Neural-network models of learning and memory: leading questions and an emerging framework

Gail A. Carpenter

Real-time neural-network models provide a conceptual framework for formulating questions about the nature of cognition, an architectural framework for mapping cognitive functions to brain regions, a semantic framework for defining terms, and a computational framework for testing hypotheses. This article considers key questions about how a physical system might simultaneously support one-trial learning and lifetime memories, in the context of neural models that test possible solutions to the problems posed. Model properties point to partial answers, and model limitations lead to new questions. Placing individual system components in the context of a unified real-time network allows analysis to move from the level of neural processes, including learning laws and rules of synaptic transmission, to cognitive processes, including attention and consciousness.

When we go to the movies, we expect to relax. Here, nonetheless, even the adult moviegoer performs formidable feats of memorization. After leaving the theatre with friends, we are able to discuss details from all the scenes, and compare these with images from movies we saw only once years earlier. This common experience brings to bear an astonishing and nearly effortless blend of perception, attention, learning, and memory – the heart of cognitive science.

How can a finite system such as the brain quickly encode large quantities of new information without erasing essential memories?

One solution to this problem invokes 'exemplar learning'[1], which places each new memory in a separate compartment where it need not disturb its neighbors. The counterpoint to this view favors 'prototype learning'[2]. The dichotomy between exemplar and prototype learning is at least partially resolved in localist models, where disjoint subsets of nodes in a coding field represent distinct input clusters. Localist dynamics are most commonly modeled by 'winner-take-all' (WTA) competitive networks. In a WTA system, the net signal pattern converging on a field of coding nodes is quickly transformed by the field's internal dynamics so that only one node remains active in the steady state (Fig. 1a). A 'competitive network' with strong inhibitory connections produces WTA coding, as the node receiving the largest total signal suppresses all other activation[3,4]. Learning laws that restrict adaptation to paths projecting to or from the single active coding node protect memories stored in all other paths. When a node is first activated, its memory is of exemplar type. Subsequent activations may transform this to a prototype which represents a set of inputs but is identical to none.

Gail A. Carpenter
Dept of Cognitive and Neural Systems, 677 Beacon Street, Boston University, Boston, MA 02215, USA.
e-mail: gail@cns.bu.edu

Localist or distributed code representations?

Page has recently published a comprehensive review of the benefits and explanatory power of localist modeling in psychology[5]. He begins by pointing out: 'Over the last decade, fully distributed models have become dominant in connectionist psychology modelling, whereas the virtues of localist models have been underestimated.' (p. 443). A notable class of fully distributed models are the multilayer perceptrons (MLPs) (Refs 6,7), which include back propagation[8]. An MLP represents the code of the current input as activation patterns at one or more hidden layers. Nodes in these layers are modeled as traditional 'McCulloch–Pitts neurons'[9], with activity taken to be directly proportional to the total signal transmitted from a previous layer (Fig. 1b). Page notes that these models have become widely used: 'It is often stated as one of the advantages of networks using distributed representations that they permit generalization, which means that they are able to deal appropriately with patterns of information they have not previously experienced by extrapolating from those patterns they have experienced and learned' (p. 454) However, as Page also shows, 'contrary to an often repeated but seldom justified assumption' (p. 455) localist networks also generalize, albeit by different rules. Moreover, MLPs are prone to catastrophic forgetting, wherein memories are lost unpredictably (see Ref. 10 for a review of catastrophic interference in neural networks.) Finally, these networks typically use 'slow learning', which produces small weight adjustments on each learning trial. The one-trial learning experience of the moviegoer is more akin to 'fast learning', which allows weights to converge to asymptote on each trial.

