *can* examples from three known animals (*canary, salmon,* and *sunfish*). Four runs were conducted, each using a different random sequence of items within each epoch.

A number of possible approaches can be taken to thinking about the most appropriate analogue for the Tse et al. (2011) NM condition. In the NM condition of their experiment, the arena was placed in a novel location in their laboratory so that extramaze cues were completely different, and the experiment employed a new arrangement of foodwells as well as novel landmarks at new locations within the arena. However, there were still many residual similarities between the original and new conditions—the arena itself had the same geometry and was made of the same materials, rewards were found by digging to find foodwells under a bed of sawdust, the start box was the same and was always placed in the middle of one side of the arena, and general conditions of gravitation, atmosphere, illumination, temperature, and so on were not different between the two environments. This led me to construct new items for use in a simulation analogue of this condition that maintained the completely general features (*isa-living-thing, can-grow*) from the original environment but otherwise recombined item features in impossible ways. Accordingly, a set of six new training examples were constructed: Each set contained one *isa* and one *can* example for each of three new items called *xyzzx, yzxxy,* and *zxyyz*; each of these items was assigned one of the three added input units. Each item included the *isa-living-thing* attribute and the *can-grow* attribute but otherwise contained a mixture of characteristics of plants and animals. The patterns were *xyzzx-isa-*{*livingthing-plant-fish*}, *xyzzx-can-*{*grow-move-sing*}, *yzxxy-isa-*{*livingthing-animal-tree*}, *yxzzy-can-*{*grow-fly*}, *zxyyz-isa-*{*livingthing-plant-flower*}, and *zxyyz-can-*{*grow-move-swim*}. Alternative analogues to the NM condition are considered after analysis of this version of the NM condition. As with the OPA condition, four runs were conducted, each using a different random sequence of items within each epoch.

Simulation procedure. For each run of each simulation, as in Simulation 1, connection weights were set at the beginning of each run to the values obtained at the end of pretraining. The network was then trained for five epochs, with all six examples from the given environment presented once per epoch, considered to be based on one direct experience during a single exposure as in the Tse et al. (2011) experiment, together with a few possible replays from the hippocampus that might have occurred shortly after this single exposure. I then examined the extent of change to incoming connection weights (including bias weights as well as connections from other layers) to the representation, hidden, and output layers. Although Tse et al. inserted a delay between the first four and last two training items, it is possible that any synaptic plasticity induced by the first four as well as the last two items in the test session contributed to the observed IEG levels, and some replay may well have occurred throughout the time between the start of the test session and the moment when the animals were sacrificed. Thus, I examined the totality of cortical changes that might have been produced by any of the items used in the critical session. The extent of change was indexed by the mean over the set of included connection weights of the square of the difference between the value of the weight before and after the five epochs of training.

Results and Discussion

Though interest in this simulation focuses on the changes in connection weights as a result of new learning, it is important to begin by assessing how the network's performance changed over the five epochs of exposure in the simulation. As expected, learning occurred quickly for new schema-consistent information in the NPA condition but progressed far more slowly for the schema-inconsistent information in the NM condition. To assess learning of the specific featural information associated with the new items in the NM and NPA conditions, I examined the output error at the critical *tree, flower, bird,* or *fish* unit for the *item-isa* input patterns or at the critical *swim, fly,* or *sing* unit for the *item-can* input patterns. For all new patterns, these critical units were activated relatively weakly by the input pattern before any learning of these patterns had occurred. For example, *trout-isa* and *xyzzx-isa* both activated *fish* to the same weak extent (∼.25), so that the absolute value of the error at the critical unit for this output unit was about .75 in both cases; similarly, *trout-can* and *zxyyz-can* both activated *swim* to the same weak extent (∼.15) before learning. As expected, these values changed quickly for the *trout-isa* and *trout-can* items in the NPA runs where the *trout* items were used and similarly for the *cardinal-isa* and *cardinal-can* items in the NPA runs where these items were used. Overall, before learning started, the average across runs and patterns of the absolute error at the relevant critical output unit was .81; after five epochs, it dropped to .47. In contrast, these values changed very little for the *xyzzx, yzxxy,* and *zxyyz* items. In this case, before learning started, the average across runs and items of the absolute error at the relevant critical output unit was .862; after five epochs, it dropped only a little, to about .825.[3] As expected, performance changed little over the five epochs either in the OPA condition (average across runs and items of the largest error across all the units listed above dropped from .17 to .15 in the OPA condition and from .20 to .17 for the old items in the NPA condition).

To examine the magnitude of connection weight change in the network, the four runs for each of the three conditions were averaged to produce the change index values shown in Figure 6. Results are presented separately for the connection weights coming into the representation units (these include the bias weights to these units, as well as the weights from the input units to the hidden unit), the weights coming into the hidden units (including these units' bias weights, the representation-to-hidden connections, and the relation-to-hidden connections), and the weights coming into the attribute output units (including these units' bias weights and the hidden-to-output connections—refer to Figure 2 for a visualization of the connections). For runs using identical patterns, there were very slight variations between runs due to variation in the order of pattern presentation (standard deviations of change scores across runs with identical patterns were calculated separately for each set of runs using the same patterns; the largest of these standard deviations was 1.43×10^{-5}). Results differed slightly for the two variants of the new consistent condition (*trout* vs. *cardinal*). The two contributing values are indicated by the error bars shown in the figure.

As the figure indicates, exposure to new schema-consistent information resulted in large changes in the connections into the representation layer in Rumelhart network, but schema-inconsistent information produced relatively little effect on these connections. The figure reveals that, for these connections, far more change occurred in the NPA condition than in the NM condition. This is true even though, as in Tse et al. (2011), there were only two new patterns in the NPA condition while there were six new patterns in the NM condition and even though the values of all parameters including the learning rate parameter were the same in both simulations. Also, as previously noted, Tse et al. attempted to isolate the IEG activity produced by the last two patterns in each condition by inserting a delay between the first four patterns and the last two patterns. As a result, their measurements may have excluded some of the IEG expression induced by four of the six associations in the NM condition. Thus, the simulation likely overrepresented the amount of synaptic change their analysis would have captured in the NM condition. Presentation of known patterns produced hardly any change (see bars in graph for the OPA condition) since these items were already very well learned, indicating that the change in connection weights observed in the NPA condition was almost entirely due to the presentations of the two new but consistent training items (either *trout-isa* and *trout-can* or *cardinal-isa* and *cardinal-can*).[4]

The dramatic difference between the NPA and NM conditions occurred in the weights coming into the representation units from the item input units. It is interesting that the change occurred precisely where it was needed for learning the new schema-consistent items: The network learns the new patterns primarily by assigning connection weights that map these patterns onto representations already used for known birds (in the case of the *cardinal*) or fish (in the case of the *trout*). This cannot work for the patterns in the NM environment because it has not acquired representations that can support their output patterns, which are highly inconsistent with the outputs for known items.

Changes in other connections were more similar for the NPA and NM conditions—in fact, for the NM condition, there was more change in the connections to the output units than in the NPA condition. The effect is not dramatic—it is about the same on a per-new-pattern basis in the NPA and NM conditions. Nevertheless, it still reflects, in my view, a possible shortcoming of the Rumelhart model as an adequate model from the point of view of the desirability of avoiding interference when exposed to information that is inconsistent with prior knowledge. It is changes in the representation-to-hidden connections and particularly the hidden-to-output connections that lead to interference with knowledge of known patterns. By the design of the Rumelhart network, changes from the localist input units to the representation units cannot produce interference with performance on other items, since the connection weights involved are not shared across different items. However, all of the connection weights forward from the representation layer are shared, and thus, it is the changes made to these weights to accommodate new schema-inconsistent items that can lead to catastrophic interference. I return to this issue below.

