I enthusiastically concur with the bulk of what Page has to say, but I would like to elaborate on the localist approach outlined in the target article based on my own involvement with the approach. In my opinion, the localist position and the localist computational model presented in the target article are overly restrictive. The author has focused primarily on the representation of entities that can be expressed as soft (weighted) conjunctions of features. The localist model described in section 4 deals almost exclusively with the acquisition and retrieval of higher-level concepts (nodes) that are soft conjunctions of lower-level features. Even the more advanced example discussed in section 4.5 focuses on learning associations between such entities. This limited representational focus is also reflected in the examples of entities enumerated by the author, namely, "words, names, persons, etc." (sect. 2.5). What Page leaves out are more complex conceptual items such as events, situations, actions, and plans, which form the grist of human cognition.

Events, situations, actions, and plans involve *relational* and *procedural* knowledge, and hence, cannot be encoded as mere soft conjunctions of features; their encoding requires more *structured* representations. Working toward a representation of such complex and structured items leads to a more articulated view of the localist approach than the one presented in the target article. I will briefly comment on this view. For more details, the reader is referred to specific models that instantiate this view (see Ajjanagadde & Shastri 1991; Bailey 1997; Shastri 1991; 1997; 1999a; 1999b; 1999c; Shastri & Ajjanagadde 1993; Shastri et al., in press).

In the enriched representational context of events, situations, actions, and plans the operative representational unit is often a *circuit* of nodes rather than a node. Moreover, only some of the nodes in such a circuit correspond to cognitively meaningful entities (as the latter are characterized in sect. 2.5). Most of the other nodes in the circuit serve a *processing* function or perform an ancillary representational role. For example, such nodes glue together simpler items in systematic ways to form composite relational items, they provide a handle for systematically accessing specific components of a composite item, they provide a handle for systematically accessing specific components of a composite item, and they allow actions and plans to be expressed as partially ordered structures of subactions and subplans. Thus the encoding of an event (E1) "John gave a book to Mary" in long-term memory would involve not only nodes corresponding to cognitively meaningful entities such as John, Mary, book, giver, recipient, and object, but also *functionally* meaningful nodes such as: a node for asserting belief in E1, a node for querying E1, *binder* nodes for encoding role-entity bindings in E1 (for example, a node for encoding the binding giver = John), *binding-extractor* nodes for selectively retrieving role-fillers in E1 (for example, a node for activating "John" in response to the activation of the role "giver" in the context of E1), and nodes for *linking* the encoding of E1 to a generic perceptual-motor schema for the *give* action. Furthermore, the localist encoding of the *give* schema would involve specific nodes and circuits for encoding a partially ordered sequence of perceptual-motor subactions comprising the *give* action.

In the expanded localist framework, individual nodes continue to have well-defined localist interpretations. However, these interpretations are best couched in terms of a node's *functional* significance rather than its semantic significance (cf. sect. 2.5).

The learning framework presented by the author has a strong overlap with work on recruitment learning (Diederich 1989; Feldman 1982; Shastri 1988; 1999b; 1999c; Valiant 1994; Wickelgren 1979). The architecture described in Figure 9 of the target article is in many ways analogous to that sketched out in (Shastri 1988, pp. 182–92). This overlap merits further exploration. In the recruitment learning framework, learning occurs within a network of quasi-randomly connected nodes. Recruited nodes are those nodes that have acquired a distinct meaning or functionality by virtue of their *strong* interconnections to other recruited nodes and/or other sensorimotor nodes. Nodes that are not yet recruited are nodes "in waiting" or "free" nodes. Free nodes are connected via weak links to a large number of free, recruited, and/or sensorimotor nodes. These free nodes form a primordial network from which suitably connected nodes may be recruited for representing new items. For example, a novel concept y which is a conjunct of existing concepts x_1 and x_2 can be encoded in long-term memory by "recruiting" free nodes that receive links from both x_1 and x_2 nodes. Here recruiting a node simply means strengthening the weights of links incident on the node from x_1 and x_2 nodes. In general, several nodes are recruited for each item.

The recruitment process can transform quasi-random networks into structures consisting of nodes tuned to specific functionalities. Typically, a node receives a large number of links, and hence, can potentially participate in a large number of functional circuits. If, however, the weights of selected links increase, and optionally, the weights of other links decrease, the node can become more selective and participate in a limited number of functional circuits.