Why, then, have fully distributed models such as the MLP become so popular? One reason is their ability to cope with certain types of noisy training data: even the two-layer perceptron can construct an optimal hyperplane to separate two overlapping

http://tics.trends.com 1364-6613/01/$ – see front matter PII: S1364-6613(00)01591-6

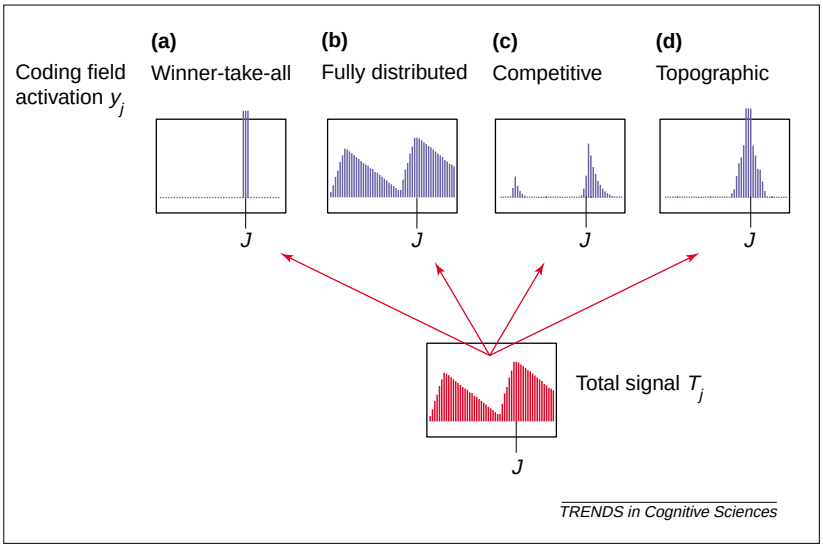

Fig. 1. Coding field activation patterns. Internal dynamics of a field of nodes determine a steady-state activation pattern, or code, $\mathbf{y}_j \equiv (y_1 \ldots y_j \ldots y_N)$. Vector **y** represents the network response to an incoming signal pattern $\mathbf{T}_j \equiv (T_1 \ldots T_j \ldots T_N)$, where component T_j is the sum of signals projecting to the j^{th} coding node. In the boxes (top row), with nodes of the coding field arranged from $j = 1$ at the left to $j = N$ at the right, the height of the graph indicates the activation (y_j) of each node. The total incoming signal is maximal at node $j = J$, and the activation y_j is also maximal in each code. (a) With winner-take-all coding, $y_J = 0$ at all nodes where $j \neq J$. (b) With fully distributed coding, the active pattern **y** is directly proportional to the signal pattern **T**. (c) Competition at the coding field enhances relative differences in the signal pattern and suppresses activation at nodes receiving a small signal. When internal competitive feedback is strong relative to the external signals T_j, the coding pattern is normalized. That is, total activation $\sum_{j=1}^{N} y_j$ is approximately equal to a constant which is independent of **T**. (d) A topographic map distributes activation to nodes adjacent to the maximally activated node *J*.

Gaussian distributions. A fast-learning WTA network such as fuzzy ARTMAP (Ref. 11) is designed to treat each wrong prediction as a potentially informative rare case, rather than as an outlier. Such a system might construct an adequate solution to an overlapping Gaussian problem, but these solutions tend to be inefficient, requiring more memory than the perceptron. On the other hand, fuzzy ARTMAP memories are stable, with weights converging with fast or slow learning.

Can the shared representations of a distributed code improve performance, efficiency or biological plausibility of a fast-learning system, whilst retaining desirable characteristics of localist codes?

The search for a design that integrates the best properties of models with fully distributed and WTA representations suggests consideration of these code types as two extremes of a continuum of competitive systems. If the internal dynamics of a coding field are parameterized in terms of the degree of competition between nodes, a fully distributed code (Fig. 1b) is found at the limit of zero competition. Increasing interaction strengths produces steady-state codes that represent progressively contrast-enhanced versions of the pattern of incoming signals (Fig. 1c), until the WTA limit (Fig. 1a) is reached[4]. A variant of the WTA case activates nodes adjacent to the winner (Fig. 1d), which produces a topographic relationship among nodes, as in the self-organizing map[12], a type of competitive learning[13–15]. The strength-of-competition parameter introduces an extra degree of freedom, and hence a new design question.

The fact that a WTA network may support fast learning and stable memories suggests consideration of coding patterns near this parametric limit. Such a network, where internal feedback signals are strong compared to external signals, has the property of 'normalization', which means that total steady-state activation across all nodes in the coding field is approximately constant. In a field of many nodes (N), with the dynamic range of each node scaled to the interval [0,1] and with competition so strong that total activation also bounded by 1, the average nodal activation is small ($1/N$). Thus, although such a normalized code is distributed in the sense that all nodes may be somewhat active at once, only a small number of nodes can be even moderately active simultaneously.

