BEHAVIORAL AND BRAIN SCIENCES (2000) **23,** 443–512
Printed in the United States of America

Connectionist modelling in psychology: A localist manifesto

Mike Page
Medical Research Council Cognition and Brain Sciences Unit, Cambridge, CB2 2EF, United Kingdom
mike.page@mrc-cbu.cam.ac.uk www.mrc-cbu.cam.ac.uk/

Abstract: Over the last decade, fully distributed models have become dominant in connectionist psychological modelling, whereas the virtues of localist models have been underestimated. This target article illustrates some of the benefits of localist modelling. Localist models are characterized by the presence of localist representations rather than the absence of distributed representations. A generalized localist model is proposed that exhibits many of the properties of fully distributed models. It can be applied to a number of problems that are difficult for fully distributed models, and its applicability can be extended through comparisons with a number of classic mathematical models of behaviour. There are reasons why localist models have been underused, though these often misconstrue the localist position. In particular, many conclusions about connectionist representation, based on neuroscientific observation, can be called into question. There are still some problems inherent in the application of fully distributed systems and some inadequacies in proposed solutions to these problems. In the domain of psychological modelling, localist modelling is to be preferred.

Keywords: choice; competition; connectionist modelling; consolidation; distributed; localist; neural networks; reaction-time

1. Introduction

The aim of this target article is to demonstrate the power, flexibility, and plausibility of connectionist models in psychology which use localist representations. I will take care to define the terms "localist" and "distributed" in the context of connectionist models and to identify the essential points of contention between advocates of each type of model. Localist models will be related to some classic mathematical models in psychology and some of the criticisms of localism will be addressed. This approach will be contrasted with a currently popular one in which localist representations play no part. The conclusion will be that the localist approach is preferable whether one considers connectionist models as psychological-level models or as models of the underlying brain processes.

At the time of writing, it is thirteen years since the publication of *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* (Rumelhart, McClelland, & the PDP Research Group 1986). That two-volume set has had an enormous influence on the field of psychological modelling (among others) and justifiably so, having helped to revive widespread interest in the connectionist enterprise after the seminal criticisms of Minsky and Papert (1969). In fact, despite Minsky and Papert's critique, a number of researchers (e.g., S. Amari, K. Fukushima, S. Grossberg, T. Kohonen, C. von der Malsburg) had continued to develop connectionist models throughout the 1970s, often in directions rather different from that in which the 1980's "revival" later found itself heading. More specifically, much of the earlier work had investigated networks in which *localist representations* played a prominent role, whereas, by contrast, the style of modelling that received most attention as a result of the PDP research group's work was one that had at its centre the concept of *distributed representation*. It is more than coincidental that the word "distributed" found itself centrally located in both the name of the research group and the title of its major publication but it is important to note that in these contexts the words "parallel" and "distributed" both refer to processing rather than to representation. Although it is unlikely that anyone would deny that processing in the brain is carried out by many different processors in parallel (i.e., at the same time) and that such processing is necessarily distributed (i.e., in space), the logic that leads from a consequent commitment to the idea of *distributed processing,* to an equally strong commitment to the related, but distinct, notion of *distributed representation,* is more debatable. In this target article I hope to show that the *thoroughgoing* use of distributed representations, and the learning algorithms associated with them, is very far from being mandated by a general commitment to parallel distributed processing.

As indicated above, I will advocate a modelling approach that supplements the use of distributed representations (the existence of which, in some form, nobody could deny) with the additional use of localist representations. The latter have acquired a bad reputation in some quarters. This cannot be directly attributed to the PDP books themselves, in which several of the models were localist in flavour (e.g.,

MIKE PAGE studied Engineering Science at Oxford University before obtaining a doctorate in connectionist modelling of music perception at the University of Wales, Cardiff. He is currently a nontenured scientist at the Medical Research Council Cognition and Brain Sciences Unit in Cambridge, United Kingdom, where he works on connectionist modelling of memory, in particular, memory for serial order.

 0140-525X/00 $12.50

interactive activation and competition models, competitive learning models). Nonetheless, the terms "PDP" and "distributed" on the one hand, and "localist" on the other, have come to be seen as dichotomous. I will show this apparent dichotomy to be false and will identify those issues over which there is genuine disagreement.

A word of caution: "Neural networks" have been applied in a wide variety of other areas in which their plausibility as models of cognitive function is of no consequence. In criticizing what I see to be the overuse (or default use) of fully distributed networks, I will accordingly restrict discussion to their application in the field of connectionist modelling of cognitive or psychological function. Even within this more restricted domain there has been a large amount written about the issues addressed here. Moreover, it is my impression that the sorts of things to be said in defence of the localist position will have occurred independently to many of those engaged in such a defence. I apologize in advance, therefore, for necessarily omitting any relevant references that have so far escaped my attention. No doubt the *BBS* commentary will set the record straight.

The next section will define some of the terms to be used throughout this target article. As will be seen, certain subtleties in such definitions becloud the apparent clarity of the localist/distributed divide.

2. Defining some terms

2.1. Basic terms

Before defining localist and distributed representations, we establish some more basic vocabulary. In what follows, the word *nodes* will refer to the simple units out of which connectionist networks have traditionally been constructed. A node might be thought of as consisting of a single neuron or a distinct population of neurons (e.g., a cortical minicolumn). A node will be referred to as having a level of activation, where a loose analogy is drawn between this activation and the firing rate (mean or maximum firing rate) of a neuron (population). The activation of a node might lead to an output signal's being projected from it. The projection of this signal will be deemed to be along one or more weighted connections, where the concept of weight in some way represents the variable ability of output from one node to affect processing at a connected node. The relationship between the weighted input to a given node (i.e., those signals projected to it from other nodes), its activation, and the output which it in turn projects, will be summarized using a number of simple, and probably familiar, functions. All of these definitions are, I hope, an uncontroversial statement of the basic aspects of the majority of connectionist models.

2.2. Localist and distributed representations

The following definitions, drawn from the recent literature, largely capture the difference between localist and distributed representations [see also Smolensky: "On the Proper Treatment of Connectionism" *BBS* 11(1) 1988; Hanson & Burr: "What Connectionist Models Learn" *BBS* 13(3) 1990; Van Gelder: "The Dynamical Hypothesis in Cognitive Science" *BBS* 21(5) 1998; O'Brien & Opie: "A Connectionist Theory of Phenomenal Experience" *BBS* 22(1) 1999]. First, distributed representations:

> Many neurons participate in the representation of each memory and different representations share neurons. (Amit 1995, p. 621)
>
> The model makes no commitment to any particular form of representation, beyond supposing that the representations are distributed; that is, each face, semantic representation, or name is represented by multiple units, and each unit represents multiple faces, semantic units or names. (Farah et al. 1993, p. 577)

The latter definition refers explicitly to a particular model of face naming, but the intended nature of distributed representations in general is clear. To illustrate the point, suppose we wished to represent the four entities "John," "Paul," "George," and "Ringo." Figure 1a shows distributed representations for these entities. Each representation involves a pattern of activation across four nodes and, importantly, there is overlap between the representations. For instance, the first node is active in the patterns representing both John and Ringo, the second node is active in the patterns representing both John and Paul, and so on. A corollary of this is that the identity of the entity that is currently represented cannot be unambiguously determined by inspecting the state of any single node.

Now consider the skeleton of a definition of a localist representation, as contrasted with a distributed coding:

> With a local representation, activity in individual units can be interpreted directly . . . with distributed coding individual units cannot be interpreted without knowing the state of other units in the network. (Thorpe 1995, p. 550)

For an example of a localist representation of our four entities, see Figure 1b. In such a representation, only one node is active for any given entity. As a result, activity at a given unit can unambiguously identify the currently represented entity.

When nodes are binary (i.e., having either activity 1 or 0), these definitions are reasonably clear. But how are they affected if activity can take, for example, any value between these limits? The basic distinction remains: in the localist model, it will still be possible to interpret the state of a given node independent of the states of other nodes. A natural way to "interpret" the state of a node embedded in a localist model would be to propose, as did Barlow (1972), a monotonic mapping between activity and confidence in the presence of the node's referent:

> The frequency of neural impulses codes subjective certainty: a high impulse frequency in a given neuron corresponds to a high degree of confidence that the cause of the percept is present in the external world. (Barlow 1972, p. 381)

It may be that the significance of activating a given node is assessed in relation to a threshold value, such that only superthreshold activations are capable of indicating nonzero

John	1	1	0	0
Paul	0	1	1	0
George	0	0	1	1
Ringo	1	0	0	1

(a)

John	1	0	0	0
Paul	0	1	0	0
George	0	0	1	0
Ringo	0	0	0	1

(b)

Figure 1. Four names represented (a) in a distributed fashion and (b) in a localist fashion.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

confidence. Put another way, the function relating activation to "degree of confidence" would not necessarily be linear, or even continuously differentiable, in spite of being monotonic nondecreasing.

Having offered both Thorpe's and Barlow's descriptions of localist representation, I must point out that interpreting a node's activation as "degree of confidence" is potentially inconsistent with the desire to interpret a given node's activation "directly," that is, independent of the activation of other nodes. For example, suppose, in a continuous-activation version of Figure 1b, that two nodes have near maximal activity. In some circumstances we will be happy to regard this state as evidence that both the relevant referents are present in the world: in this case the interpretation of the node activations will conform to the independence assumption. In other cases, we might regard such a state as indicating some ambiguity as to whether one referent or the other is present. In these cases it is not strictly true to say that the degree of confidence in a particular referent can be assessed by looking at the activation of the relevant node alone, independent of that of other nodes. One option is to assume instead that activation maps onto *relative* degree of confidence, so that degree of activation is interpreted relative to that of other nodes. Although strictly inconsistent with Thorpe's desire for direct interpretation, this preserves what is essential about a localist scheme, namely that the entity about which relative confidence is being expressed is identified with a single node. Alternatively, both Thorpe's and Barlow's definitions can be simultaneously maintained if some competitive process is implemented directly (i.e., mechanically), so that it is impossible to sustain simultaneously high activations at two nodes whose interpretations are contradictory. A scheme of this type would, for example, allow two nodes to compete for activation so as to exclusively identify a single person.

As an aside, note that a simple competitive scheme has some disadvantages. Such a scheme is apparently inadequate for indicating the presence of two entities, say, John and Paul, by strongly activating the two relevant nodes simultaneously. One solution to this apparent conundrum might be to invoke the notion of binding, perhaps implemented by phase relationships in node firing patterns (e.g., Hummel & Biedermann 1992; Roelfsema et al. 1996; Shastri & Ajjanagadde 1993). (Phase relationships are only one candidate means of perceptual binding and will be assumed here solely for illustrative purposes.) Thus, in the case in which we wish both John and Paul to be simultaneously indicated, both nodes can activate fully but out of phase with each other, thus diminishing the extent to which they compete. This out-of-phase relationship might stem from the fact that the two entities driving the system (John and Paul) must be in two different spatial locations, allowing them to be "phased" separately. In the alternative scenario, that is, when only one individual is present, the nodes representing alternative identifications might be in phase with each other, driven as they are by the same stimulus object, and would therefore compete as required.