Further Exploration of Changes in the Input-to-Representation Connections

Across the two reported simulations, something that might seem mysterious occurred. In spite of using the same learning rate parameter and the same number of presentations, learning occurred far more rapidly for items consistent with what was already known than for items inconsistent with what was already known. This rapid learning was exhibited in terms of the output of the network (faster reduction in output error) and by larger changes to the connection weights that allow the network to link new items into its existing knowledge structure (the input-to-representation connections). Why did this occur? The reasons are that (a) the target patterns for a new schema-

[3] There were only very minor differences across runs due to the random sequence of training items for all conditions, but there were consistent differences for different output features. Specifically, the initial activation was weakest for *sing* in new *item-can* inputs, since only one object known to the network could sing. Such differences can explain the tendency for the initial error to be slightly larger on average for the items used in the NM condition compared to the new items in the NPA condition. However, the change in error at the critical output unit was small for all items in the NM condition (mean change = −0.037; range 0.020 to −0.086).

[4] Some readers may note that some evidence of IEG induction over cage-control baseline occurred for the OPA group but that hardly any change at all occurred in the simulation in the OPA condition. This may reflect greater overtraining of the Rumelhart network or the absence of weight decay in the model. Either difference would lead to less error in the model's output with old items than there was in Tse et al.'s (2011) animals in the OPA condition.

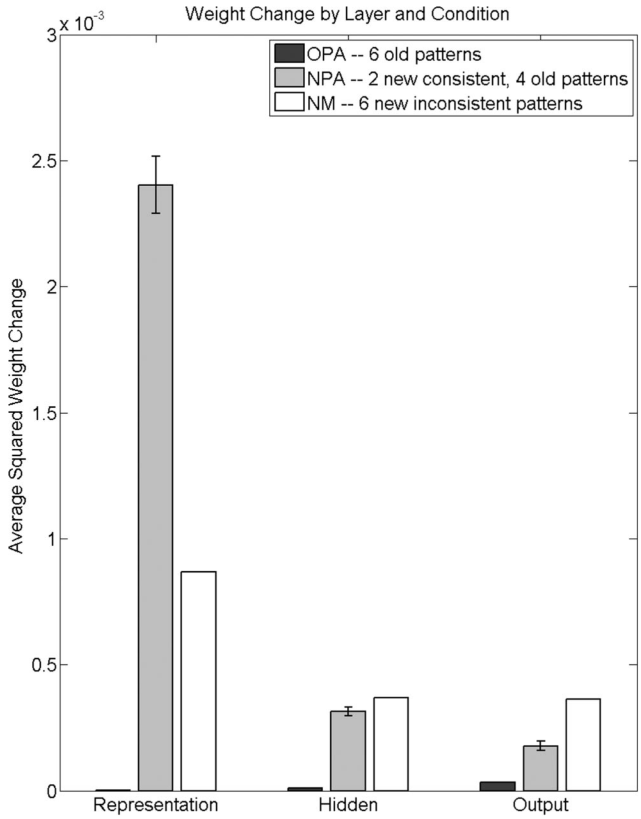

Figure 6. Total amount of connection change (sum of squared differences between before and after) in connection weights coming into different layers of units in the Rumelhart network, after five presentations of six new inconsistent training examples corresponding to Tse et al.'s (2011) NM condition (dark grey bars), two new consistent examples and four known examples corresponding to the NPA condition (light grey bars), or six known examples corresponding to the OPA condition (white bars). A single set of six training examples was used in four runs of the NM condition and similarly for the OPA condition. For the new consistent condition, two sets of six training examples were used in two runs each. One set contained the *cardinal-isa* and *cardinal-can* examples, along with the *isa* and *can* examples for *canary* and *salmon*. In the other set, the *trout* examples were used with the *canary* and *salmon* examples. The error bars show the range of weight change magnitudes in the NPA condition; the variation reflects slight differences between the *trout* and *cardinal* items (*cardinal* induced slightly more change in the connection weights than *trout*). Variability between runs with the same input patterns is too slight to visualize on this scale. OPA = old paired associate; NPA = new paired associate; NM = new map.

consistent item (e.g., the *trout*) generate error signals at the output layer that propagate coherently back to the representation layer, thereby efficiently cumulating to drive learning at these connections, and (b) learning at these connections is sufficient to produce correct outputs for new consistent items. The changes to these connections rapidly reduce the network's output error for the new items, so that less change ends up occurring elsewhere.

To demonstrate that the target patterns for a new schema-consistent item generate error signals that propagate coherently, I examined a measure of the consistency of weight changes incoming to the representation layer in the three conditions of Simulation 2. The measure of consistency for a given condition is the total net change to these connection weights over the five epochs of training divided by the sum of the presentation-by-presentation changes to these connection weights.[5] The logic of this analysis is simple: If the changes are perfectly coherent and efficient, then the sum of the individual changes should exactly equal the total change. If, on the other hand, the changes are incoherent and mutually contradictory, then the individual changes will tend to cancel each other out, as connection weights vacillate back and forth from item to item. The measure of change used is the sum, across all of the relevant connections, of the absolute value of the difference between the connection weight's value before versus after the single learning trial or before versus after the entire ensemble of learning trials. To be explicit, the measure of total net change was

$$TC = \sum_{i,j} \left| w_{ija} - w_{ijb} \right|,$$

and the measure of the sum of the individual changes was

$$SC = \sum_{p,e} \sum_{i,j} \left| w_{ija} - w_{ijb} \right|,$$

while the value of the efficiency index was simply the ratio of these two measures:

$$EI = TC/SC.$$

In these equations, the indices *i, j* index the individual connection weight, *a* and *b* index after versus before the full set of training trials (*TC*) or before versus after the individual training trial (*SC*), and *p* and *e* index the individual pattern and epoch in which the adjustment occurred (*SC* only). The finding from this analysis is clear: Efficiency is very high in the NPA condition (*EI* = .93), and much lower for the NM condition (*EI* = .27).

To demonstrate that learning in the input-to-representation connections is sufficient to produce correct outputs for new consistent items, I froze all connection weights in the network except those from a single new item input unit to the representation layer; then, I presented the *trout-can* and *trout-isa* training examples, allowing learning to occur only at the connections from the new input unit dedicated to *trout* to the representation hidden layer. While learning was slower than in Simulation 1, after 20 epochs of training, all output units were closer to the correct than the incorrect value (0 or 1) for this item. I repeated the simulation with similar results for the *cardinal.* I also repeated the simulation for the *xyzzx, yzxxy,* and *zxyyz* items and for the *penguin,* and in these cases, the network was unable to get all of the output values closer to correct than incorrect, even after 100 epochs of training.

It is true in the network that changes to the input-to-hidden connections alone are not sufficient to fully learn even the completely schema-consistent new items—even for these items, some adjustment to connections elsewhere in the network is needed, and indeed, the error signals generated during presentations of consistent items do call for changes to connections elsewhere in the network, albeit to a lesser degree than with inconsistent items. These are the reasons why learning based on input-to-hidden connections alone is slower than learning based on allowing

[5] I thank Ken Norman for suggesting this analysis.

changes to all connections. While these additional changes facilitate learning of the new consistent items, they can produce slight interference with some related items (e.g., changes to these connections tend to push the network toward treating similar items as more troutlike). If learning is interleaved, such changes tend to be cancelled by ongoing exposure to these other items in the training set—this is why interleaving slightly retards acquisition of new schema-consistent items, as observed in Simulation 1.