Normalization does not stabilize memory

Normalization of total coding-field activation points to a strategy for memory stabilization that uses the activity of each coding node to limit adaptation in paths projecting to and from that node. However, normalization alone does not accomplish this task. Consider, for example, a typical competitive learning system (Fig. 2). An input pattern **I** is transmitted to a coding field via converging weighted paths which transform **I** to a net signal pattern **T**. Strong intrafield competition transforms **T** to the normalized and contrast-enhanced code **y**. A type of gated steepest descent ('instar') learning adjusts weights according to the equation:

$$\frac{d}{dt}\mathbf{w}_j = y_j(\mathbf{I} - \mathbf{w}_j) \qquad (1)$$

where $\mathbf{w}_j$ is the vector of weights projecting from the input field to the j^{th} node of the coding field. According to Eqn 1, the weight vector $\mathbf{w}_j$ converges toward the input vector **I** wherever $y_j > 0$. As the rate of convergence is proportional to y_j, a weight vector $\mathbf{w}_j$ projecting to a highly active node will track the input **I** more closely than will a less active node – *provided that learning is slow*. With fast learning, *all* vectors $\mathbf{w}_j$ will converge to the *same* input **I** wherever y_j is even slightly positive. This is an extreme form of catastrophic forgetting in which each input can wipe out all prior memories. MLPs, which typically learn via another type of gated steepest descent ('back-coupled error correction'[6]), require slow learning for a similar reasons.

Rules of synaptic transmission

Coding field normalization does not immediately solve the catastrophic forgetting problem. Analysis of the competitive learning example does, however, point the way toward a reconsideration of the fundamental components that govern network dynamics at the synaptic level and the implicit

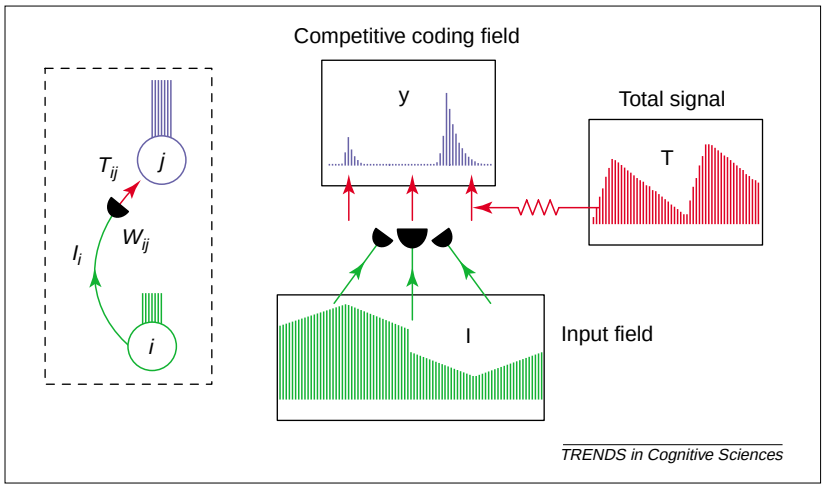

Fig. 2. Competitive learning example. A typical competitive learning network maps an input pattern $\mathbf{I} \equiv (I_1 \ldots I_i \ldots I_M)$ to a compressed recognition code $\mathbf{y} \equiv (y_1 \ldots y_j \ldots y_N)$ via an adaptive filter $w_1 \ldots w_j \ldots w_N$, where $\mathbf{w}_j \equiv (w_{1j} \cdots w_{ij} \cdots w_{Mj})$ is the pattern of weights in paths projecting from an input field to the j^{th} node of a coding fiel d. The j^{th} component of the net signal pattern T is a sum $T_j = \sum_{i=1}^{M} T_{ij}$, where T_{ij} is the signal transmitted to the j^{th} coding node from the i^{th} input node, via a weighted path. Competitive feedback within the coding field transforms the signal T to the code y. Instar learning sends w_j toward I, with the rate of convergence depending on the activation (y_j) of the j^{th} target node.

assumptions that define learning laws, the signal transmitted across a synapse, and even the basic unit of memory.