A similar binding scheme might also be useful if distributed representations are employed. On the face of it, using the representations in Figure 1a, the pattern for John and George will be the same as that for Paul and Ringo. It may be possible to distinguish these summed patterns on the basis of binding relationships as before – to represent John and George the first and second nodes would be in phase with each other while the third and fourth nodes would both be out of phase with the first two nodes and in phase with each other. But complications arise when we wish to represent, say, John and Paul: would the second node be in phase with the first node or the third? It is possible that in this case the second node would fire in both the phases associated with nodes one and two (though this would potentially affect its firing rate as well as its phase relationships). Mechanisms for binding are the focus of a good deal of ongoing research, so I shall not develop these ideas further here.

2.3. Grandmother cells . . .

In a discussion of localist and distributed representations it is hard to avoid the subject of "grandmother cells." The concept can be traced back to a lecture series delivered by Jerome Lettvin in 1969 (see Lettvin's appendix to Barlow 1995), in which he introduced to a discussion on neural representation an allegory in which a neuroscientist located in the brains of his animal subjects "some 18,000 neurons . . . that responded uniquely only to the animal's mother, however displayed, whether animate or stuffed, seen from before or behind, upside down or on a diagonal, or offered by caricature, photograph or abstraction" (from appendix to Barlow 1995).

The allegorical neuroscientist ablated the equivalent cells in a human subject, who, postoperatively, could not conceive of "his mother," while maintaining a conception of mothers in general. The neuroscientist, who was intent on showing that "ideas are contained in specific cells," considered his position to be vulnerable to philosophical attack, and rued not having searched for grandmother cells instead, grandmothers being "notoriously ambiguous and often formless."

The term "grandmother cell" has since been used extensively in discussions of neural representation, though not always in ways consistent with Lettvin's original conception. It seems that (grand)mother cells are considered by some to be the necessary extrapolation of the localist approach and thereby to demonstrate its intrinsic folly. I believe this conclusion to be entirely unjustified. Whatever the relevance of Lettvin's allegory, it certainly does not demonstrate the necessary absurdity of (grand)mother cells and, even if it did, this would not warrant a similar conclusion regarding localist representations in general. Given the definitions so far advanced, it is clear that, while (grand)mother cells are localist representations, not all localist representations necessarily have the characteristics attributed by Lettvin to (grand)mother cells. This depends on how one interprets Lettvin's words "responded uniquely" (above). A localist representation of one's grandmother might respond partially, but subthreshold, to a similar entity (e.g., one's great aunt), thus violating one interpretation of the "unique response" criterion that forms part of the grandmother-cell definition.

2.4. . . . and yellow Volkswagen cells

A related point concerns the "yellow Volkswagen cells" referred to by Harris (1980). Harris's original point, which dates back to a talk given in 1968, illustrated a concern regarding a potential proliferation in the types of selective

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

cells hypothesized to be devoted to low-level visual coding. Such a proliferation had been suggested by experiments into, for instance, the "McCollough Effect" (McCollough 1965), which had led to the positing of detectors sensitive to particular combinations of orientation and colour. The message that has been extrapolated from Harris's observation is one concerning representational capacity: that while "yellowness" cells and "Volkswagen cells" may be reasonable, surely specific cells devoted to "yellow Volkswagens" are not. The fear is that if yellow VWs are to be locally represented then so must the combinatorially explosive number of equivalent combinations (e.g., lime-green Minis). There is something odd about this argument. In accepting the possibility of Volkswagen cells, it begs the question as to why the fear of combinatorial explosion is not already invoked at this level. Volkswagens themselves must presumably be definable as a constellation of a large number of adjective-noun properties (curved roof, air-cooled engine, etc.), and yet accepting the existence of Volkswagen cells does not presume a vast number of other cells, one for each distinct combination of feature-values in whatever feature-space VWs inhabit. On a related point, on occasions when the (extrapolated) yellow-VW argument is invoked, it is not always clear whether the supposed combinatorial explosion refers to the number of possible percepts, which is indeed unimaginably large, or to the vanishingly smaller number of percepts that are witnessed and, in some sense, worth remembering. Since the latter number is likely to grow only approximately linearly with lifespan, fears of combinatorial explosion are unwarranted. It is perfectly consistent with the localist position that different aspects of a stimulus (e.g., colour, brand name, etc.) can be represented separately, and various schemes have been suggested for binding such aspects together so as to correctly represent, in the short term, a given scene (e.g., Hummel & Biedermann 1992; Roelfsema et al. 1996; see earlier). This systematicity (cf. Fodor & Pylyshyn 1988) in the perceptual machinery addresses the problem of combinatorial explosion regarding the number of possible percepts. It in no way implies, however, that in a localist model each *possible* percept must be allocated its own permanent representation, that is, its own node. A similar point was made by Hummel and Holyoak (1997) who noted that "it is not necessary to postulate the preexistence of all possible conjunctive units. Rather a novel binding can first be represented dynamically (in active memory), with a conjunctive unit created only when it is necessary to store the binding in LTM" (p. 434).

It is entirely consistent with the localist position to postulate that cells encoding specific combinations will be allocated only when needed: perhaps in an experiment in which pictures of yellow VWs and red bikes require one response, while red VWs and yellow bikes require another (cf. XOR); or, more prosaically, in establishing the memory that one's first car was a yellow VW. When one restricts the number of localist representations to those sufficient to describe actual percepts of behavioural significance (i.e., those that require long-term memorial representation), the threat of combinatorial explosion dissipates. Later I shall show how new localist nodes can be recruited, as needed, for the permanent representation of previously unlearned configurations (cf. the constructivist learning of Quartz & Sejnowski 1997, and the accompanying commentary by Grossberg 1997; Valiant 1994).

2.5. Featural representations

The above discussion of yellow VWs illustrates the issue of *featural representation.* A featural representation will be defined here as a representation comprising an array of localist nodes in appropriate states. Figure 2 shows the featural representations of Tony Blair, Glenda Jackson, Anthony Hopkins, and Queen Elizabeth II, where the relevant features are "is-a-woman," "is-a-politician," and "is/was-a-film-actor." Clearly, the representations of these four entities are distributed, in the sense that the identity of the currently present entity cannot be discerned by examining the activity of any individual node. Nonetheless, the features themselves are locally represented (cf. "is-yellow," "is-a-Volkswagen"). Whether or not a politician is currently present can be decided by examining the activity of a single node, independent of the activation of any other node.

It is curious that researchers otherwise committed to the thoroughgoing use of distributed representations have been happy to use such featural representations. For instance, Farah et al. (1993), whose commitment to distributed representations was quoted earlier [see Farah: "Neuropsychological Inference with an Interactive Brain" *BBS* 17(1) 1994], used a distributed representation for semantic information relating to particular people. To continue the earlier quotation:

> The information encoded by a given unit will be some "microfeature" . . . that may or may not correspond to an easily labeled feature (such as eye color in the case of faces). The only units for which we have assigned an interpretation are the "occupation units" within the semantic pool. One of them represents the semantic microfeature "actor" and the other represents the semantic microfeature "politician." (Farah et al. 1993, p. 577)

It would be odd to be comfortable with the idea of nodes representing "is-an-actor," and yet hostile to the idea of nodes representing "is-Tony-Blair" or "is-my-grandmother." If "is-an-actor" is a legitimate microfeature (though one wonders what is micro about it), then why is "is-Tony-Blair" not? Is there any independent rationale for what can and cannot be a microfeature? Moreover, to anticipate a later discussion, by what learning mechanism are the localist (micro)featural representations (e.g., "is-an-actor") themselves deemed to be established? The most natural assumption is that, at some level, local unsupervised featural learning is carried out. But a commitment to fully distributed representation of identity, if not profession, would therefore require that at some arbitrary stage just before the level at which identity features (e.g., "is-Tony-Blair")

	woman	politician	actor
Tony Blair	0	1	0
Glenda Jackson	1	1	1
Anthony Hopkins	0	0	1
Elizabeth II	1	0	0

Figure 2. Four persons represented in a featural fashion with regard to semantic information.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

might emerge, a different, supervised learning mechanism cuts in.

Whether or not we choose to define featural representations as a subclass of distributed representations has little to do with the core of the localist/distributed debate. No localist has ever denied the existence of distributed representations, especially, but not exclusively, if these are taken to include featural representations. To do so would have entailed a belief that percepts "go local" in a single step, from retina directly to grandmother cell, for instance. The key tenet of the localist position is that, on occasion, localist representations of meaningful entities in the world (e.g., words, names, people, etc.) emerge and allow, among other things, distributed/featural patterns to be reliably classified and enduringly associated.

I should make clear that restricting the definition in the preceding paragraph to "meaningful entities in the world" is simply a rather clumsy way of avoiding potentially sterile discussions of how far localist representation extends down the perceptual hierarchy. To take a concrete example, one might ask whether an orientation column (OC) in the visual cortex should be considered a localist representation of line segments in a particular part of the visual field and at a particular angular orientation. An opponent of such a localist description might argue that in most everyday circumstances nothing of cognitive significance (nothing of meaning, if you like) will depend on the activation state of an individual OC and that later stages in the perceptual path will best be driven by a distributed pattern of activation across a number of OCs so as to preserve the information available in the stimulus. I am sympathetic to this argument – there seems little point in describing a representation as localist if it is never interpreted in a localist manner. Nonetheless, to temper this conclusion somewhat, imagine an experiment in which a response is learned that depends on which of two small line segments, differing only in orientation, is presented. Assuming that such a discrimination is learnable, it does not seem impossible a priori that a connectionist model of the decision task would depend rather directly on the activations of specific OCs. (The issue is related to the decoding of population vectors, discussed briefly in section 4.3.1 and in the accompanying footnote.) I have not modelled performance in this rather contrived task and hence cannot say what should be concluded from such a model. One can simply note that certain models might lead to a more charitable view toward an interpretation that treated single OCs as localist representations. The general point is that a representation might be labelled localist or not depending on the particulars of the modelled task in which the corresponding nodes are taken to be involved. Whether one chooses to reserve the term localist for representations that are habitually involved in processes/tasks that highlight their localist character or, alternatively, whether one allows the term to apply to any representational unit that can at some time (albeit in unusual or contrived circumstances) usefully be treated as localist, is probably a matter of taxonomic taste. For fear of getting unnecessarily involved in such matters, I will retreat to using the term localist to refer, as above, to a form of representation of meaningful entities in the world whose localist character is habitually displayed. I do so in the hope and belief that, at least in the modelling of most types of cognitive-psychological task, it will be clear what the relevant meaningful entities are.