In Shastri (1999b; 1999c) it is shown that recruitment learning can be firmly grounded in the biological phenomena of *long-term potentiation* (LTP) and *long-term depression* (LTD) that involve rapid, long-lasting, and specific changes in synaptic strength (Bliss & Collingridge 1996; Linden 1994). Moreover, as explained in Shastri (1999c) the specification of a learning algorithm amounts to choosing a suitable network architecture and a set of appropriate parameter values for the induction of LTP and LTD.

The recruitment learning framework also offers an alternate explanation for the age-of-acquisition effect discussed in section 4.4. It suggests that (on an average) more cells are recruited for items acquired earlier in a learning cycle than for items acquired later in the cycle. Thus items acquired earlier in the learning cycle have greater neuronal mass and it is this greater mass that gives these items their competitive edge.

To conclude, Page must be commended for systematically and comprehensively presenting a strong case for the localist models. The localist position and computational model presented in the target article, however, can be enriched by considering the representation of complex items involving relational and procedural knowledge. Work on representing such items leads to a more articulated view of the localist approach than that presented in the target article.

ACKNOWLEDGMENT

This work was supported in part by NSF grants SBR-9720398 and ECS-9970890.

Prototypes and portability in artificial neural network models

Thomas R. Shultz

Department of Psychology, McGill University, Montreal, Quebec, Canada H3A 1B1. **shultz@psych.mcgill.ca**
www.psych.mcgill.ca/labs/lnsc/html/Lab-Home.html

Abstract: The Page target article is interesting because of apparent coverage of many psychological phenomena with simple, unified neural techniques. However, prototype phenomena cannot be covered because the strongest response would be to the first-learned stimulus in each category rather than to a prototype stimulus or most frequently presented stimuli. Alternative methods using distributed coding can also achieve portability of network knowledge.

The Page target article is surprisingly interesting. I use the term "surprisingly" because, with all of the deep controversies in cognitive science, it is difficult to care much about whether network representations are local or distributed. In any given simulation, choice of representation is of key importance, but it is rarely regarded as a life-and-death ideological issue whether these codes are local or distributed. Many modelers adopt an eclectic approach that enables them to use representations that (a) work in terms of covering psychological data and (b) can be justified by psychological evidence.

What is significant about Page's article is the fact that such a simple, unified, nonmainstream neural model can apparently capture so many phenomena, from unsupervised learning to age-of-acquisition effects, in such a natural fashion. That the coding is local is somewhat incidental to that source of interest, even though local coding happens to be critical to the functioning of Page's particular networks.

It might be that Page has dichotomized and polarized the field too much. For example, a reader could easily get the impression from section 4.3.2 that conventional attractor networks always or typically employ distributed codes. But there are many instances of local encoding in successful attractor network models that are quite different from the networks that Page proposes. Such models cover, for example, analogical retrieval and mapping (Holyoak & Thagard 1989; Thagard et al. 1990), explanation (Read & Marcus-Newhall 1993; Thagard 1989), decision making (Thagard & Millgram 1995), attitude change (Spellman et al. 1993), person impression (Kunda & Thagard 1996; Read & Miller 1998; Smith & DeCoster 1998), and cognitive dissonance (Shultz & Lepper 1996).

Page argues that local coding is to be preferred for psychological modeling over distributed coding. A less polarizing conclusion would be that both local and distributed encoding techniques are legitimate within a variety of different neural network techniques. Page himself notes that many localist models use some distributed coding. Because eclectic use of local and distributed codes is so common, it is somewhat difficult to accept Page's strongly localist argument. In the end, Page is willing to call a coding system local even if only some of its codes are local. With so many modelers willing to use both local and distributed codes, a strict dichotomy seems unwarranted.