A 'synaptic transmission rule' specifies the model function that transforms a presynaptic input, or spiking frequency (I_i), to a postsynaptic signal (T_{ij}) transmitted to the j^{th} target node (Fig. 2). Rosenblatt's original perceptron axioms[6] postulated a general class of transmission rules, with:

$$T_{ij} = f(I_i, w_{ij}) \qquad (2)$$

Since 1960, the vast majority of neural-network models have taken the unit of long-term memory (LTM) to be a multiplicative weight, or adaptive gain (w_{ij}), corresponding to the particular synaptic transmission rule:

$$T_{ij} = I_i \cdot w_{ij} \qquad (3)$$

This hypothesis is also implicit in the experimental investigation of long-term potentiation (LTP): 'Changes in the amplitude of synaptic responses evoked by single-shock extracellular electrical stimulation of presynaptic fibres are usually considered to reflect a change in the gain of synaptic signals, and are the most frequently used measure for evaluating synaptic plasticity.' (Ref. 16, p. 807). That is, LTM change is assumed to be characterized by testing only with low-frequency ('single-shock') presynaptic inputs (small I_i), with the response to high-frequency inputs inferred via Eqn 3.

Although the multiplicative-weight hypothesis has proved computationally useful for decades, it is neither axiomatic nor experimentally required, which opens the question: *what rules of synaptic transmission support global computational goals in model systems and in their physiological counterparts?*

Recently, Markram and Tsodyks[16] have critically challenged the universality of the adaptive gain hypothesis by demonstrating redistribution of synaptic efficacy (RSE) in pairing experiments in neocortical layer-5 pyramidal cells. In this preparation, the elevated synaptic efficacy characteristic of the single-pulse LTP test disappears for test pulses of higher frequencies. In fact, the post-pairing response to test pulses above 20 Hz falls below the pre-pairing level. These important experiments suggest the possibility that LTM adaptation should be modeled as redistribution of synaptic efficacy rather than as a gain change, which would have implications for global pattern learning in neural networks.

Cortical feedback loops

Let us now return to a consideration of network-level design. Pollen[17], in a wide-ranging review of the neural correlates of visual perception, resolves various past and current views of cortical function by placing them in a framework he calls 'adaptive resonance theories.' This unifying perspective postulates resonant feedback loops as the substrate of phenomenal experience. Adaptive resonance offers a core module for the representation of hypothesized processes underlying learning, attention, search, recognition, and prediction[18]. At the model's field of coding neurons, the continuous stream of information pauses for a moment, holding a fixed activation pattern long enough for memories to change. Intrafield competitive loops fixing the moment are broken only by active reset, which flexibly segments the flow of experience according to the demands of perception and environmental feedback.

Pollen further suggests that 'it may be the consensus of neuronal activity across ascending and descending pathways linking multiple cortical areas that in anatomical sequence subserves phenomenal visual experience and object recognition and that may underlie the normal unity of conscious experience.' (Ref. 17, pp. 15–16). Despite its appeal, as well as manifold experimental demonstrations of feedback in the visual system[19], achieving an interfield feedback consensus presents formidable computational challenges, including the question: what designs for feedback loop dynamics and the matching of bottom-up and top-down signals guarantee convergence or other interpretable network states?

The interfield feedback problem is illustrated by the following example (Fig. 3). Suppose that feedforward signals activate a code which returns top-down feedback, thereby transforming the original input pattern in some way. This design generates a cascade of questions. Will new feedforward signals then produce a new code, which will send new feedback, etc.? Will the interfield activation cycle converge? If so, how would learning affect the code representing this input? Would the input make a correct prediction the next time it is presented? The first ART model[20] contained the module shown in Fig. 3, and learning laws and