2.6. What is a localist model?

Given the definitions of localist and distributed representations discussed so far, what are we to understand by the term "a localist model"? The first and most crucial point, alluded to above, is that a localist model is not well defined as one that uses localist rather than distributed representations: localist models almost always use both localist and distributed representations. More explicitly, any entity that is locally represented at layer n of a hierarchy is sure to be represented in a distributed fashion at layer $n - 1$. To illustrate, take as an example the interactive activation (IA) model of visual word recognition (McClelland & Rumelhart 1981; Rumelhart & McClelland 1982), which is generally agreed to be localist. It uses successive processing layers: In the "lowest" of these are visual-feature detectors, which respond selectively to line segments in various orientations; in the next layer are nodes that respond selectively to letters in various positions in a word; in the third are nodes that respond maximally to individual familiar words. Thus, a given word is represented locally in the upper layer and in a distributed fashion at the two previous layers. Letters-in-position are likewise represented locally in the second layer but in a distributed manner in the first layer. It accordingly makes no sense to define a localist model as one that precludes distributed representation. A better definition relies only on whether or not there are localist representations of the relevant entities.

It so happens that, in the IA example, the distributed representations at lower layers are of the featural variety, as discussed above. This, however, is not a crucial factor in the IA model's being labelled localist: The lower layers might have used distributed representations unamenable to a featural characterization without nullifying the fact that in the upper layer a localist code is used. The difference between localist and distributed models is most often not in the nature or status of the representation of the input patterns, which depends ultimately (in vivo) on the structure and function of the relevant sense organ(s), but in the nature of representation at the later stages of processing that input. As stated above, localists posit that certain cognitively meaningful entities will be represented in a local fashion at some, probably late, level of processing, and it is at this level that decisions about which entities are identifiable in any given input can best be made.

So can the term "localist model" be universally applied to models using localist representations? Not without care. Consider the model of reading proposed by Plaut et al. (1996). This was developed from the seminal model of Seidenberg and McClelland (1989), in which neither letters at the input nor phonemes at the output were represented in a local fashion. According to Plaut et al., it was this aspect of the model, among others, which manifested itself in its relatively poor nonword reading. Plaut et al. referred to this as the "dispersion problem." Perhaps, as Jacobs and Grainger (1994) rather archly suggest, it might better have been termed the distribution problem, given that Plaut et al.'s solution entailed a move to an entirely local scheme for both input orthography (letters and letter clusters) and output phonemes. And yet, even with this modification, it would be very misleading to call Plaut et al.'s a localist model: The most powerful and theoretically bold contribution of that model was to show that the mapping between orthographic representations of both words and nonwords and their pro-

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

nunciations could be carried out in a distributed fashion, that is, without any recourse to either a locally represented mental lexicon or an explicit system of grapheme-to-phoneme correspondence rules. So whereas the Plaut et al. model was certainly localist at the letter and phoneme levels, it was undeniably distributed at the lexical level. It is for this reason that calling that model localist would be thoroughly misleading. I conclude that the term "localist model" should be used with care. In most cases, it will be better to be explicit about the entities for which localist coding is used (if any), and to identify the theoretical significance of this choice.

A further point should be made regarding localist models, again taking the IA model as our example. When a word is presented to the IA model, a large number of nodes will be maximally active – those representing certain visual features, letters-in-position, and the word itself. A number of other nodes will be partially active. On presentation of a nonword, no word-node will attain maximal activation but otherwise the situation will be much the same. The simple point is this: The fact that activity is distributed widely around the network should not lead to the incautious suggestion that the IA model is a distributed rather than a localist model. As noted earlier, it is important to distinguish between distributed processing and distributed representation. Having made this distinction we can better interpret labels that have been applied to other models in the literature, labels that might otherwise have the potential to confuse.

As an illustration, consider the Learning and Inference with Schemas and Analogies (LISA) model of Hummel and Holyoak (1997), as applied to the processing of analogy. The title of the paper ("Distributed representations of structure: A theory of analogical access and mapping") might suggest that LISA is a fully distributed model, but a closer look reveals that it uses localist representation. For instance, in its representation of the proposition "John loves Mary," there is a node corresponding to the proposition itself and to each of the constituents "John," "Mary," and "loves"; these nodes project in turn onto a layer of semantic units that are crucially involved in the analogy processing task. The whole network is hierarchically structured, with activity distributed widely for any given proposition and, in this case, organized in time so as to reflect various bindings of, for example, subjects with predicates. (Most, if not all, models that use phase binding do so in the context of localist representation.) LISA thus constitutes a clear example of the interaction between localist representations of entities and a distributed or featural representation of semantics. As in the IA model, there is no contradiction between distributed processing and localist representation. At the risk of overstating the case, we can see exactly the same coexistence of local representation, distributed representation, and distributed processing in what is often considered a quintessentially localist model, namely, Quillian's (1968) model of semantics. Quillian's justly influential model did indeed represent each familiar word with a localist "type" unit. But a word's meaning was represented by an intricate web of structured connections between numerous tokens of the appropriate types, resulting, on activation of a given word-type, in a whole plane of intricately structured spreading activation through which semantic associative relationships could become apparent.

To summarize, a localist model of a particular type of entity (e.g., words) is characterized by the presence of (at least) one node that responds maximally to a given familiar (i.e., learned) example of that type (e.g., a given familiar word), all familiar examples of that type (e.g., all familiar words) being so represented. This does not preclude some redundancy in coding. For example, in the word example used here, it may be that various versions of the same word (e.g., different pronunciations) are each represented locally, though in many applications these various versions would be linked at some subsequent stage so as to reflect their lexical equivalence.

It is hoped that this definition of what constitutes a localist model will help to clarify issues of model taxonomy. Under this taxonomy, the term "semilocalist" would be as meaningless as the term "semipregnant." But what are we to make of representations that are described as "sparse distributed" or "semidistributed"? It is rather difficult to answer this question in general because there is often no precise definition of what is meant by these terms. Sparse distributed representational schemes are frequently taken to be those for which few nodes activate for a given stimulus with few active nodes shared between stimuli, but this definition begs a lot of questions. For example, how does the definition apply to cases in which nodes have continuous rather than binary activations? To qualify as a sparse distributed representational scheme, are nodes required to activate to identical degrees for several different stimuli (cf. Kanerva's binary sparse distributed memory, Kanerva 1988; Keeler 1988)? Or are nodes simply required to activate (i.e., significantly above baseline) for more than one stimulus? Certainly in areas in which the term "sparse distributed" is often employed, such as in the interpretation of the results of single-cell recording studies, the latter formulation is more consistent with what is actually observed. As will be pointed out later, however, it is not really clear what distinction can be made between a sparse distributed scheme defined in this way and the localist schemes discussed above – after all, the localist IAM model would be classified as sparse distributed under this looser but more plausible definition. If the class of sparse distributed networks is defined so as to include both localist and nonlocalist networks as subclasses (as is often the case), then statements advocating the use of sparse distributed representation cannot be interpreted as a rejection of localist models.

A similar problem exists with the term "semidistributed." French (1992) discusses two systems he describes as semidistributed. The first is Kanerva's sparse distributed memory (Kanerva 1988), a network of binary neurons inspired more by a digital computer metaphor than by a biological metaphor, but which nonetheless shows good tolerance to interference (principally due to the similarities it shares with the localist models described here). The second is Kruschke's (1990) ALCOVE model, which (in its implemented version at least) would be classified under the present definition as localist. French developed a third type of semidistributed network, using an algorithm that sought to "sharpen" hidden unit activations during learning. Unfortunately, this semidistributed network only semisolved the interference problem to which it was addressed, in that even small amounts of later learning could interfere drastically with the ability to perform mappings learned earlier. What advantage there was to be gained from using a semidistributed network was primarily to be found in a

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

measure of time to relearn the original associations – some compensation but hardly a satisfactory solution to the interference problem itself.

It is informative to note that French's (1992) motivation for using a semidistributed rather than a localist network was based on his assumption that localist models acquire their well-known resistance to interference by sacrificing their ability to generalize. In what follows I will question this common assumption and others regarding localist models, thus weakening the motivation to seek semidistributed solutions to problems that localist networks already solve.

3. Organization of the argument

Before launching into the detail of my remaining argument, I will first signpost what can be expected of the remainder of this target article. This is necessary because, as will be seen, on the way to my conclusion I make some moderately lengthy, but I hope interesting, digressions. These digressions may seem especially lengthy to those for whom mathematical modelling is of little interest. Nonetheless, I hope that the end justifies the means, particularly since the approach adopted here results in a model that is practically equivalent to several mathematical models, but with most of the mathematics taken out.

It seems essential, in writing an article such as this, to emphasize the positive qualities of localist models as much as to note the shortcomings of their fully distributed counterparts. In the next part of the paper, I accordingly develop a generalized localist model, which, despite its simplicity, is able to exhibit generalization and attractor behaviour – abilities more commonly associated with fully distributed models. This is important because the absence of these abilities is often cited as a reason for rejecting localist models. The generalized localist model is also able to perform stable supervised and unsupervised learning and qualitatively to model effects of age of acquisition, both of which appear difficult for fully distributed models. The model is further shown to exhibit properties compatible with some mathematical formulations of great breadth, such as the Luce choice rule and the "power law of practice," thus extending the potential scope of its application.

In later sections I consider why, given the power of localist models, some psychological modellers have been reluctant to use them. These parts of the paper identify what I believe to be common misconceptions in the literature, in particular those based on conclusions drawn from the domain of neuroscience. Finally, I address some of the problems of a fully distributed approach and identify certain inadequacies in some of the measures that have been proposed to overcome these problems.

4. A generalized localist model

In this section I shall describe, in general terms, a localist approach to both the unsupervised learning of representations and the supervised learning of pattern associations. In characterizing such a localist approach I have sought to generalize from a number of different models (e.g., Burton 1994; Carpenter & Grossberg 1987a; 1987b; Foldiak 1991; Kohonen 1984; Murre 1992; Murre et al. 1992; Nigrin 1993; Rumelhart & Zipser 1986). These models differ in their details but are similar in structure and I shall attempt to draw together the best features of each. The resulting generalized model will not necessarily be immediately applicable to any particular research project but it will, I hope, have sufficient flexibility to be adapted to many modelling situations.