In summary, error signals supporting the linking of new items into the existing knowledge structure in the network propagate coherently and are sufficient to link the new item into the knowledge structure. In contrast, error signals generated by the schema-inconsistent items used in the simulations are both inconsistent and insufficient to allow the new items to be learned. It may be worthwhile noting that the cancellation of error signals can occur both within and across training examples. For the *penguin,* it occurs across the *isa* and *can* examples. The target output for *penguin-isa* is consistent with known birds, while the target output for *penguin-can* is consistent with known fish—so within each, the error signals are coherent, but across them, they partially cancel out. The same is true for *zyxxz,* whose *isa* properties are consistent with a *flower* but whose *can* properties are consistent with a *fish.* But for the *xyzzy* and the *yzxxy* items, there are inconsistencies with prior knowledge both within and between the two training examples (*xyzzy-isa-*{*livingthing-plant-fish*}; *xyzzy-can-*{*grow-move-sing*}; *yzxxy-isa-*{*livingthing-animal-tree*}; *yxzzy-can-*{*grow-fly*}). Either type of inconsistency can cause cancellation of error signals, resulting in weak changes at the input-to-representation connection weights.[6]

Simulation 2′: Alternative Simulation Analogues of the NM Condition

As a check on the generality of the finding that new schema-consistent information produces far more weight change in the weights coming into representation units than new schema-inconsistent information, I conducted additional simulations in the Rumelhart model with two additional possible NM analogue conditions.[7] In both, I selected the training example pairs used in the OPA condition of Simulation 2 (*canary-isa/can, salmon-isa/can, sunfish-isa/can*), to control for consistency among the training examples per se, since there is no reason to believe that Tse et al.'s (2011) NM environment was any less well structured as an environment than their original training environment.

The first of the new conditions, the *new network* (NN) condition, was based on the possibility that rodents' spatial representation systems might represent different environments using completely nonoverlapping sets of units and connections. Perhaps the intact hippocampus present during learning provides such nonoverlapping inputs to neocortex, and neocortex then recruits new neurons and connections to represent structured relationships within each environment separately. If so, the best simulation analogue would be to explore learning when a set of items is presented to a completely fresh, untrained network.

The second new condition explored the possibility that novel input and output units would be used to represent the features of the new environment but that these would ultimately have to rely on the same internal units, thereby potentially allowing for both interference and also knowledge sharing across the two environments (in Rogers & McClelland, 2008, we argued that the potential for such knowledge sharing across domains or environments is a major advantage of networks like the Rumelhart network). Accordingly, for this *new units* (NU) condition, I provided entirely new sets of input units (both for items and relations) and output units and connected these to the representation and hidden units according to the same policies as for the original input and output units. In this condition, I randomly initialized only the weights from/to these units, leaving the internal connections from the representation units to the hidden units and the bias weights to both the representation and hidden units unaffected. For comparison to the two new conditions, I also repeated the NPA condition.

Method. For all conditions, four simulation runs were conducted. The network architecture used in Simulation 2, with the same pretraining, was used again as the base for these simulations. For each run in each condition, as in Simulation 2, the six training examples used in a given condition were presented once per epoch in permuted order for five epochs.

The replication of the NPA condition employed the *trout-isa* and *trout-can* training examples in two runs and the *cardinal-isa* and *cardinal-can* examples in two runs; all four runs also included the *canary-isa* and *canary-can* examples as well as the *salmon-isa* and *salmon-can* examples. One of the three previously unused input units was used as the input unit for the *trout* or the *cardinal* as appropriate.

For the NN condition, the *isa* and *can* examples for *canary, salmon,* and *sunfish* were used as in Simulation 2's OPA condition. However, in this case, instead of using the weights acquired after pretraining, the network was reinitialized with random weights at the start of each run of the simulation. Four separate runs were conducted to allow assessment of the variability of results over different random starting weight values. It made no difference for this condition whether the units originally assigned to *canary, salmon,* and *sunfish* or the three previously unused input units were used for these items; for ease of implementation, I just used the units originally assigned to these items.

Conceptually, the NU condition required adding new sets of item input, relation input, and attribute output units, with fresh connections from the item input to the representation units, from the relation input to the hidden units, and from the hidden units to the attribute units, presenting inputs and targets only to these units during training and testing. However, for simplicity of implementation, no additional units or connections were actually added to the network. Instead, the input item, input relation, and output attribute units were reused, but the connection weights from the

[6] Considered literally, cancellation within a training example produces no weight changes in the input-to-hidden connections at the weight update after processing the item, while cancellation between training examples produces two weight changes that cancel each other out. If gene induction indexed the summed magnitude of weight changes, then one of these cases would produce gene induction, and the other would not. The fact that the cancellation is between examples in the simulation, however, should not be construed as a claim on my part that cancellation is a between-example matter in Tse et al.'s (2011) novel map condition. It seems likely that at every moment, the animal is receiving previously unfamiliar constellations of input features, something more like the within- than between-example situation. See below for further discussion of this point.

[7] I thank Dharshan Kumaran for suggestions leading to the exploration of these additional conditions.

item input units to the representation units, from the relation input units to the hidden units, and from the hidden units to the output units were reinitialized. Input and output conventions were the same as in the NN condition, and again, four separate runs were conducted to allow assessment of variability of results over different initial random weight values.

Results. In both the NN and NU conditions, only very weak changes occurred to the connections to the representation units (see Figure 7). The changes were far smaller than in the NPA condition, especially in the NN case. Both new conditions produced weaker changes than did the NPA condition in the connections coming into the hidden layer as well. However, both conditions exhibited relatively large changes in the connection weights from the hidden to the output layer of the network.

The reason for the very small changes in the connections to the representation layer is that the small random initial weights forward from these units propagate error signals very weakly and incoherently. The propagation is the weakest in the NN condition because both layers of weights forward of the representation layer are small and random in this case, whereas only the hidden-to-output weights are small and random in the NU condition. Error signals propagate slightly less weakly from the output units to the hidden units, since these signals must pass through only one matrix of weights in that case, but they still propagate relatively poorly.

The situation at the output layer is somewhat different. Given the small initial random weights, the activation patterns produced by different input patterns are initially very similar to each other (Rogers & McClelland, 2004) and only become differentiated slowly as learning progresses. Given this, early weight adjustments from internal units to output units primarily have the effect of producing a pattern of activation across the output units that is nearly the same for all input patterns and that matches each output unit's average target value across the set of training examples (McClelland, 2011, Chapter 5; Rogers & McClelland, 2004; Servan-Schreiber, Cleeremans, & McClelland, 1991). The five epochs of training used in these simulations are still within this early learning period; the changes to the output weights have moved the activations of the output units partway toward their average target value, and the output patterns are very similar for all six input patterns.[8] Such adjustments occurred to a lesser extent in the NPA condition and in the OPA and NM conditions of Simulation 2, since learning in these conditions occurred in a network that had already acquired connection weights encoding this general knowledge about each output unit's average target value, as well as more specific knowledge about the target activation values associated with each training example.

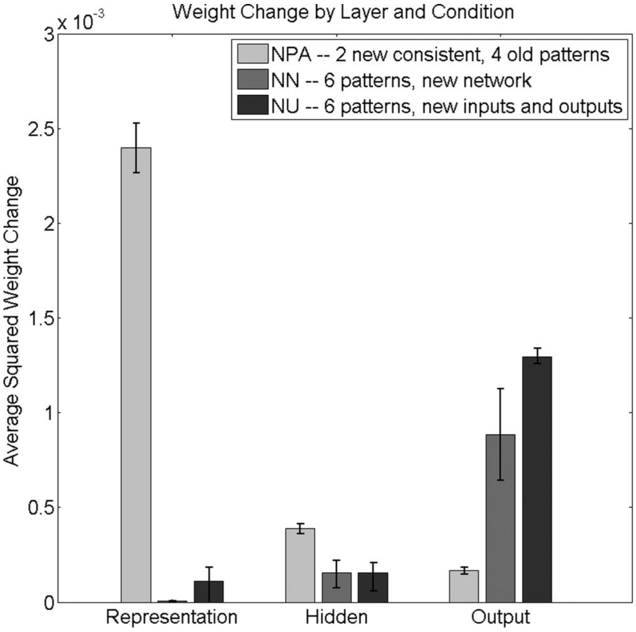

Figure 7. Comparison of weight changes in the new paired associates (NPA) condition compared to the new network (NN) condition and the new units (NU) condition. Error bars show the range of results produced over four runs of the simulation. As before, in the NPA condition, two runs involved the *cardinal-isa* and *cardinal-can* patterns, and two runs involved the *trout-isa* and *trout-can* patterns, along with the *isa* and *can* patterns for two old items. The patterns used in all of the NN and NU runs were the same, but each involved different random values for the newly initialized connections used in these two conditions.