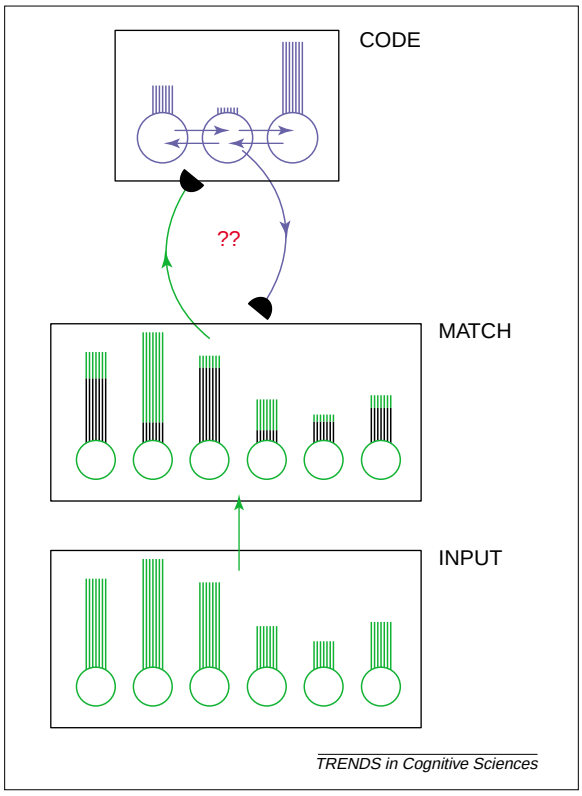

Fig. 3. Interfield and intrafield feedback. In the competitive learning example of Fig. 2, bottom-up signals from an input pattern (green bars) activate a coding pattern (purple bars). Intrafield feedback loops (purple arrows) implement competitive dynamics within the coding field. Questions arise concerning how to design a learning system that incorporates an interfield feedback loop, where the code would project top-down signals that transform the active pattern at a matching field (black bars), which would then send new signals to the coding field, and so on. Unless the code is WTA, these questions remain unanswered.

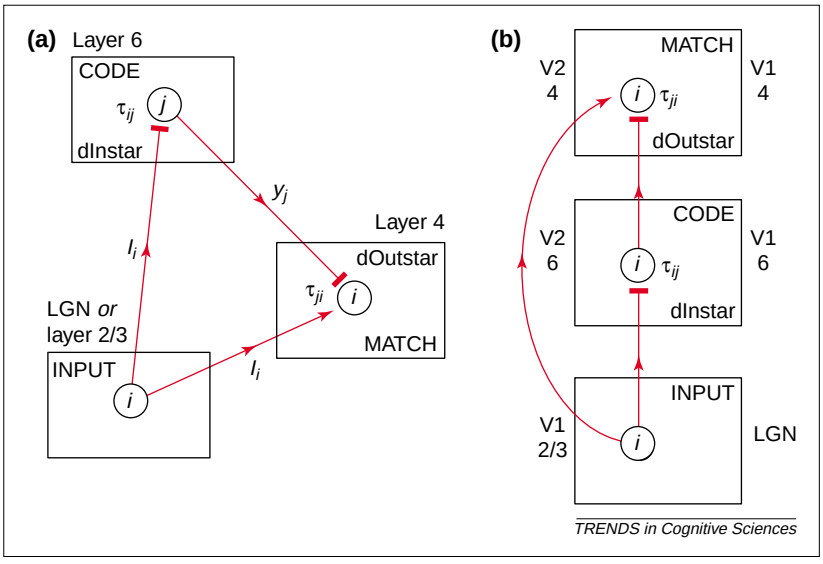

Fig. 4. dART network configuration and cortical layers. (a) In the distributed ART (dART) network, an input pattern I projects directly to a coding field, which transforms the net signal pattern to a normalized code y, which may be distributed across arbitrarily many nodes. Activity at a matching field registers the degree of similarity between the bottom-up input I and a top-down signal pattern, or expectation, transmitted from the coding field. Long-term memories are stored in paths projecting to the coding field as thresholds τ_{ij}, which adapt according to a distributed instar (dInstar) learning law; and in paths projecting from the coding field, as thresholds τ_{ji}, which adapt according to a distributed outstar (dOutstar) learning law. (b) The dART network configuration is isomorphic to modular components of a laminar model[26] of visual cortex. Comparing dART with the first level of the laminar model hierarchy, the input field may be identified with LGN, the coding field with V1 cortical layer 6, and the matching field with the V1 layer 4. This anatomical equivalence indicates how learning laws and other dynamic components of the dART network might be incorporated into a cortical model, and suggests new functional roles for the various layers. Since the laminar model features isomorphic structures in a cortical hierarchy, dART functions may be tested at each corresponding level. Note that the reconfiguration of the dART architecture blurs the distinction between 'top-down' expectation and 'bottom-up' input at the matching field in diagram (a), since both sets of signals are drawn 'bottom-up' in diagram (b). Note, too, that the laminar cortex model includes other top-down attentional signals (e.g. from V2 layer 6 to V1 layer 6) as part of a 'folded-feedback' circuit.