4.1. A learning module

As a first step in building a localist system, I will identify a very simple module capable of unsupervised, self-organized learning of individual patterns and/or pattern classes. This work draws heavily on the work of Carpenter and Grossberg and colleagues (e.g., Carpenter & Grossberg 1987a; 1987b; a debt that is happily acknowledged), with a number of simplifications. The module (see Fig. 3) comprises two layers of nodes, L_1 and L_2, fully connected to each other by modifiable, unidirectional (L_1-to-L_2) connections, which, prior to learning, have small, random weights, w_{ij}. (Throughout the paper, w_{ij} will refer to the weight of the connection from the ith node in the originating layer to the jth node in the receiving layer.) For simplicity of exposition, the nodes in the lower layer will be deemed to be binary, that is, to have activations (denoted a_i) either equal to zero or to one. The extension to continuous activations will usually be necessary and is easily achieved. The input to the nodes in the upper layer will simply be sum of the activations at the lower layer weighted by the appropriate connection weights. In fact, for illustrative purposes, I shall assume here that this input to a given node is divided by a value equal to the sum of the incoming weights to that node plus a small constant (see, e.g., Marshall 1990) – this is just one of the many so-called "normalization" schemes typically used with such networks. Thus the input, I_j, to an upper-layer node is given by

$$I_j = \frac{\sum_{\text{all i}} a_i w_{ij}}{\left(\sum_{\text{all i}} w_{ij}\right) + \alpha}, \qquad (1)$$

where α is the small constant. Learning of patterns of activation at the lower layer, L_1, is simply achieved as follows. When a pattern of activation is presented at L_1, the inputs, I_j, to nodes in the upper layer, L_2, can be calculated. Any node whose vector of incoming weights is parallel (i.e., a constant multiple of) the vector of activations at L_1 will have input, I_j, equal to $\frac{1}{1+\alpha/(\Sigma_{\text{all i}} w_{ij})}$. Any L_2 node whose vector of incoming weights is orthogonal to the current input vector (that is, nodes for which $w_{ij} = 0$ where $a_i = 1$) will have zero input. Nodes with weight vectors between these two extremes, whose weight vectors "match" the current activation vector to some nonmaximal extent, will have intermediate values of I_j. Let us suppose that, on presentation of a given L_1 pattern, no L_2 node achieves an input, I_j, greater than a threshold θ. (With θ set appropriately, this supposition will hold when no learning has yet been carried out in the L_1-to-L_2 connections.) In this case, learning of the current input pattern will proceed. Learning will comprise setting the incoming weights to a single currently "uncommitted" L_2 node (i.e., a node with small, random incoming weights) to equal the corresponding activations at L_1 – a pos-

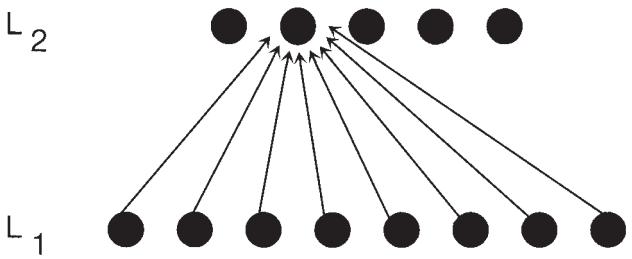

Figure 3. A simple two-layer module.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

sible mechanism is discussed later. The learning rule might thus be stated

$$\frac{dw_{iJ}}{dt} = \lambda(a_i - w_{iJ}), \quad (2)$$

where J indexes the single L_2 node at which learning is being performed, λ is the learning rate, and, in the case of so-called "fast learning," the weight values reach their equilibrium values in one learning episode, such that $w_{iJ} = a_i$. The L_2 node indexed by J is thereafter labelled as being "committed" to the activation pattern at L_1, and will receive its maximal input on subsequent presentation of that pattern.

The course of learning is largely determined by the setting of the threshold, θ. This is closely analogous to the vigilance parameter in the adaptive resonance theory (ART) networks of Carpenter and Grossberg (1987a; 1987b), one version of which (ART2a, Carpenter et al. 1991) is very similar to the network described here. More particularly, if the threshold is set very high, for example, $\theta = 1$, then each activation pattern presented will lead to a learning episode involving commitment of a previously uncommitted L_2 node, even if the same pattern has already been learned previously. If the threshold is set slightly lower, then only activation patterns sufficiently different from previously presented patterns will provoke learning. Thus novel patterns will come to be represented by a newly assigned L_2 node, without interfering with any learning that has previously been accomplished. This is a crucial point in the debate between localist and distributed modellers and concerns "catastrophic interference," a topic to which I shall return in greater depth later.

The value of the threshold, θ, need not necessarily remain constant over time. This is where the concept of vigilance is useful. At times when the vigilance of a network is low, new patterns will be unlikely to be learned and responses (see below) will be based on previously acquired information. In situations where vigilance, θ, is set high, learning of the current L_1 pattern is likely to occur. Thus learning can, to some extent, be influenced by the state of vigilance in which the network finds itself at the time of pattern presentation.

In order to make these notions more concrete, it is necessary to describe what constitutes a response for the network described above. On presentation of a pattern of activation at L_1, the inputs to the L_2 nodes can be calculated as above. These inputs are then thresholded so that the net input, I_j^{net}, to the upper-layer nodes is given by

$$I_j^{\text{net}} = \max(0,[I_j - \theta]), \quad (3)$$

so that nodes I_j with less than θ will receive no net input, and other nodes will receive net input equal to the degree by which I_j exceeds θ. Given these net inputs, there are several options as to how we might proceed. One option is to allow the L_2 activations to equilibrate via the differential equation

$$\frac{da_j}{dt} = -a_j + f(I_j^{\text{net}}), \quad (4)$$

which reaches equilibrium when $\frac{da_j}{dt} = 0$, that is, when a_j is some function, f, of the net input. A common choice is to assume that f is the identity function, so that the activations equal the net inputs. Another option, which will be useful when some sort of decision process is required, is to allow the L_2 nodes to compete in some way. This option will be developed in some detail in the next section because it proves to have some very attractive properties. For the moment it is sufficient to note that a competitive process can lead to the selection of one of the L_2 nodes. This might be achieved by some *winner-takes-all* mechanism, by which the L_2 nodes compete for activation until one of them quenches the activation of its competitors and activates strongly itself. Or it may take the form of a "horse-race," by which the L_2 nodes race, under the influence of the bottom-up inputs, I_j^{net}, to reach a criterion level of activation, χ. Either way, in the absence of noise, we will expect the L_2 node with the greatest net input, I_j^{net}, to be selected. In the case where the Jth L_2 node is selected, the pattern at L_1 will be deemed to have fallen into the Jth pattern class. (Note that in a high-θ regime there may be as many classes as patterns presented.) On presentation of a given input pattern, the selection (in the absence of noise) of a given L_2 node indicates that the current pattern is most similar to the pattern learned by the selected node, and that the similarity is greater than some threshold value.

To summarize, in its noncompetitive form, the network will respond so that the activations of L_2 nodes, in response to a given input (L_1) pattern, will equilibrate to values equal to some function of the degree of similarity between their learned pattern and the input pattern. In its competitive form the network performs a classification of the L_1 activation pattern, where the classes correspond to the previously learned patterns. This results in sustained or super-criterion activation ($a_j > \chi$) of the node that has previously learned the activation pattern most similar to that currently presented. In both cases, the network is *self-organizing* and *unsupervised.* "Self-organizing" refers to the fact that the network can proceed autonomously, there being, for instance, no separate phases for learning and for performance. "Unsupervised" is used here to mean that the network does not receive any external "teaching" signal informing it how it should classify the current pattern. As will be seen later, similar networks will be used when supervised learning is required. In the meantime, I shall simply assume that the selection of an L_2 node will be sufficient to elicit a response associated with that node (cf. Usher & McClelland 1995).

4.2. A competitive race

In this section I will give details of one way in which competition can be introduced into the simple module described above. Although the network itself is simple, I will show that it has some extremely interesting properties relating to choice accuracy and choice reaction-time. I make no claim to be the first to note each of these properties; nonetheless, I believe the power they have in combination has either gone unnoticed or has been widely underappreciated.

Competition in the layer L_2 is simulated using a standard "leaky integrator" model that describes how several nodes, each driven by its own input signal, activate in the face of decay and competition (i.e., inhibition) from each of the other nodes:

$$\frac{da_j}{dt} = -Aa_j + (I_j^{\text{net}} + N_1) + f_1(a_j) - C\sum_{k\neq j} f_2(a_k) + N_2 \quad (5)$$

where A is a decay constant; I_j is the excitatory input to the jth node, which is perturbed by zero-mean Gaussian noise, N_1, with variance σ_1^2; $f_1(a_j)$ is a self-excitatory term; $C\ \Sigma_{k\neq j}\ f_2\ (a_k)$ represents lateral inhibition from other nodes in L_2; and N_2 represents zero-mean Gaussian noise with variance σ_2^2. The value of the noise term, N_1, remains constant over the time course of a single competition since it is intended to represent inaccuracies in "measurement" of I_j. By contrast, the value of N_2 varies with each time step, representing moment-by-moment noise in the calculation of the derivative. Such an equation has a long history in neural modelling, featuring strongly in the work of Grossberg from the mid 1960s onwards and later in, for instance, the cascade equation of McClelland (1979).

4.2.1. Reaction time. Recently, Usher and McClelland (1995) have used such an equation to model the time-course of perceptual choice. They show that, in simulating various two-alternative forced choice experiments, the above equation subsumes optimal classical diffusion processes (e.g., Ratcliff 1978) when a response criterion is placed on the difference between the activations, a_j, of two competing nodes. Moreover, they show that near optimal performance is exhibited when a response criterion is placed on the

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

absolute value of the activations (as opposed to the difference between them) in cases where, as here, mutual inhibition is assumed. The latter case is easily extensible to multiway choices. This lateral inhibitory model therefore simulates the process by which multiple nodes, those in L_2, can activate in response to noisy, bottom-up, excitatory signals, I^{net}, and compete until one of the nodes reaches a response criterion based on its activation alone. Usher and McClelland (1995) have thus shown that a localist model can give a good account of the time course of multiway choices.

4.2.2. Accuracy and the Luce choice rule. Another interesting feature of the lateral inhibitory equation concerns the accuracy with which it is able to select the appropriate node (preferably that with the largest bottom-up input, I_j) in the presence of input noise, N_1, and fast-varying noise, N_2. Simulations show that in the case where the variances of the two noise terms are approximately equal, the effects of the input noise, N_1, dominate – this is simply because the leaky integrator tends to act to filter out the effects of the fast-varying noise, N_2. As a result, the competitive process tends to "select" that node which receives the maximal noisy input, $(I_j^{net} + N_1)$. This process, by which a node is chosen by adding zero-mean Gaussian noise to its input term and picking the node with the largest resultant input, is known as a Thurstonian process (Thurstone 1927, Case V). Implementing a Thurstonian process with a lateral-inhibitory network of leaky integrators, as above, rather than by simple computation, allows the dynamics as well as the accuracy of the decision process to be simulated.