In summary, although little change occurred in the connections into the representation or hidden layers in either the NN or NU conditions, both conditions produced fairly substantial changes in the connections to the output layer. Overall, as with the NM condition of Simulation 2, both new conditions still produced less overall weight change than the NPA condition, even though both the NN and NU conditions involved 3 times as many previously unfamiliar associations as the NPA condition. To varying degrees and in slightly different ways, therefore, any of these conditions might be viewed as capturing the reduced overall plasticity (as indexed by levels of IEG expression) in the NM condition of Tse et al. (2011) relative to their OPA condition.

On the other hand, the relatively high degree of plasticity induced in the output connections of the Rumelhart network in all three of the possible NM analogue conditions considered here does underscore a limitation of the network as a model of cortical learning. Allowing completely new experiences to restructure these connections is potentially deleterious, especially when (as in the NM condition analogue employed in Simulation 2) the output feature units are shared across environments. The findings from Tse et al.'s (2011) study are consistent with the idea that their NM condition induced no more change at cortical connections than did their OPA condition. This suggests that cortical networks may well be more fully buffered against rapid change on first exposure to a new environment than are the connections in the output layer of the Rumelhart network. These and other limitations of the Rumelhart network were the focus of one additional simulation.

[8] As evidence of this, I calculated the correlation of the pattern of activation produced by each training example with the average target activation value across all six training examples in one run of the NN condition. After five epochs of learning, all six correlation coefficients were between .54 and .57. The six patterns were all very similar, such that the smallest correlation among them was larger than .98. The evidence that the activations produced by the network had not yet converged to the average target values is that the average targets ranged from 0 to .5, while the activations varied only between .05 and .27.

Interim Summary

Simulations 1 and 2 demonstrated that new patterns consistent with a previously acquired schema or knowledge base of living things stored in the Rumelhart network can be learned quickly and result in large, concentrated changes to connection weights that integrate new information into the existing knowledge base without interfering with existing knowledge. Exposure to new information that is partially inconsistent with existing knowledge (such as the *penguin* used in Simulation 1 or the *xyzzx, yzxxy,* and *zxyyz* items used in the NM condition of Simulation 2) results in far less learning and far less overall change in connection weights. Simulation 2′ showed that learning in a completely fresh network results in changes in the network's output weights that capture the overall frequency that output units should be active in an environment but otherwise results in very slow learning at first, due to the weak propagation of error signals in initially unstructured networks. This tendency for error signals to propagate weakly is the primary basis for the tendency for learning to occur very slowly at first in multilayer back-propagation networks (McClelland, 2011; Saxe et al., 2013) and has been used to address stagelike progressions in development. While the simulation was not intended to address the slow learning seen in acquiring a set of flavor–place associations either in the original environment used by Tse et al. (2007, 2011) or in the second environment they used in their NM condition, the overall slowness of progress in learning in back-propagation networks does bear some similarity to the gradual pattern of schema acquisition exhibited in the Tse et al. experiments.

Limitations of the Rumelhart Network

The Rumelhart model, as useful as I believe it has been in many ways, nevertheless falls short in some respects. First of all, the model propagates activation in only one direction, and this runs counter to the view that perception, cognition, and memory are all highly interactive processes, requiring bidirectional propagation of activation, as Rumelhart and I have both argued (McClelland, Mirman, & Holt, 2006; McClelland & Rumelhart, 1981). Related to this, the model treats high-level aspects of the external environment as though they were directly represented, but other inputs and outputs from the environment likely undergo substantial preprocessing—something that I assume is highly experience dependent (consider, as just one example, the difficulty of perceiving nonnative speech sounds). Second, the back-propagation learning rule has been criticized on both computational and biological grounds, making it useful to consider simpler, perhaps more biologically plausible learning rules. Third, and more specifically relevant to the specifics of the issues raised in the Tse et al. (2007, 2011) studies, although the Rumelhart model partially addresses the problem of avoiding catastrophic interference when new learning is inconsistent with what is already known, it is imperfect in this regard, as indicated by the fact that even just a few presentations of schema-inconsistent items such as the *penguin* or any of the *xyzzy*-type items will begin to produce interference with prior knowledge. Possible fixes (allowing changes only in certain layers of weights or using different learning rates in different layers) could be proposed, but ultimately, it will be more satisfying to explore a model that addresses all of these limitations.

A fully adequate response to all of these considerations will require an extensive program of research, and is clearly beyond the scope of the current article. However, it may be useful to consider how a subset of the issues raised above might begin to be addressed using a very simple, arguably more biologically plausible network model. To this end, I present such a model, showing how connection weight change in the model can be buffered against large changes in response to inputs inconsistent with the structure of its previous experience, and point out how this could contribute to reducing the interference we have observed in the Rumelhart model if incorporated into a larger system.

For this demonstration, I used the *competitive learning* model (Grossberg, 1976; Rumelhart & Zipser, 1985), as implemented in the **pdptool** simulation environment (McClelland, 2011) with extensions that I have introduced with the goal of simultaneously increasing the biological realism of the model and leading it to exhibit experience dependence in activation of units in the network and therefore in learning. The goal of the analysis to show that this alternative to back propagation can exhibit (a) the gradual acquisition of sensitivity to the structure present in its training environment, (b) acceleration of learning as experience accumulates, and (c) a marked reduction of activation and return to slow learning if the network is exposed to a novel training environment.

Simulation 3

Method

For this simulation, I used the **cl** model as implemented in Version 2.07 of the **pdptool** software (McClelland, 2011), with two modifications to be mentioned below.

A **cl** network consists of two layers of units: an input layer and a competitive representation layer, with connection weights running from the input layer to the representation layer only. The network used here had 25 input units and five representation units (see Figure 8). Each representation unit receives a full set of connections from each input unit, initialized to random positive values in the [0, 1] interval and then normalized to sum to 1. Patterns from a training environment are presented in permuted order in a series of training epochs as in the **bp** program. Each pattern consists of a specification of activation values (either 0 or 1) for the units in the input layer only—there is no target pattern in a **cl** network. In each pattern presentation, the input pattern is used to set the activation values of the input units, and then activation is propagated to the units at the representation layer, where a net input is calculated for each representation unit using the standard formula $net_i = \Sigma_j a_j w_{ij} + bias_i$ (the inclusion of a bias is part of one of the modifications, as explained below). The representation unit with the largest activation is thought to be selected by a competitive, mutual inhibition process and is designated as the *winning unit.* The incoming connections to the winning unit from the input layer are then adjusted according to the following weight change equation (Rumelhart & Zipser, 1985):

$$\Delta w_{ij} = \varepsilon(a_i)(a_j/n - w_{ij}),$$

where a_i is the activation of the winning representation unit, a_j is the activation of the jth input unit, w_{ij} is the weight to the winning representation unit from the jth input unit, n is the number of active

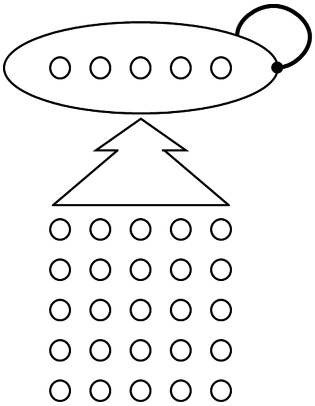

Figure 8. The competitive learning network used in Simulation 3. There is a connection from each input unit in the 5 × 5 input array to each unit in the representation layer of the network. Units within the representation layer (enclosed in oval) are mutually inhibitory and compete to represent the current input.

input units, and ε is the learning rate parameter. Note that this learning rule is Hebbianlike, in that it depends on the product of the activation of the receiving unit and another term that depends on the activation of the sending unit, and can exhibit analogs of synaptic plasticity phenomena including long-term potentiation and depression. This rule is simpler that other biologically supported learning rules (e.g., the BCM learning rule; Bienenstock, Cooper, & Munro, 1982; Cooper & Bear, 2012). It is likely that some versions of such rules would capture similar effects.