Because Page's models can apparently cover such a wide range of effects, it would be useful to examine this coverage in more detail than was possible in his article. For example, the basic learning module described in section 4.1 would seem to have considerable difficulty simulating common prototype effects. This difficulty stems from the fact the strongest second-layer (output) responses would be the first stimulus learned in each category, rather than to a prototype stimulus or the most frequent stimuli. This is because each new stimulus is responded to most by the second-layer unit that first learned to respond to the most similar previously learned stimulus. Only stimuli that are sufficiently different from previously learned stimuli will provoke new learning by an uncommitted second-layer unit. In contrast, psychological evidence has found the largest recognition responses to occur to prototypic or especially frequent stimuli, not to first-learned stimuli (Hayes-Roth & Hayes-Roth 1977). These psychological prototype findings are more readily accounted for by a variety of neural network models that are different from Page's models. For example, auto-associator networks learning with a Hebbian or delta rule (McClelland & Rumelhart 1986) or encoder networks learning with the back-propoagation rule can cover prototype phenomena. Interestingly, it does not matter whether these successful network models use local or distributed codes. It might prove interesting to examine in more detail the psychological fitness of the rest of Page's models, all of which build on this basic learning module.

One of the major apparent advantages of Page's localist models is the relative ease with which local representations can be manipulated (sect. 7.5), as compared to representations that are distributed over many units. It is possible that this feature could be exploited to achieve portability of knowledge. People seem capable of porting their knowledge to novel problems in creative ways, and this portability is sometimes regarded as a significant challenge for artificial neural network models (Karmiloff-Smith 1992). Local representations like those advocated by Page might be good candidates for portability. Building or learning connection weights from a single unit, perhaps representing a complex idea, seems much easier than establishing connection weights from many such representation units.

This is not to admit, however, that local coding is required for knowledge portability in neural networks. Alternative techniques for achieving knowledge portability with distributed codes might well be possible too. One example is the work we are currently doing on a system called Knowledge-based Cascade-correlation (KBCC) Shultz 1998). Ordinary Cascade-correlation (CC) is a generative feed-forward network algorithm that grows as it learns, by recruiting new hidden units into a network as needed to reduce error (Fahlman & Lebiere 1990). The hidden units recruited by CC are virginal, know-nothing units until they are trained to track current network error during the recruitment process. However, KBCC has the ability to store and possibly recruit old CC networks that do already know something. In KBCC, old networks compete with new single units to be recruited. This makes old knowledge sufficiently portable to solve new problems, if the old knowledge is helpful in tracking and ultimately reducing network error. Moreover, because the stored networks are retained in their original form, KBCC is much more resistant to the catastrophic interference caused by new learning in most static feed-forward networks. It is noteworthy once again that all of this can be accomplished regardless of whether the coding is local or distributed in KBCC systems. Actually, even ordinary CC networks are quite resistant to catastrophic interference because of the policy of freezing input weights to hidden units after recruitment (Tetewsky et al. 1994). This ensures that each hidden unit never forgets its original purpose, even though it may eventually play a new role in learning to solve current problems.

Hidden Markov model interpretations of neural networks

Ingmar Visser

Developmental Psychology Institute of the University of Amsterdam, 1018 WB Amsterdam, The Netherlands. **ingmar@dds.nl**
develop.psy.uva.nl/users/ingmar/op_visser@macmail.psy.uva.nl

Abstract: Page's manifesto makes a case for localist representations in neural networks, one of the advantages being ease of interpretation. However, even localist networks can be hard to interpret, especially when at some hidden layer of the network distributed representations are employed, as is often the case. Hidden Markov models can be used to provide useful interpretable representations.

In his manifesto for the use of localist neural network models, Page mentions many advantages of such a scheme. One advantage is the ease of interpretation of the workings of such a network in psychologically relevant terms (sect. 7.6).

As Page justly remarks, a localist model does not imply that distributed representations are not used in *any* part of the model; rather a localist model is characterized by employing localist representations at some (crucial) points such as the output level of the network. More specifically he states that "any entity that is locally represented at layer n of the hierarchy is sure to be represented in a distributed fashion at layer $n - 1$" (sect. 2.6). Why should the problem of interpretation not apply to these distributed representations at lower levels as well? I think it does, and it's best to illustrate this with an example.

Following the work of Elman (1990), Cleeremans and McClelland (1991) used a simple recurrent network SRN to model implicit learning behavior using localist representations at both input and output layers, but a distributed representation at the hidden layer of the network. As they show in their paper the SRN model captures the main features of subjects' performance by "growing increasingly sensitive to the temporal context [of the current stimulus]." This sensitivity to the temporal context of stimuli is somehow captured by representations formed at the hidden layer of the network. In exactly what sense differences in temporal context affect activity at the hidden layer is unclear: What does a certain pattern of activity of the hidden layer units mean?