The fact that the competitive process is equivalent to a classical Thurstonian (noisy-pick-the-biggest) process performed on the inputs, I_j, is extremely useful because it allows us to make a link with the Luce choice rule (Luce 1959), ubiquitous in models of choice behaviour. This rule states that, given a stimulus indexed by i, a set of possible responses indexed by j, and some set of similarities, η_{ij}, between the stimulus and those stimuli associated with each member of the response set, then the probability of choosing any particular response, J, when presented with a stimulus, i, is

$$p(J \mid i) = \frac{\eta_{iJ}}{\sum_{\text{all k}} \eta_{ik}}. \tag{6}$$

Naturally this ensures that the probabilities add up to 1 across the whole response set. Shepard (1958; 1987) has proposed a law of generalization which states, in this context, that the similarities of two stimuli are an exponential function of the distance between those two stimuli in a multidimensional space. (The character of the multidimensional space that the stimuli inhabit can be revealed by multidimensional scaling applied to the stimulus-response confusion matrices.) Thus $\eta_{ij} = e^{-d_{ij}}$, where

$$d_{ij} = \left[\sum_{m=1}^{M} c_m \mid i_m - j_m \mid^r \right]^{1/r}, \tag{7}$$

where the distance is measured in M-dimensional space, i_m represents the coordinate of stimulus i along the mth dimension, and c_m represents a scaling parameter for distances measured along dimension m. The scaling parameters simply weigh the contributions of different dimensions to the overall distance measure, much as one might weigh various factors such as reliability and colour when choosing a new car. Equation 7 is known as the Minkowski power model formula and d_{ij} reduces to the "city-block" distance for $r = 1$ and the Euclidean distance for $r = 2$, these two measures being those most commonly used.

So how does the Luce choice rule acting over exponentiated distances relate to the Thurstonian (noisy choice) process described above? The illustration is easiest to perform for the case of two-alternative choice, and is the same as that found in McClelland (1991). Suppose we have a categorization experiment in which a subject sees one exemplar of a category A and one exemplar of a category B. We then present a test exemplar, T, and ask the subject to decide whether it should be categorized as being from category A or category B. Suppose further that each of the three exemplars can be represented by a point in an appropriate multidimensional space such that the test exemplar lies at a distance d_A from the A exemplar and d_B from the B exemplar. This situation is illustrated for a two-dimensional space in Figure 4. Note that coordinates on any given dimension are given in terms of the relevant scaling parameters, with the increase of a given scaling parameter resulting in an increase in magnitude of distances measured along that dimension and contributing to the overall distance measures d_A and d_B. (It is for this reason that an increase of scaling parameter along a given dimension is often described as a stretching of space along that dimension.) The Luce choice rule with exponential generalization implies that the probability of placing the test exemplar in category A is

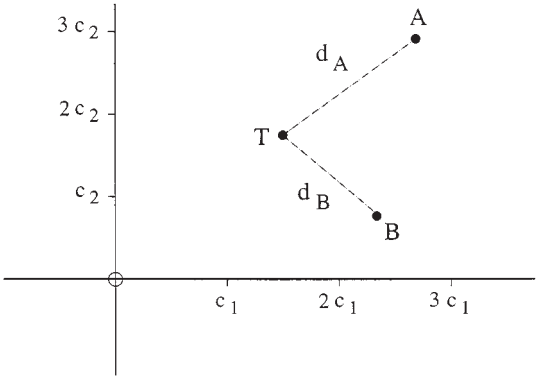

Figure 4. Locations of exemplars A and B and test pattern T in a two-dimensional space. The distances d_A and d_B are Euclidean distances, and coordinates on each dimension are given in terms of scaling parameters, c_i.

$$p(\text{test is an A}) = \frac{e^{-d_A}}{e^{-d_A} + e^{-d_B}}; \tag{8}$$

dividing top and bottom by e^{-d_A} gives

$$p(\text{test is an A}) = \frac{1}{1 + e^{d_A - d_B}}, \tag{9}$$

which equals 0.5 when $d_A = d_B$. This function is called the logistic function, and it is extremely similar to the function describing the (scaled) cumulative area under a normal (i.e., Gaussian) curve. This means that there is a close correspondence between the two following procedures for probabilistically picking one of two responses at distances d_A and d_B from the current stimulus: one can either add Gaussian noise to d_A and d_B and pick the category corresponding to, in this case, the smallest resulting value (a Thurstonian process); or one can exponentiate the negative distances and pick using the Luce choice rule. The two procedures will not give identical results, but in most experimental situations will be indistinguishable (e.g., Kornbrot 1978; Luce 1959; Nosofsky 1985; van Santen & Bamber 1981; Yellott 1977). (In fact, if the noise is double exponential rather than Gaussian, the correspondence is exact; see Yellott 1977.)

The consequences for localist modelling of this correspondence, which extends to multichoice situations, are profound. Two things should be kept in mind. First, a point in multidimensional space can be represented by a vector of activations across a layer of nodes, say the L_1 layer of the module discussed earlier, and/or by a vector of weights, perhaps those weights connecting the set of L_1 nodes to a given L_2 node. Second, taking two nodes with activations d_A and d_B, adding zero-mean Gaussian noise, and pick-

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

ing the node with the smallest resulting activation is equivalent, in terms of the node chosen, to taking two nodes with activations $(k - d_A)$ and $(k - d_B)$ (where k is a constant), adding the same zero-mean Gaussian noise and this time picking the node with the largest activation. Consequently, suppose that we have two L_2 nodes, such that each has a vector of incoming weights corresponding to the multidimensional vector representing one of the training patterns, one node representing pattern *A*, the other node representing pattern *B*. Further, suppose that on presentation of the pattern of activations corresponding to the test pattern, *T*, at layer L_1, the inputs, I_j, to the two nodes are equal to $(k-d_A)$ and $(k-d_B)$, respectively (this is easily achieved). Under these circumstances, a Thurstonian process, like that described above, which noisily picks the L_2 node with the biggest input will give a pattern of probabilities of classifying the test pattern either as an *A* or, alternatively, as a *B*, which is in close correspondence to the pattern of probabilities that would be obtained by application of the Luce choice rule to the exponentiated distances, $e^{-d_{ij}}$ (Equation 9).

4.2.3. Relation to models of categorization. This correspondence means that mathematical models, such as Nosofsky's generalized context model (Nosofsky 1986), can be condensed, doing away with the stage in which distances are exponentiated, and the stage in which these exponentiated values are manipulated, in the Luce formulation, to produce probabilities (a stage for which, to my knowledge, no simple "neural" mechanism has been suggested[1]), leaving a basic Thurstonian noisy choice process, like that above, acting on (a constant minus) the distances themselves. Since the generalized context model (Nosofsky 1986) is, under certain conditions, mathematically equivalent to the models of Estes (1986), Medin and Schaffer (1978), Oden and Massaro (1978), Fried and Holyoak (1984), and others (see Nosofsky 1990), all these models can, under those conditions, similarly be approximated to a close degree by the simple localist connectionist model described here. A generalized exemplar model is obtained under high-θ (i.e., high threshold or high vigilance) conditions, when all training patterns are stored as weight vectors abutting a distinct L_2 node (one per node).

4.2.4. Effects of multiple instances. We can now raise the question of what happens when, in the simple two-choice categorization experiment discussed above, the subject sees multiple presentations of each of the example stimuli before classifying the test stimulus. To avoid any ambiguity I will describe the two training stimuli as "exemplars" and each presentation of a given stimulus as an "instance" of the relevant exemplar. For example, in a given experiment a subject might see ten instances of a single category A exemplar and ten instances of a single category B exemplar. Let us now assume that a high-θ classification network assigns a distinct L_2 node to each of the ten instances of the category A exemplar and to each of the ten instances of the category B exemplar. There will thus be twenty nodes racing to classify any given test stimulus. For simplicity, we can assume that the learned, bottom-up weight vectors to L_2 nodes representing instances of the same exemplar are identical. On presentation of the test stimulus, which lies at distance d_A (in multidimensional space) from instances of the category A exemplar and distance d_B from instances of the category B exemplar, the inputs to the L_2 nodes representing category A instances will be $(k - d_A) + N_{1j}$, where N_1 is, as before, the input-noise term and the subscript j $(0 \leq j < 10)$ indicates that the zero-mean Gaussian noise term will have a different value for each of the ten nodes representing different instances. The L_2 nodes representing instances of the category B exemplar will likewise have inputs equal to $(k - d_B) + N_{1j}$, for $10 \leq j < 20$. So what is the effect on performance of adding these extra instances of each exemplar? The competitive network will once again select the node with the largest noisy input. It turns out that, as more and more instances of each exemplar are added, two things happen. First, the noisy-pick-the-biggest process becomes an increasingly better approximation to the Luce formulation, until for an asymptotically large number of instances the correspondence is exact. Second, performance (as measured by the probability of picking the category whose exemplar falls closest to the test stimulus) improves, a change that is equivalent to stretching the multidimensional space in the Luce formulation by increasing by a common multiplier the values of all the scaling parameters, c_m, in the distance calculation given in Equation 7. For the mathematically inclined, I note that both these effects come about owing to the fact that the maximum value out of N samples from a Gaussian distribution is itself cumulatively distributed as a double exponential, $\exp(-e^{-ax})$, where a is a constant. The distribution of differences between two values drawn from the corresponding density function is a logistic function, comparable with that implicit in the Luce choice rule (for further details see, e.g., Yellott 1977; Page & Nimmo-Smith, in preparation).

To summarize, for single instances of each exemplar in a categorization experiment, performance of the Thurstonian process is a good enough approximation to that produced by the Luce choice rule such that the two are difficult to distinguish by experiment. As more instances of the training exemplars are added to the network, the Thurstonian process makes a rapid approach toward an asymptotic performance that is precisely equivalent to that produced by application of the Luce choice rule to a multidimensional space that has been uniformly "stretched" (by increasing by a common multiplier the values of all the scaling parameters, c_m, in Equation 7) relative to that space (i.e., the set of scaling parameters) that might have been inferred from the application of the same choice rule to the pattern of responses found after only a single instance of each exemplar had been presented. It should be noted that putting multiple noiseless instances into the Luce choice rule will not produce an improvement in performance relative to that choice rule applied to single instances – in mathematical terminology, the Luce choice rule is insensitive to uniform expansion of the set (Yellott 1977).