Two adjustments were made to the standard **cl** simulation model. First, instead of treating the activation of the winning unit as 1, the activation of the wining unit was set based on the *hedged softmax* function, a graded version of the simple *max* or winner-take-all function:

$$a_i = \frac{e^{\gamma net_i}}{\sum_{i'} e^{\gamma net_{i'}} + e^C},$$

where γ is a gain parameter and C is a positive constant. This function is often used in neural network and cognitive models (Kumaran & McClelland, 2012; Morton, 1969; Nosofsky, 1984). If connection weights have appropriate values, the resulting activations of units in the competitive layer can be seen as an estimate of the probability that the given input pattern is an example of the *i*th category of input patterns, where each competing unit plays the role of representing one of the alternative (mutually exclusive possible) categories. In the absence of the e^C term in the denominator, the normalization ensures that the sum of the activations of the units is equal to 1. The e^C term can be viewed as allowing for the possibility that the pattern is novel, so that one minus the sum of the activations of the units corresponds to the probability that the item is not a member of any of the known categories. The presence of this term causes hedging of the activation values, reducing them in accordance with this possibility, and has the effect of keeping all activations low when input patterns are unfamiliar and therefore net inputs are weak (see Kumaran & McClelland, 2012, for further discussion). This choice of activation function in the current context means that when net inputs are weak, connection weights will not change very much, since the activation of the winning unit is a factor in the weight update equation.

The second adjustment to the standard **cl** model was to introduce the bias term in the net input to each unit and to make this bias term activity dependent, so that the bias becomes positive if units are less active than others and negative if they are more active than others. This ensures equal utilization of units in the representation layer. Something like this mechanism, sometimes called *conscience,* has previously been used in other competitive-learning models (Desieno, 1988) and plays a role similar to the adjustment of the activity-dependent plasticity threshold in the BCM model (Bienenstock et al., 1982). The value of this bias term started at 0 for all units. The bias is incremented by the learning rate ε every epoch for all units, then decremented by the same amount each time a representation unit is chosen as the winner. Thus, if a unit wins more than once per epoch, its bias would tend to become negative, reducing its tendency to be chosen, while if it wins less than once per epoch, its bias will tend to become positive, increasing its tendency to be chosen.

Environments and Training Regime

I created two training environments for the network, each consisting of five recurring patterns. In the first, the five patterns each contained a row of five active units on the input layer shown in Figure 8, with one pattern for each of the five rows of the grid. In the second environment, the five patterns each contained a column of active input units, with one pattern for each of the five columns of the grid. These patterns can be viewed as capturing different patterns of feature co-occurrence in each of the two training environments. That is, the horizontal rows can be seen as one arbitrary clustering of subsets of input features; the vertical rows can be seen as a different arbitrary clustering of subsets of input features. By selection, the two clusterings are antithetical to each other, so that a learned model based on one environment (extracting the co-occurring clusters there) will prove unhelpful in the other environment. While real environments certainly have more complex structure than this, these two environments certainly have very different structure.

The network was trained first on the row environment for 100 epochs and then switched to the column environment. Each epoch consisted of one presentation of each of the five patterns in the environment; weights were updated after each pattern presentation. The parameter values used were as follows: γ = 1.7, C = 4, ε = .1. The results shown are based on a single run; other runs with different random initial connection weights produced similar results.

Results and Discussion

The network captures the critical property of familiarity dependence of learning. Previously unfamiliar input patterns initially produce very weak activations and correspondingly weak connection weight changes (see Figure 9). As these changes cumulate, activations become stronger, and connection weight changes gradually become larger, until a point is reached where the connection weights match the input patterns,

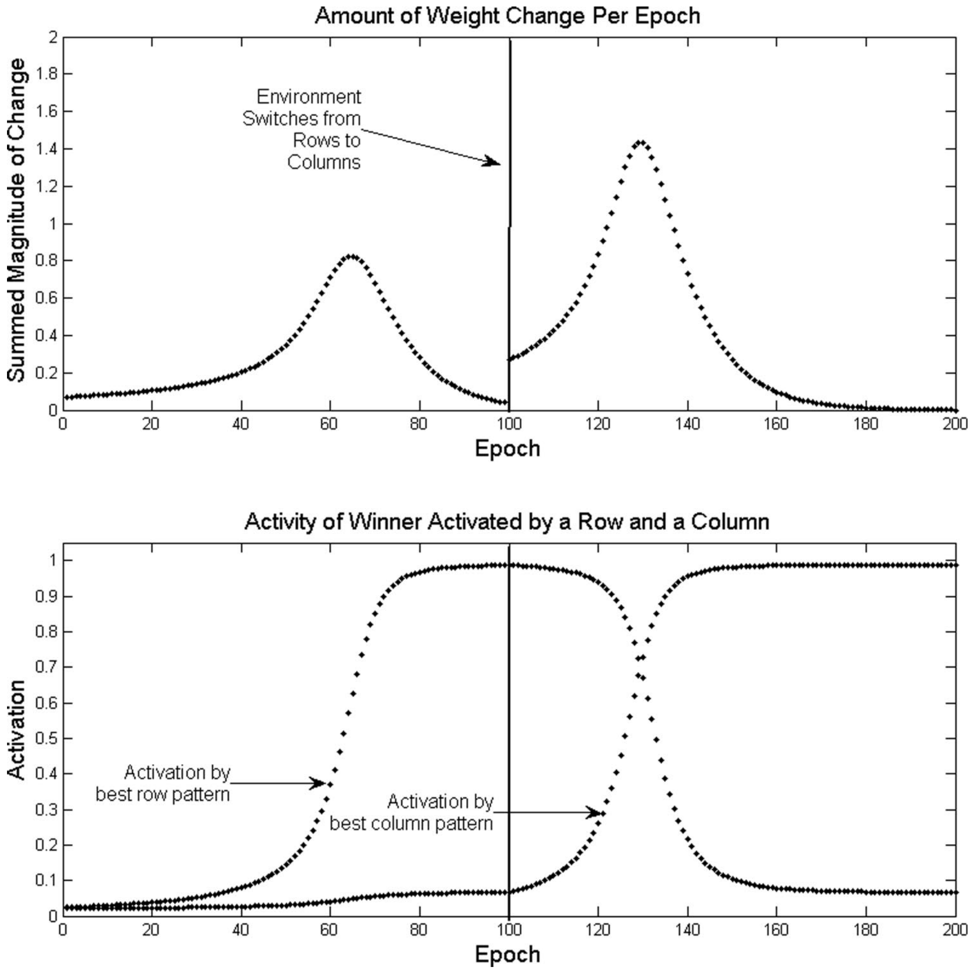

Figure 9. Effective learning rate (upper panel) and winning unit activation (lower panel) in the competitive learning network used in Simulation 3. Magnitude of weight change is the sum of the absolute values of the differences between weights before and after each epoch of training. Winning unit activation is based on the *hedged softmax* function for one example member of the row and column training set (all other units show a similar pattern) and is assessed after each epoch of training with learning turned off during the assessment. Note that after the switch from exposure to rows to exposure to columns, the network maintains connection weights that can process and represent the row patterns quite effectively even though it is now receiving training with the column patterns rather than the row patterns.