network dynamics were chosen explicitly to guarantee convergence and orderly learning. Moreover, the search process of the supervised ARTMAP network[21] was designed to ensure that subsequent learning corrects predictive errors. However, these solutions apply only if coding is WTA.

A quasi-localist fast-learning network

The series of questions discussed in the previous sections range from large-scale problems of pattern learning to small-scale problems of synaptic computation. Starting with a WTA ART module, step-by-step consideration of these questions has led to a new network configuration, new rules of synaptic transmission, new learning laws, and a new unit of memory. The resulting distributed ART ('dART') model is one working example of a neural system that produces stable memories with fast learning and with code representations that may be distributed across arbitrarily many nodes (Refs 22–24).

New learning laws and rules of synaptic transmission in a reconfigured network architecture (Fig. 4a) sidestep the interfield feedback problems caused by distributed coding in a traditional ART network. Despite their different architectures, however, dART with fast learning and WTA coding is algorithmically equivalent to fuzzy ART. The critical design element that allows dART to solve the catastrophic forgetting problem is the 'dynamic weight'[25]. This quantity equals the rectified difference between coding node activation and an adaptive threshold, thus combining short-term and long-term memory in the network's fundamental computational unit. Thresholds in paths projecting from an input field to a coding field obey a distributed instar ('dInstar') learning law, which reduces to an instar law (Eqn 1) when coding is WTA. Learning in these paths resembles Markram and Tsodyks redistribution of synaptic efficacy, rather than adaptive gain change. Thresholds in paths projecting from the coding field to a matching field obey a different learning law ('dOutstar'), encoding the network's learned expectations with respect to the coding field activation pattern. As in other ART systems, dART compares the top-down expectation with the bottom-up input at the matching field, and quickly searches for a new code if the match fails to meet a criterion determined by a parameter called 'vigilance'.

Where in the brain might model components be found?

A comparison between the dART network and a recent laminar computing model of bottom-up, top-down, and horizontal interactions among layers in the visual cortex[26,27] has suggested some preliminary identifications between model components and cortical layers (Fig. 4b). In turn, this identification suggests how the laminar model, which has been applied primarily to earlier levels of the visual cortex, might be extended to include fast, stable, distributed learning in later cortical areas that participate in recognition,

Table 1. Dynamic balance of memory design elements

System dynamics	
Bottom-up signals	Top-down signals
Feedforward inflow	Feedback outflow
Perception	Expectation
Localist activation	Distributed activation
Rules and symbols	Real-time processing
Specific signals	Nonspecific signals
Signal	Noise
Environmental input	Critical features
Prototypes	Exemplars
Generalization	Encoding rare cases
Present features	Absent features
On-cells	Off-cells
Search	
Attention	Orientation
Familiarity	Novelty
Match	Reset
Learning	
Stability	Plasticity
Invariance	Change
Limited capacity of STM	Unlimited capacity of LTM
Dynamic weight	Fixed weight
Online, incremental learning	Offline, batch learning
Unsupervised learning	Supervised learning
Fast learning	Slow adaptation
Cognition	
Coding	Action
One-to-many mapping	Many-to-one mapping
Consistent worldview	Inconsistent perceptions
Lifetime memory	Amnesia

learning and prediction, including inferotemporal cortex. In particular, this identification would predict RSE (which was measured by Markram and Tsodyks in layer 5) in certain paths projecting to layer 6, but different synaptic computations at layer 4.