Simulations using the Luce formulation (e.g., Nosofsky 1987) have typically used uniform multiplicative increases in the values of the dimensional scaling parameters (the c_m in Equation 7) to account for performance improvement over training blocks. The Thurstonian process described here, therefore, has the potential advantage that this progressive space-stretching is a natural feature of the model as more instances are learned. Of course, until the model has been formally fitted to experimental data, the suggestion of such an advantage must remain tentative – indeed early simulations of data from Nosofsky (1987) suggest that some parametric stretching of stimulus space is still required to maintain the excellent model-to-data fits that Nosofsky achieved (Page & Nimmo-Smith, in preparation). Nonetheless, the present Thurstonian analysis potentially unites a good deal of data, as well as raising a fundamental question regarding Shepard's "universal law of generalization." Could it be that the widespread success encountered when a linear response rule (Luce) is applied to representations with exponential generalization gradients in multidimensional stimulus space is really a consequence of a Thurstonian decision process acting on an exemplar model, in which each of the exemplars actually responds with a linear generalization gradient? It is impossible in principle to choose experimentally between these two characterizations for experiments containing a reasonable number of instances of each exemplar. The Thurstonian (noisy-pick-the-biggest) approach has the advantage that its "neural" implementation is, it appears, almost embarrassingly simple.

4.2.5. The power law of practice. On the basis of the work described above, we can conclude that a simple localist, competitive model is capable of modelling data relating to both choice reaction-time and choice probability. In this section I will make a further link with the so-called "power law of practice." There is a large amount of data that support the conclusion that reaction-time varies as a power function of practice, that is, RT = A +

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

BN^{-c}, where N is the number of previous practice trials, and A, B, and c are positive constants (see Logan 1992 for a review). In a number of papers, Logan (1988; 1992) has proposed that this result can be modelled by making the following assumptions. First, each experience with a stimulus pattern is obligatorily stored in memory and associated with the appropriate response. Second, on later presentation of that stimulus pattern, all representations of that stimulus race with each other to reach a response criterion, the response time being the time at which the first of the representations reaches that criterion. The distribution of times-to-criterion for a given representation is assumed to be the same for all representations of a given stimulus and independent of the number of such representations. Logan has shown that if the time-to-criterion for a given stimulus representation is distributed as a Weibull function (which, within the limits of experimental error, it is – see Logan 1992), then, using the minimum-value theorem, the distribution of first-arrival times for a number of such representations is a related Weibull function, giving a power function of practice. He has collected a variety of data that indicate that this model gives a good fit to data obtained from practiced tasks and thus a good fit to the power law of practice (e.g., Logan 1988; 1992). At several points, Logan makes analogies with the exemplar models of categorization and identification discussed above but does not develop the point further.

This link between Logan's instance model and exemplar models of categorization has, however, been most fully developed in two recent papers by Robert Nosofsky and Thomas Palmeri (Nosofsky & Palmeri 1997; Palmeri 1997). Although they draw on Logan's instance theory of reaction-time (RT) speed-up, they identify a flaw in its behaviour. Their criticism involves observing that Logan's theory does not take stimulus similarity into account in its predictions regarding RT. To illustrate their point, suppose we have trained our subject, as above, with ten instances of a category A exemplar and ten instances of a category B exemplar. Consistent with Logan's instance theory, if we had tested performance on the test exemplar at various points during training, we would have observed some RT speed-up – the RT is decided by the time at which the first of the multiple stored instances crosses the winning line in a horse race to criterion. The time taken for this first crossing decreases as a power law of the number of instances. But what happens if we now train with one or more instances of an exemplar that is very similar to the category A exemplar, but is indicated as belonging to category B? In Logan's model, the addition of this third exemplar can only speed up responses when the category A exemplar is itself presented as the test stimulus. But intuitively, and actually in the data, such a manipulation has the opposite effect – that is, it slows responses.

Nososfsky and Palmeri (1997) solve this problem by introducing the exemplar-based random walk model (EBRW). In this model, each training instance is represented in memory, just as in Logan's model. Unlike in Logan's model, however, the response to a given stimulus is not given by a single race-to-criterion. For exemplar similarities to have the appropriate effect, two changes are made: First, each exemplar races with a speed proportional to its similarity to the test stimulus. Second, the result of a given race-to-criterion does not have an immediate effect on the response but instead drives a random-walk, category-decision process similar to that found in Ratcliff's (1978) diffusion model – multiple races-to-criterion are held consecutively, with their results accumulating until a time when the cumulative number of results indicating one category exceeds the cumulative number indicating the other category by a given margin. Briefly, as the number of instances increases, each of the races-to-criterion takes a shorter time, as in the Logan model; to the extent that the test stimulus falls clearly into one category rather than another, the results of consecutive races will be consistent in indicating the same category, and the overall response criterion will be more quickly reached.

I now suggest an extension of the Thurstonian model developed earlier, which accounts, qualitatively at least, for the data discussed by Nosofsky and Palmeri (1997) and addresses the problems inherent in Logan's instance model. The network is depicted in Figure 5. The activation pattern across the lower layer, L_1, is equivalent to the vector in multidimensional space representing the current test stimulus. Each node in the middle layer, L_2, represents a previously learned instance and activates to a degree inversely and linearly related to the distance between that learned instance and the current test stimulus (as before), plus zero-mean Gaussian noise, N_1. The third layer of nodes, L_3, contains a node for each of the possible category responses; L_2 nodes are linked by a connection of unit strength to the appropriate L_3 response node and to no other. The effective input to each of the L_3 nodes is given by the maximum activation value across the connected L_2 nodes. Thus the input driving a category A response will be the maximum activation across those L_2 nodes associated with a category A response. The response nodes in L_3 compete using lateral-inhibitory leaky-integrator dynamics, as before. This competition, in the absence of large amounts of fast-varying noise, will essentially pick the response with the largest input, thus selecting, as before, the category associated with the L_2 instance node with the largest noise-perturbed activation. The selection process therefore gives identical results, in terms of accuracy of response, to the simple Thurstonian model developed earlier and hence maintains its asymptotic equivalence with the Luce choice rule. The reaction time that elapses before a decision is made depends on two things: the number of instance representations available in L_2 and the strength of competition from alternative responses. As more instances become represented at L_2, the maximum value of the noisy activations creeps up, according to the maximum-value theorem, thus speeding highly practiced responses. To the extent that a given test stimulus falls near to instances of two different categories, the lateral inhibitory signals experienced by the response node that is eventually selected will be higher, thus delaying the response for difficult-to-make decisions, as required by the data.

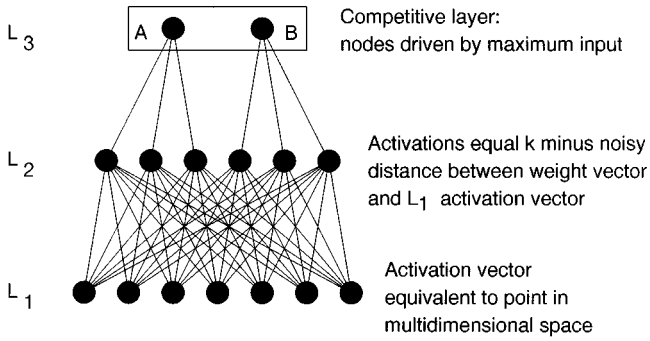

Figure 5. A module that displays power law speed-up with practice.

Does the reaction-time speed-up with practice exhibited by this model fit the observed power law? I have done many simulations, under a variety of conditions, all of which produced the pattern of results shown in the graph in Figure 6. The graph plots mean reaction time, taken over 1000 trials, against number of instances of each response category. As can be seen, the speed-up in reaction time with practice is fitted very well by a power function of practice, $A + BN^{-c}$. The fact that the time axis can be arbitrarily scaled, and the exponent of the power curve can be fitted by varying the signal-to-noise ratio on the input signals, bodes well for the chances of fitting this Thurstonian model to the practice data – this is the subject of current work. We know already that this signal-to-noise ratio also determines the accuracy with which the network responds, and the speed with which this accuracy itself improves with practice. While it is clear that it will be possible to fit either the accuracy performance or the RT performance with a given set of parameters, it remains to be seen whether a single set of parameters will suffice for fitting both simultaneously.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

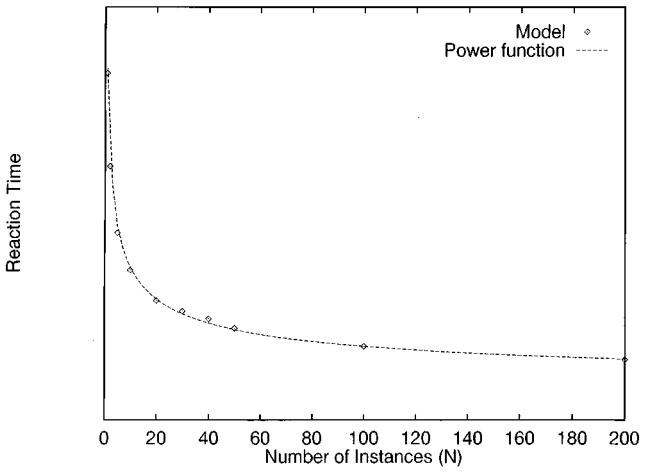

Figure 6. The model's performance compared with a power function ($A + BN^{-c}$). The time-axis can be scaled arbitrarily.

4.2.6. Summary. To summarize this section of the paper, I have illustrated how a simple, localist, lateral inhibitory network, fed by appropriate noisy bottom-up inputs delivered by some, presumably hierarchical, feature-extraction process, inherits considerable power from its close relationship to a number of classic mathematical models of behaviour. The network implements a Thurstonian choice-process which gives accuracy results that are indistinguishable (at asymptote) from those produced by application of the Luce choice rule to representations with exponential generalization gradients. It shows accuracy that improves with an increase in the number of training instances, equivalent in the Shepard-Luce-Nosofsky formulation to the uniform stretching of multidimensional space. With regard to reaction time, the network's RT distributions are very similar to those produced by Ratcliff's (1978) diffusion model (see Usher & McClelland 1995) and to those found in experimental data. The RTs reflect category similarities and are speeded as a power law with practice. This simple localist model thus provides a qualitative synthesis of a large range of data and offers considerable hope that this breadth of coverage will be maintained when quantitative fitting is attempted.

To this point I have not, at times, made a clear distinction between supervised and unsupervised learning. Underlying this conflation is the belief that the mechanisms underlying the two types of behaviour largely overlap – more particularly, unsupervised learning is a necessary component of supervised, or association, learning. This will, I hope, become more clear later. First I shall describe qualitatively how variants of the localist model discussed above can exhibit a number of behaviours more commonly associated with distributed models, and at least one behaviour that has proved difficult to model.