so that the weights no longer change, and activation values for the patterns in the training environment have reached maximal levels. At this point, the environment is switched. The pattern of intercorrelation of input pixels is now completely different. The new items each produce weak activations, though these activations are a bit larger than they were when the network was first initialized (the slightly increased activation results from the fact that one of the strong connection weights formed in learning the row inputs is engaged by each of the column inputs). The weaker activations in turn have the consequence that the resulting weight changes are relatively small at first as well (though again, not as small as when the network was freshly initialized). For a period of time after the switch to the column environment, then, the network continues to robustly respond when tested with the items from the row environment. Gradually, however, the network readapts its connection weights so that they become consistent with robust processing of the patterns in the new, column environment and inconsistent with processing of the patterns from the old, row environment.

The results of this very simple simulation address several of the issues raised in the Tse et al. (2007, 2011) experiments. First, they address to a degree how learning consistent with what is already known can progress more rapidly than new inconsistent learning, though in this case, we are exploring not the learning of new consistent versus new inconsistent items but the continued learning of consistent items versus the initial learning of new inconsistent items. A similar prior-knowledge dependence occurs in the Rumelhart model (Rogers & McClelland, 2004; Saxe et al., 2013) and other back-propagation networks, and I have suggested elsewhere (McClelland, 1989) that this pattern of prior-knowledge dependence of learning may help explain stagelike transitions in cognitive development (Siegler, 1976).

Perhaps more important for current purposes, the simulation demonstrates one simple way in which the cortex may be protected against allowing inputs completely inconsistent with the structure of previous experience to rapidly rewrite neocortical connections; at first, the totally new environment makes relatively little impact, since the new inputs produce weak activations, which in turn result in relatively small connection adjustments.

A second relevant point arises if we think of the little network simulated here as residing at the lowest level of a hierarchy of learned representations, such that the patterns produced across the representation layer of this network would have to be associated with patterns across the representation layers of other such networks to represent superordinate relationships (as would occur if one had to learn associations between previously unfamiliar objects to form a schema for a completely novel environment). In this case, there would be a further buffering of the system's prior knowledge from interference: The inputs to these higher levels of representation would themselves be sufficiently weak that virtually no learning at all would occur at the higher levels.

While this model captures several features of schema dependence in learning, it is certainly very drastically limited. One simple problem is the minimal number of units provided; in a more realistic system, there would be enough neurons available to represent both the rows and the columns patterns (in this case, the switch in environment would be expected to recruit, over time, some of the row units to represent the column inputs but leave others to continue to represent rows). More broadly, the model lacks the necessary multilayer structure of realistic brainlike networks, lacks the interactivity I consider to be an intrinsic characteristic of real neural networks, and likely oversimplifies many other aspects of realistic processing and learning. Its utility is only to point toward future models that could overcome some of these limitations and to illustrate in rudimentary form one aspect of how experience dependence and buffering against deleterious consequences of exposure to a radically restructured environment might arise outside the limitations of the particular assumptions of the Rumelhart network.

General Discussion

The simulations and analyses presented above address many of the key findings of Tse et al. (2007, 2011) and extend the analysis of neocortical learning within the CLST. In McClelland et al. (1995), my colleagues and I emphasized the idea that the neocortex should learn new information slowly, to allow the effective (if gradual) acquisition of structured knowledge and to avoid interference with structured knowledge previously acquired. Here, however, we see that acquisition of new information need not proceed slowly if it is consistent with what is already known. Indeed, the simulations reported here demonstrate that new information consistent with previously acquired structured knowledge can actually be learned quite quickly in an artificial neural network of the kind previously used to capture both the gradual emergence of a structured representation of knowledge in a putatively neocortexlike network and the interference with that knowledge that can occur when the same network is forced to learn new inconsistent information rapidly. The Rumelhart network allows new information consistent with what is already known to be learned quickly, producing little interference with what is already known. It is only information that is inconsistent with what is already known that must be learned more slowly to avoid catastrophic interference. Interestingly, the network's structure buffers it against the deleterious effects of exposure to items inconsistent with its previous experience. It tends to allow new information consistent with what is already known to be assimilated rapidly while avoiding being too strongly affected by new inconsistent information.

Tse et al. (2007, 2011) deserve credit for showing that schema-consistent information can be learned rapidly in real neocortical networks and for bringing out the importance of schema consistency for new learning. Their findings serve to uncover an important feature of neocortical learning that was not previously emphasized in the statement of the CLST. However, the challenge to the theory has now been addressed, relying on previously underemphasized properties of the neural network model previously used to illustrate the key elements of the theory: Specifically, Simulation 1 reported here demonstrates that the Tse et al. findings are compatible with the neocortical learning model used in the CLST: Like the rats in the Tse et al. study, the network acquires new schema-consistent knowledge rapidly and without interference with known items. It is only schema-inconsistent information that, if learned too rapidly (and without interleaving), produces catastrophic interference. Thus, the earlier statement of the core claim of the theory remains, although with a qualification: Complementary learning systems and interleaved learning are necessary when learning new information if that information is not consistent with the structure of known information.

Tse et al. (2011) also found that synaptic change in neocortex, as indexed by gene expression, is strongly induced by exposure to new schema-consistent, but not schema-inconsistent, information. Simulation 2 reported here produced a strikingly similar pattern. This pattern is highly adaptive, according to the CLST: If exposure to new inconsistent information without interleaving produced large changes, this would interfere with existing knowledge. Such changes are not problematic for new consistent items, since, as we saw above, these changes do not produce interference.

A crucial question arises: How does the brain know when to allow experience to produce large synaptic changes in neocortex? The Rumelhart model provides a partial answer to this question: Gradually acquired schema-based connections propagate strong, coherent learning signals about consistent information, inducing large connection changes; schema-inconsistent information generates signals that cancel each other out, so little net change is induced. It is also notable that early in training a Rumelhart networklike model, before a schema has been acquired, error signals are weak and incoherent, even for schema-consistent items. Such networks therefore explain the fact that what can be learned from an experience depends on the state of prior knowledge. Importantly, these features are not restricted to feed-forward networks trained with back propagation: Many similar properties are also exhibited in the arguably more biologically plausible competitive learning network used in Simulation 3.

Neocortical Learning: Not Fast or Slow, but Prior-Knowledge Dependent

A second important amendment in the statement of the principles of the CLST is required to accommodate the findings of

Tse et al. (2007, 2011) as well as the simulation findings presented here. In previous articles describing the CLST, learning in the neocortex was described as *slow,* and learning in the hippocampus was described as *fast.* This characterization is not really adequate, however, since the rate of learning (as indexed by a change in performance or in connection weights) can vary dramatically depending on the prior state of knowledge and the consistency of what is being learned with what is already known. Given this, it is clearly misleading to characterize neocortical learning as slow. Building a new knowledge structure and integrating new information inconsistent with existing structure are slow, but building on top of existing structure is not necessarily slow. Given the above, it may be useful to characterize neocortical learning not so much as fast or slow but as *prior-knowledge dependent.* I take this statement to be the principal conceptual advance contributed by the findings of Tse et al. (2007, 2011). I view it as an important amendment to the characterization of the CLST—a theory that I hope they and others will view as an evolving ensemble of principles to which their research has made an essential contribution.