The dART synapses use the activation level at each coding node to stabilize memory by imposing limits on threshold changes permitted on any given learning trial, with fast or slow learning. The network hereby relies strongly on a coding field normalization hypothesis. Although any number of nodes may combine their activations to make a net prediction, in practice, learned change is often restricted to one active node. These networks are thus more 'quasi-localist' than fully distributed in character. Normalization helps stabilize memory in a system whose permitted codes are infinitely more varied than the WTA special case, but which represents just a preliminary solution to some of the design problems outlined above.

Conclusion: modeling as a dynamic balancing act

The functional capabilities and limitations of the dART network immediately suggest additional questions. For one, the current network tends to be weighted too heavily in favor of absolute implementation of the stability requirement, which can cause the system to resist learning new information late in training. Hence the ongoing design question: *how can a fast-learning network maintain stable codes without locking in its early memories too soon?*

A related question concerns the design of a distributed match-reset-search process. In particular: *when a network makes a predictive error, how should a distributed code be reset so that the system can learn not to repeat the error next time?* More generally, the model development process illustrated here exemplifies some of the trade-offs in a dynamic balance of memory designs (see Table 1).

Acknowledgements
The author thanks Stephen Grossberg for many years of discussion and collaboration. This research was supported by grants from the Office of Naval Research and the Defense Advanced Research Projects Agency (ONR N00014-95-1-0409 and ONR N00014-1-95-0657) and the National Institutes of Health (NIH 20-316-4304-5).

References

1 Estes, W.K. (1986) Memory storage and retrieval processes in category learning. *J. Exp. Psychol. Gen.* 115, 155–174
2 Posner, M.I. and Keele, S.W. (1970) Retention of abstract ideas. *J. Exp. Psychol.* 83, 304–308
3 Sperling, G. (1970) Binocular vision: a physical and a neural theory. *Am. J. Psychol.* 83, 461–534
4 Grossberg, S. (1973) Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Stud. Appl. Math.* 52, 217–257
5 Page, M. (2000) Connectionist modelling in psychology: a localist manifesto. *Behav. Brain Sci.* 23, 443–512
6 Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 486–408
7 Rosenblatt, F. (1962) *Principles of Neurodynamics*, Spartan Books
8 Rumelhart, D.E. *et al.* (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* (Vol. I) (Rumelhart, D.E. and McClelland, J.L., eds), pp. 318–362, MIT Press
9 McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 9, 127–147
10 French, R.M. (1999) Catastrophic forgetting in connectionist networks. *Trends Cognit. Sci.* 3, 128–135
11 Carpenter, G.A. *et al.* (1992) Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. Neural Netw.* 3, 698–713
12 Kohonen, T. (1984) *Self-organization and Associative Memory*, Springer-Verlag
13 von der Malsburg, C. (1973) Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14, 85–100
14 Grossberg, S. (1972) Neural expectation: cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik* 10, 49–57
15 Grossberg, S. (1976) Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 21, 145–159
16 Markram, H. and Tsodyks, M. (1996) Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* 382, 807–810
17 Pollen, D.A. (1999) On the neural correlates of visual perception. *Cereb. Cortex* 9, 4–19
18 Grossberg S. (1980) How does a brain build a cognitive code? *Psychol. Rev.* 87, 1–51
19 Lamme, V.A.F. and Roelfsema, P.R. (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579
20 Carpenter G.A. and Grossberg S. (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vis. Graphics Image Process.* 37, 54–115
21 Carpenter G.A. *et al.* (1991) ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Netw.* 4, 565–588
22 Carpenter, G.A. (1996) Distributed activation, search, and learning by ART and ARTMAP neural networks. In *Proc. Int. Conf. Neural Netw.* (*ICNN'96*), pp. 244–249, IEEE Press
23 Carpenter, G.A. (1997) Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Netw.* 10, 1473–1494
24 Carpenter, G.A. *et al.* (1998) Distributed ARTMAP: a neural network for fast distributed supervised learning. *Neural Netw.* 11, 793–813
25 Carpenter, G.A. (1994) A distributed outstar network for spatial pattern learning. *Neural Netw.* 7, 159–168
26 Grossberg, S. (1999) How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spat. Vis.* 12, 163–185
27 Grossberg, S. (2000) The complementary brain: unifying brain dynamics and modularity. *Trends Cognit. Sci.* 4, 233–246