4.3. Generalization, attractor behaviour, "clean-up," and categorical perception

4.3.1. Generalization. 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. In a similar vein, such networks have been said to be robust, their performance worsening only gradually in the presence of noisy or incomplete input. Generalization and robustness are essentially the same thing: both refer to the networks' ability to deal with inputs that only partially match previous experience. One very common inference has been that networks using localist representations do not share these abilities. In this section I show that this inference is unjustified.

First, the previous sections have illustrated how one of the most resilient mathematical models of the stimulus-response generalization process can be cast in the form of a simple localist network. Put simply, in the face of a novel, or noisy, stimulus, the input signals to a layer of nodes whose weights encode patterns of activation encountered in previous experiences will reflect, in a graded fashion, the degree of similarity that the current input shares with each of those learned patterns. If the current pattern is not sufficiently similar to any learned pattern to evoke superthreshold input, then no generalization will be possible, but to the extent that similarities exist, the network can choose between competing classifications or responses on the basis developed above. It will be possible to vary the breadth of generalization that can be tolerated by varying the input threshold, θ. Thus if no L_2 node receives superthreshold input, yet generalization is required, the threshold can simply be dropped until input, upon which a response can be based, is forthcoming.

Of course, the stimulus-response model described above only generalizes in the sense that it generates the most appropriate of its stock of previously learned responses on presentation of an unfamiliar stimulus. This type of generalization will not always be appropriate: imagine a localist system for mapping orthography to phonology, in which each familiar word is represented by a node which alone activates sufficiently in the presence of that word to drive a representation of that word's phonology. Would this system exhibit generalization on presentation of a novel orthographic string (i.e., a nonword)? Only in the sense that it would output the phonology of the word that best matched the unfamiliar orthographic input. This is not the sort of generalization that human readers perform in these circumstances; they are content to generate novel phonemic output in response to novel orthographic input. The localist approach to simulating this latter ability relies on the fact that in quasiregular mappings, like that between orthography and phonology, in which both the input pattern (i.e., the letter string) and the output pattern (i.e., the phonemic string) are decomposable into parts, and in which each orthographic part has a corresponding phonemic part with which it normatively corresponds, the localist model can perform generalization by input decomposition and output assembly. Specifically, although the unfamiliar orthographic string cannot activate a localist representation of the complete nonword (since by definition no such representation exists), it can lead to activation in localist units representing orthographic subparts, such as onset cluster, rime, vowel, coda, and so on, and each of these can in turn activate that portion of the phonological output pattern with which it is most usually associated. This idea of generalization by input decomposition and output assembly for nonwords, supplemented by a dominant, but not exclusive direct route for known words is, of course, the strategy used by many localist modellers of single-word reading (Coltheart et al. 1993; Norris 1994a; Zorzi et al. 1998).

In nonregular domains, where generalization by decomposition and assembly is not possible, the tendency of localist models either to fail to generalize or, when appropri-

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

ate, to perform generalization to the best matching stock response, might be seen as a distinct advantage. (Many of the points that follow can also be found in Forster 1994, but I believe they bear repetition.) Take the mapping from orthography to semantics, or the mapping from faces to proper names: Is it appropriate to generalize when asked to name an unfamiliar face? Or when asked to give the meaning of a nonword? In a localist model of the general type developed above, the threshold for activating the localist representation of a known face or a known word can be set high enough such that no stock response is generated for such unfamiliar stimuli. When a stock response is required, such as to the question "Which familiar person does this unfamiliar person most resemble?", the input threshold might still be dropped, as described above, until a response is forthcoming. It is unclear whether distributed models of, for example, face naming or orthography-to-meaning mapping (particularly those with attractor networks employed to "clean up" their output) exhibit this sort of flexibility rather than attaching spurious names to unfamiliar faces, spurious meanings to nonwords, or spurious pronunciations to unpronounceable letter strings.

Are networks that automatically show generalization the most appropriate choice for implementing irregular mappings such as that between orthography and semantics? Forster (1994) suggests not, while McRae et al. (1997), in rejecting Forster's pessimism, note that "feedforward networks . . . can learn arbitrary mappings if provided with sufficient numbers of hidden units. Networks that are allowed to memorize a set of patterns sacrifice the ability to generalize, but this is irrelevant when the mapping between domains is arbitrary" (p. 101).

McRae et al., however, do not show that "sufficient numbers of hidden units" would be significantly less than one for each word (i.e., an easily learned localist solution); and, even so, it is not clear what advantages such a distributed mapping would exhibit when compared with a localist lexical route, given that generalization is specifically not required. With regard to Forster's questions about the spurious activation of meaning by nonwords, McRae et al.'s simulations used a Hopfield-type model, with restrictions on data collection allowing them to model the learning of only 84 orthography-to-semantic mappings. A test of the resulting network, using just ten nonwords, led to "few or no [semantic] features" being activated to criterion – whether Forster would be persuaded by this rather qualified result is doubtful. Plaut et al. (1996) identify essentially the same problem in discussing their semantic route to reading. Their unimplemented solution involves semantic representations that are "relatively sparse, meaning each word activates relatively few of the possible semantic features and each semantic feature participates in the meanings of a very small percentage of words" (p. 105), and they add "this means that semantic features would be almost completely inactive without specific evidence from the orthographic input that they should be active. Notice that the nature of this input must be very specific in order to prevent the semantic features of a word like CARE from being activated by the presentation of orthographically similar words like ARE, SCARE, CAR, and so forth" (p. 105).

Since the mapping between orthography and semantics clearly requires an intermediate layer of mapping nodes, it might seem easiest to ensure this exquisite sensitivity to orthographic input by making these mapping nodes localist lexical representations. Of course this would mean that the second route to reading was the type of localist lexical route the authors explicitly deny. It remains to be demonstrated that a genuinely distributed mapping could exhibit the requisite properties and, again, what advantages such a scheme would enjoy over the rather straightforward localist solution.

Finally, another type of generalization is possible with localist networks, namely, generalization by weighted interpolation. In such a scheme, localist representations of various familiar items activate to a level equal to some function of the degree that they match an unfamiliar input pattern, the combined output being an activation-weighted blend of the individual output patterns associated with each familiar item. This type of generalization is most appropriate in domains in which mappings are largely regular. A similar arrangement has been postulated, using evidence derived from extensive cell recording, for the mapping between activation of motor cortical neurons and arm movements in primates (Georgopoulos et al. 1988). Classifying this so-called population coding as a type of localist representation is perhaps stretching the notion farther than necessary (cf. earlier comments regarding orientation columns), although it really amounts to no more than acknowledging that each cell in a population will respond optimally to some (presumably familiar) direction, albeit one located in a space with continuously varying dimensions. In some cases it might even be difficult to distinguish between this weighted-output decoding of the pattern of activation across what I'll call the coding layer and an alternative decoding strategy that imagines the cells of the coding layer as a set of localist direction-nodes racing noisily to a criterion, with the winner alone driving the associated arm movement.[2] A similar distinction has been explored experimentally by Salzman and Newsome (1994), who located a group of cells in rhesus monkey MT cortex, each of which responded preferentially to a given direction of motion manifested by a proportion of dots in an otherwise randomly moving dot pattern. The monkeys were trained on a task that required them to detect the coherent motion within such dot patterns and to indicate the direction of motion by performing an eight-alternative forced-choice task. Once trained, the monkeys were presented with a pattern containing, for example, northerly movement while a group of cells with a preference for easterly movement was electrically stimulated to appropriate levels of activation. The responses of the monkeys indicated a tendency to respond with either a choice indicating north or one indicating east, rather than modally responding with a choice indicating the average direction northeast. The authors interpreted these results as being consistent with a winner-takes-all rather than a weighted-output decoding strategy. Implicit in this interpretation is the monkeys' use of a localist coding of movement direction. It is likely that both decoding strategies are used in different parts of the brain or, indeed, in different brains: The opposite result, implying a weighted output decoding strategy, has been found for those neurons in the leech brain that are sensitive to location of touch (Lewis & Kristan 1998). More germanely, generalization by weighted output can be seen in several localist models of human and animal cognition (e.g., Kruschke 1992; Pearce 1994).

To summarize, contrary to an often repeated but seldom justified assumption, there are (at least) three ways in which

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

localist models can generalize: by output of the most appropriate stock response; by input decomposition and output assembly; or by activation-weighted output.

4.3.2. Attractors. Another much-discussed feature of networks employing distributed representations is their ability to exhibit "attractor" behaviour. In its most general sense (the one I shall adopt here) attractor behaviour refers to the ability of a dynamic network to relax (i.e., be "attracted") into one of several stable states following initial perturbation. In many *content addressable memory* networks, such as that popularized by Hopfield (1982; 1984), the stable states of the network correspond to previously learned patterns. Such attractor networks are often used to "clean up" noisy or incomplete patterns (cf. generalization). In mathematical terms, a learning algorithm ensures that learned patterns lie at the minima of some function (the Lyapunov function) of the activations and weights of the network. The activation-update rule ensures that, from any starting point, the trajectory the network takes in activation space always involves a decrease in the value of the Lyapunov function (the network's "energy"), thus ensuring that eventually a stable (but perhaps local) minimum point will be reached. Cohen and Grossberg (1983) describe a general Lyapunov function for content-addressable memories of a given form, of which the Hopfield network is a special case.

To see how attractor behaviour (again in its general sense) can be exhibited by a variant of the localist network described above, assume that we have a two-layer network, as before, in which the upper layer, L_2, acts as a dynamic, competitive, winner-takes-all layer, classifying patterns at the lower layer, L_1. Let us further assume that L_2 nodes project activation to a third layer, L_3, the same size as L_1, via connections whose weights are the same as those of the corresponding L_1-to-L_2 connections (see Fig. 7). For simplicity, let us assume that the input threshold, θ, is zero. On presentation of an input pattern at L_1, the inputs to the L_2 nodes will reflect the similarities (e.g., dot products) of each of the stored weight vectors to this input pattern. If we track the trajectory of the activation pattern at L_3 as the competition for activation at L_2 proceeds, we will find that it starts as a low-magnitude amalgam of the learned weight vectors, each weighted by its similarity to the current input pattern, and ends by being colinear with one of the learned weight vectors, with arbitrary magnitude set by the activation of the winning node. In the nondeterministic case, the L_3 pattern will finish colinear with the weight vector associated with the L_2 node receiving the largest noisy input, I_j. Thus the L_3 activation vector is attracted to one of several stable points in weight-space, each of which represents one of the learned input patterns. In the noisy case, given the results presented earlier, the probability of falling into any given attractor will be describable in terms of a Luce choice rule. This is precisely the sort of attractor behaviour we require.