It may, further, be useful to characterize learning in the medial temporal lobes as less dependent on prior knowledge—but with important caveats. Marr (1971) introduced the idea that the hippocampus may use sparse, random conjunctions of features or elements of experiences, thereby minimizing similarity (and therefore interference) between memories. This idea has been very useful and has subsequently been explored by many others (Gibson, Robinson, & Bennett, 1991; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Sutherland & Rudy, 1989; Treves & Rolls, 1992). Marr called the hippocampus a *simple* memory in that its rules of operation are very easy to define and learning is seemingly not dependent on prior knowledge. However, this characterization of learning within the medial temporal lobes, even if fully valid, may be misleading if we fail to take into account a principle articulated in McClelland and Goddard (1996; see also Kumaran & McClelland, 2012): that the features or elements that are made available to the hippocampus arise as a consequence of a gradual neocortical learning process. Given this principle, even a simple memory that is seemingly independent of prior knowledge in and of itself suddenly becomes completely knowledge dependent when placed in its proper context within the overall organization of systems of learning and memory in the brain. In this way, rapid assimilation of new information by the hippocampus may also become prior-knowledge dependent.

Another distinction that may prove heuristically useful is the distinction between *deep* and *shallow* neural networks. The difference between deep and shallow networks ultimately comes down to the idea that more than one hidden layer is needed to produce truly productive generalizations based on experience across disparate contexts. According to protagonists of deep learning (e.g., Hinton, in press), simply forming conjuncts as in Marr's (1971) simple memory is not a fully adequate approach. But learning in deep networks is not easy and (when the signals needed to learn must be passed through multiple layers or connections) results in a learning system in which the rate of learning is initially very slow (as the network begins to become sensitive to the structure in its training set) and is strongly dependent on the state of prior knowledge in the system. I believe the Rumelhart network captures some of the features of deep learning systems rather well—indeed, its use of a representation layer between the item input layer and the larger hidden layer makes it a (relatively simple) example of a deep network. In any case, it seems useful to consider the neocortex to be an example of a deep network. The hippocampus may still be a relatively shallow network, but by sitting atop the structured representations provided to it by the neocortex, it benefits from the representations learned in the deep cortical network.

Open Questions and Future Directions

While the stated amendments to the CLST may be advances, the theory is far from complete, and great deal of additional theoretical and empirical research remains to be done. Here, I raise a few of what I consider to be the principal issues.

What exactly can be learned quickly in neocortical neural networks? While the experiments in Tse et al. (2007, 2011) are important initial steps, much remains to be learned about the limits on what kind and amount of next learning can occur rapidly in neocortex. In Tse et al.'s experiments, animals mapped novel flavors onto locations in a familiar environment. While the locations were not identical to the locations to which flavors had been previously been mapped, they were very close to those locations, and the new wells replaced the old ones nearby (see Figure 1) so that search at the old locations might have led the animals to find the new wells. Thus, it is arguable, at least, that what the animals in the Tse et al. studies did was simply to map arbitrary new flavors onto established representations of well-known locations within a familiar environment. The simulations with the *trout* and the *cardinal* had very similar characteristics: The network learned to map an arbitrary new input (represented by the localist input unit assigned to either the *trout* or the *cardinal*) to outputs corresponding to previously learned items (the trained properties of *trout* are fully shared with both the *salmon* and the *sunfish,* while the trained properties of *cardinal* are fully shared with the *robin*). This is, perhaps, among the easiest kinds of things to learn. As long as the input representation of the new item to be mapped to an already-learned concept overlaps little with the input representations of known items, this is easy to do and produces little interference with existing mappings of other known items. A question arises, though: Are there other examples of new learning that can occur rapidly and with little interference? Further studies of this matter would potentially be illuminating. One specific question is the extent to which novel-structure-consistent conjunctions of inputs can be learned rapidly in human or animal neocortical networks. Suppose, for example, that a learner has been exposed to 20 items, organized into four sets of five, so that each member of Set A is paired with two members of Set B and each member of Set C is paired with two members of Set D. Additional associations between items in Set A and B or between items in Set C or D would be consistent with the regularities in the ensemble, while associations between A and D items or C and B items would not. Could new A-B or C-D but not new A-D or B-C associations be rapidly assimilated into the neocortex? Could new items be added to the A or C sets and mapped onto novel combinations of B or D items? Or would new items have to map to the same combination of associations as existing items? More generally, just what counts as schema-consistent information that can be rapidly learned by the neocortex?

What role does the hippocampus play in new schema-consistent learning? Another important question is the role that the hippocampus plays in rapid schema-consistent learning. Tse et al. (2007) found that once the hippocampus was removed, additional flavor–location associations could no longer be learned, suggesting that the hippocampus is necessary for adding new information to neocortical knowledge networks, at least in the setting of their experiment. But what exactly is this role? One possibility is that the hippocampus simply allows for several replays of new information shortly after initial exposure—in the absence of the hippocampus, only changes occurring during the exposure itself could occur, but with the hippocampus, rapid learning there coupled with replays shortly thereafter could help integrate the new information into neocortical networks. While this seems likely to be a contributing factor, it is also quite possible that the hippocampus plays other roles as well. Among others, it may well facilitate the construction of a unique representation for each environment an animal encounters, aiding the cortex in minimizing interference between environments that share similarities with each other. In any case, the nature of the role of the hippocampus in both initial learning of a representation of a new environment and in the addition of new consistent information to such a representation is an important issue for future investigation.

Toward more adequate models of neocortical learning. Although the Rumelhart network and other similar networks have proven useful in capturing many characteristics of human neocortical learning, I have already noted several ways in which the network likely differs from the neural networks in the brain. The simple competitive learning model introduced in Simulation 3 demonstrates that some aspects of the behavior of the Rumelhart network generalize to one example of a possibly more biologically plausible model, but it too has several limitations. Here, I consider just a few of the many points that will need to be addressed in making progress toward a more adequate model.

While the Rumelhart model uses feed-forward connections to propagate activation and feeds activation back through these connections to propagate error information, I have already noted that real brain networks likely utilize bidirectional connections between participating neurons within and between brain areas for activation propagation, and the competitive learning model does not address this limitation. Regarding error propagation, the situation is more complicated. Many of the characteristics of the Rumelhart model have been captured in recurrent neural network models with bidirectional propagation of activation and error information (Dilkina, McClelland, & Plaut, 2008; Rogers et al., 2004), but the back propagation of error information in these networks has never seemed biologically plausible, since it involves back propagation through time as well as through layers of units and connections.

It is important to acknowledge that the propagation of learning signals in the brain is not currently well understood. A great deal of research focuses on possible learning rules specifying how pre- and postsynaptic signals drive changes at synapses (see Shouval, 2007, for review), but most of this work focuses on information that is locally available, namely, pre- and postsynaptic activity and possibly the current value of the connection weight (as in the competitive learning model). What is unknown is whether and exactly how such signals can reflect global performance in a way that allows them to produce effective internal representations like those produced by back propagation. Some learning rules rely only on pre- and postsynaptic activity together with a more global third signal, possibly signaling sleep versus wake as in the Boltzmann machine (Ackley, Hinton, & Sejnowski, 1985), early versus late in a processing episode (Hinton & McClelland, 1988; O'Reilly, 1996), or reinforcement (Mazzoni, Andersen, & Jordan, 1991), and there are even proposals that some relevant global variation may arise from the peaks and troughs of neural oscillations (Norman, Newman, Detre, & Polyn, 2006). Some of these algorithms are effective for training bidirectionally connected networks such as those used in Rogers et al. (2004) and Dilkina et al. (2008). It is likely that simulations using some of these other learning rules would replicate the findings observed in the simulations reported here, since the weight change signals computed by such rules are similar to those computed using back propagation and their propagation from one layer to another depends of the structuring effect of prior learning and the consistency of the to-be-learned information with what was previously learned. It is also worth noting that although some form of error-correcting learning still appears critical to guide neural networks to find connection weights that fully solve demanding computational problems, current deep learning models often benefit from incorporating elements of both supervised and unsupervised learning (Hinton, in press; O'Reilly, 1996).

As a secondary issue, I hope that a future model will avoid the distinction between inputs and outputs that is one of my own frustrations with the Rumelhart network. Although it is true that experiments often involve cues to which participants make responses, much of learning may involve bidirectional associations between representations of entities such as items and locations, with responses to cues potentially involving responding by navigating to the location whose representation was retrieved by presentation of the cue.

Distributed input and output representations. The Rumelhart model uses *localist* input and output units corresponding to familiar human concepts—individual units correspond to objects, relations, and verbally labeled attributes of objects (the competitive learning model used in Simulation 3 suffers from a version of this same limitation), while to me (if not to others, cf. Bowers, 2009), it seems clear that the brain uses distributed representations, and such representations have been used in related models (Dilkina et al., 2008; Rogers & McClelland, 2004, Chapter 7). The use of a localist representation forces the modeler to add new units for each new item, relation, and attribute—something that was done for the new items in the simulations reported here but that I consider to be a violation of one of the key principles of neural representation.

An advantage of distributed representations is that a new pattern of activation over a set of units can be learned in a network without adding new units. For example, in Dilkina et al. (2008), patterns of activation corresponding to the spellings and pronunciations of words were used. The spelling or sound of a word is a particular pattern over a set of orthographic or phonological units. For the reported simulations to work using distributed item input representations, it would be necessary for these representations to be sparse and approximately orthogonal to existing item input representations. The input patterns used by Tse et al. (2007, 2011) may have these characteristics, since olfactory cortex is thought to provide sparse, orthogonalized representations for unique odors and flavors (Poo & Isaacson, 2009), and perhaps a future model could be constructed around this possibility.

Conclusion

Tse et al. (2007, 2011) have advanced our understanding by showing that the neocortex can acquire new schema-consistent knowledge quickly. Their work also shows that learning is far more gradual for inconsistent information. These findings have occasioned amendments to the stated principles of the CLST. Equally important, simulated neural network models such as the Rumelhart model capture these findings and indicate how different rates of acquisition of schema-consistent versus schema-inconsistent information may arise as emergent consequences of neocortical learning. While the Rumelhart model is not a fully adequate model of neocortex, an additional simulation using a competitive learning network points the way toward the possibility that more biologically plausible models may retain these important properties. Overall, the simulations presented here help to explain how schema-consistent information can be learned quickly by the neocortex, while schema-inconsistent information and new schemas will be learned more slowly.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147–169. doi:10.1207/s15516709cog0901_7

Bartlett, F. C. (1932). *Remembering*. Cambridge, England: Cambridge University Press.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2,* 32–48.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

Bransford, J. D. (1979). *Human cognition: Learning, understanding, and remembering*. Belmont, CA: Wadsworth.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Cooper, L. N., & Bear, M. F. (2012). The BCM theory of synapse modification at 30: Interaction of theory with experiment. *Nature Reviews Neuroscience, 13,* 798–810. doi:10.1038/nrn3353

Desieno, D. (1988), Adding a conscience to competitive learning. *Proceedings of the IEEE International Conference on Neural Networks, 1,* 117–124.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology, 25,* 136–164. doi:10.1080/02643290701723948

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209. doi:10.1016/0010-0277(86)90034-X

Gibson, W., Robinson, J., & Bennett, M. (1991). Probabilistic secretion of quanta in the central nervous system: Granule cell synaptic control of pattern separation and activity regulation. *Philosophical Transactions of the Royal Society of London: Series B. Biology*, *332,* 199–220.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134. doi:10.1007/BF00344744

Hinton, G. E. (in press). Where do features come from? *Cognitive Science*.

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

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression of semantic dementia: Implications for the organisation of semantic memory. *Memory, 3,* 463–495. doi:10.1080/09658219508253161

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review, 119,* 573–616. doi:10.1037/a0028681

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development, 6,* 17–46. doi:10.1016/0885-2014(91)90004-W

Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive Development, 3,* 247–264. doi:10.1016/0885-2014(88)90011-1

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London: Series B. Biological Sciences*, *262,* 23–81. doi:10.1098/rstb.1971.0078

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *PNAS: Proceedings of the National Academy of Sciences, USA, 88,* 4433–4437. doi:10.1073/pnas.88.10.4433

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 8–45). New York, NY: Oxford University Press.

McClelland, J. L. (2011). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises* (2nd ed.). Draft. Retrieved from http://www.stanford.edu/group/pdplab/pdphandbook

McClelland, J. L., & Goddard, N. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus, 6,* 654–665. doi:10.1002/(SICI)1098-1063(1996)6:6<654::AID-HIPO8>3.0.CO;2-G

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. doi:10.1037/0033-295X.102.3.419

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 363–369. doi:10.1016/j.tics.2006.06.007

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407. doi:10.1037/0033-295X.88.5.375

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. 109–165). New York, NY: Academic Press. doi:10.1016/S0079-7421(08)60536-8

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10,* 408–415. doi:10.1016/0166-2236(87)90011-7

Mervis, C. A., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development, 53,* 258–266. doi:10.2307/1129660

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76,* 165–178. doi:10.1037/h0027366

Norman, K. A., Newman, E. L., Detre, G. J., & Polyn, S. M. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation, 18,* 1577–1610. doi:10.1162/neco.2006.18.7.1577

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114. doi:10.1037/0278-7393.10.1.104

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8,* 895–938. doi:10.1162/neco.1996.8.5.895

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus, 4,* 661–682. doi:10.1002/hipo.450040605

Patterson, K., Lambon Ralph, M. A., Jefferies, E., Woollams, A., Jones, R., Hodges, J. R., & Rogers, T. T. (2006). "Presemantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience, 18,* 169–183. doi:10.1162/jocn.2006.18.2.169

Poo, C., & Isaacson, J. S. (2009). Odor representations in olfactory cortex: "Sparse" coding, global inhibition and oscillations. *Neuron, 62,* 850–861. doi:10.1016/j.neuron.2009.05.022

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111,* 205–235. doi:10.1037/0033-295X.111.1.205

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2008). Précis of *Semantic Cognition: A Parallel Distributed Processing Approach*. *Behavioral and Brain Sciences, 31,* 689–714. doi:10.1017/S0140525X0800589X

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornitzer & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). San Diego, CA: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, October 9). Learning representations by back-propagating errors. *Nature, 323,* 533–536. doi:10.1038/323533a0

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, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9,* 75–112. doi:10.1207/s15516709cog0901_5

Saxe, A., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical categories in deep networks. In M. Knauff, M. Paulen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 1271–1276). Austin, TX: Cognitive Science Society.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7,* 161–193. doi:10.1007/BF00114843

Shouval, H. Z. (2007). Models of synaptic plasticity. *Scholarpedia, 2*(7), 1605. doi:10.4249/scholarpedia.1605

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520. doi:10.1016/0010-0285(76)90016-5

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology, 17,* 129–144.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus, 2,* 189–199. doi:10.1002/hipo.450020209

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., . . . Morris, R. G. M. (2007, April 6). Schemas and memory consolidation. *Science, 316,* 76–82. doi:10.1126/science.1135935

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., . . . Morris, R. G. M. (2011, August 12). Schema-dependent gene activation and memory encoding in neocortex. *Science, 333,* 891–895. doi:10.1126/science.1205274

Received January 2, 2013
Revision received May 6, 2013
Accepted June 3, 2013 ■