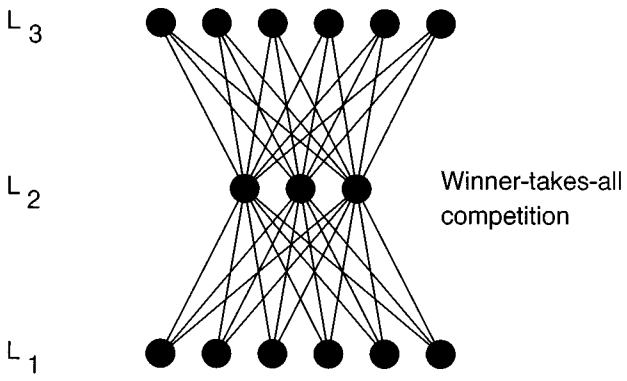

Figure 7. An attractor network.

In certain cases we might allow the L_2 nodes to project back down to the nodes in L_1 rather than to a third layer, L_3. In this case (reminiscent of the ART networks referred to earlier), the activation pattern at L_1 is attracted toward one of the stable, learned patterns. This network is essentially an autoassociation network with attractor dynamics. Such an implementation has some advantages over those autoassociative attractor networks used by Hopfield and others. For instance, it should be fairly clear that the capacity of the network, as extended by competing localist representations, is, in the deterministic case, equal to the maximum number of nodes available in L_2 to learn an input pattern. In contrast with the Hopfield network, the performance of the localist network is not hindered by the existence of mixture states, or false minima, that is, minima of the energy function that do not correspond to any learned pattern. Thus localist attractor networks are not necessarily the same as their fully distributed cousins, but they are attractor networks nonetheless: whether or not a network is an attractor network is independent of whether or not it is localist.

4.3.3. Categorical perception. Since one can view the lateral inhibitory module as performing a categorization of the L_1 activation pattern, the category being signalled by the identity of the winning L_2 node, the network can naturally model so-called *categorical perception* effects (see, e.g., Harnad 1987). Figure 8 illustrates the characteristic sharp category-response boundary that is produced when two representations, with linear generalization gradients, compete to classify a stimulus that moves between ideal examples of each category. In essence, the treatment is similar to that of Massaro (1987), who makes a distinction between categorical perception, and "categorical partitioning," whereby a decision process acts on a continuous (i.e., noncategorical) percept. This distinction mirrors the one between a linear choice rule acting on representations with

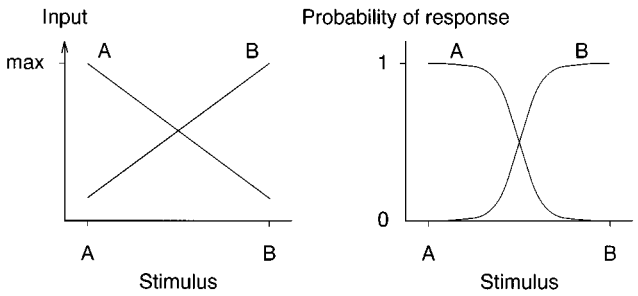

Figure 8. The left-hand graph shows the strength of input received by two L_2 nodes, A and B, as a stimulus moves between their learned patterns. The right-hand graph shows the probabilities of choosing either node when Gaussian noise is added to the input and the node with the largest resultant input chosen. The steepness of the crossover is determined by the signal-to-noise ratio.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

exponential generalization gradients and a Thurstonian choice-process acting on representations with linear generalization gradients, as seen above. The fact that Massaro describes this partitioning process in Thurstonian terms, yet models it using the Fuzzy Logic Model of Perception (Oden & Massaro 1978), serves to emphasize the strong mathematical similarities between the two approaches.

4.4. Age-of-acquisition effects

Finally, I will briefly mention a specific effect, which is potentially difficult to model in connectionist terms, namely, the effect of age of acquisition. It has been demonstrated that subjects in word-naming and lexical decision experiments respond faster to words learned earlier in life (e.g., Morrison & Ellis 1995). This is independent of any effect of word frequency, with which age of acquisition is strongly negatively correlated. The potential difficulty in accounting for this effect with a connectionist model concerns how age of acquisition might plausibly be represented. Various schemes have been suggested to model frequency effects but age of acquisition appears to present a much stiffer challenge, particularly for models that learn via an error-based learning rule and whose weights, therefore, tend to reflect the history of learning with a bias toward what has occurred most recently.

In suggesting a potential solution to this problem, I shall make three assumptions. First, word naming and lexical decision involve competitive processes. Second, word acquisition is also a competitive process. Third, there is some variation in the competitive capacity of nodes in a network and the relative competitive capacity of a given node endures over time. Taking the last of these assumptions first, it is perhaps not too fanciful to talk of a node's intrinsic ability to compete and that, given such a concept, it is uncontroversial to assume that there will be some variation in the competitive capacity of a set of nodes. Competitive capacity might be influenced by the physical location of a node relative to its potential competitors, the breadth or strength of its lateral inhibitory connections, its ability to sustain high activations, and the function relating those activations to an outgoing, lateral inhibitory signal. Given such influences, it is at least possible that certain of these aspects of the node's situation endure over time, so that, for instance, a node of high competitive capacity in one time period tends to have high competitive capacity in the next. In this context we wish the time over which relative competitive capacity endures to be of the order of tens of years.

The second assumption is that the process by which a word comes to be represented in long-term memory by a given node is a competitive one. To support this assumption it is necessary to suggest some possibilities regarding the learning of localist representations in general. Earlier in this paper it was assumed that whenever learning of an entity was required – for instance, when no committed node received superthreshold input on presentation of that entity – an uncommitted node would come to represent that entity. How might this be achieved? Let us assume that uncommitted nodes can respond to input, whether or not the magnitude of the input they receive is above the threshold θ; in other words, the threshold for uncommitted nodes is effectively zero, though the magnitude of their incoming weights may be small. Further assume that committed nodes that receive superthreshold input are able to quench significant activation at uncommitted nodes via lateral inhibitory connections. In cases where a pattern is presented under sufficiently high threshold conditions such that no committed node receives superthreshold input, numbers of uncommitted nodes will activate. If learning is presumed to be enabled under these circumstances, then let each of the uncommitted nodes adapt its incoming weights such that

$$\frac{dw_{ij}}{dt} = \lambda a_j(a_i - w_{ij}), \tag{10}$$

where λ is a learning rate, w_{ij} represents the weight from the ith L_1 node to the jth L_2 node, and a_i and a_j represent the corresponding node activations. This learning rule, almost the same as that given earlier but with the additional product term a_j, is the *instar learning rule* (Grossberg 1972), which simply states that the weights incoming to a given L_2 node will change so as to become more like the current L_1 activation pattern at a rate dependent on the activation of that L_2 node. Just as for the committed nodes, the uncommitted nodes will be subject to lateral inhibition from other uncommitted nodes, thus establishing a competition for activation, and hence a competition, via Equation 10, for representation of the current pattern. Those uncommitted nodes that, either by chance or thanks to some earlier learning, activate relatively strongly to a given pattern tend to change their incoming weights faster in response to that pattern, and thus accrue more activation – there is a positive feedback loop. Eventually, the connection weights to one of the L_2 nodes become strong enough so that that node is able to suppress activation at other uncommitted nodes. At this point, that node will be deemed to be committed to its pattern, and further learning at that node will effectively be prevented. The process by which an uncommitted node competes to represent a novel pattern might be accomplished in a single presentation of a pattern (high λ, fast learning) or several presentations (low λ, slow learning).

Two things are worth mentioning with regard to this learning procedure. One is that it is compatible with the generalization procedure described earlier. On a given test trial of, say, a category learning experiment, the network might have its threshold set low enough to allow committed nodes to receive input, permitting a best-guess response to be made. If the guess is later confirmed as correct, or, more important, when it is confirmed as incorrect, the threshold can be raised until no committed node receives superthreshold input, allowing a competition among previously uncommitted nodes to represent the current activation pattern, with that representation then becoming associated with the correct response. This is very similar to the ARTMAP network of Carpenter et al. (1991). The main difference in emphasis is in noting that it might be beneficial for the network to learn each new pattern (i.e., run as an exemplar model) even when its best-guess response proves correct. The second point worth mentioning is that the learning process suggested above will result in a number of nodes that come close to representing a given pattern yet ultimately fail to win the competition for representation. These nodes will be well placed to represent similar patterns in the future and may, for example, in single-cell recording studies, appear as large numbers of cells that seem to cluster (in terms of their preferred stimulus) around recently salient input patterns.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

The final assumption in this account of the age-of-acquisition effects is that word naming depends on a competitive process similar to that described above. It is absolutely in keeping with the modelling approach adopted here to assume that this is the case. Moreover, a number of recent models of the word-naming and lexical decision processes make similar assumptions regarding competition (see Grainger & Jacobs 1996, for a review and one such model).

Age-of-acquisition effects can now be seen to be a natural feature of any system that is consistent with these three assumptions. Those nodes, which have a high intrinsic competitive capacity, will tend to become committed to those words encountered early, since this process is a competitive one. If competitive capacity endures, then nodes that happen to represent words acquired early will have an advantage in subsequent competitions, all else being equal. If word naming and lexical decision engage competitive processes, words acquired early will tend to be processed faster than words acquired late, just as the age-of-acquisition effect demands. Note that nothing needs to be known about the distribution of competitive capacity for this account to be true. The only requirement is that there be significant variability in these nodes' competitive capacity that is consistent over time.

4.5. Supervised learning

The process of pattern compression and classification described so far is an unsupervised learning mechanism. This unsupervised process effectively partitions the space of input patterns into distinct regions on the basis of pattern similarities. By contrast, supervised learning involves the learning of pattern associations, this term extending to a wide variety of tasks including stimulus-response learning, pattern labelling, and binary (e.g., yes/no) or multiway decision making. Pattern association is, of course, the domain of application of the most common of the PDP networks, namely, the multilayer perceptron trained by backpropagation of error (henceforth abbreviated as BP network; Rumelhart et al. 1986). In the framework developed here, supervised learning is a simple extension of the unsupervised classification learning described previously. Essentially, once two patterns that are to be associated have been compressed sufficiently so that each is represented by the supercriterion activation of a single, high-level node, then the association of those two nodes can proceed by, for example, simple Hebbian learning. (Indeed, one might even view supervised learning as autoassociative learning of the amalgam of the compressed, to-be-associated patterns, permitting subsequent pattern-completing attractor behaviour.) Geometrically speaking, the classification process orthogonalizes each of the patterns of a pair with reference to the other patterns in the training set, the subsequent association between those patterns being trivially accomplished without interfering with previously acquired associations. The general scheme is shown in Figure 9 and is functionally almost identical to the ARTMAP network developed by Carpenter et al. (1991), as well as to many other networks (e.g., Hecht-Nielson 1987; Burton et al. 1990; McLaren 1993; Murre 1992; Murre et al. 1992). It is also functionally equivalent to a noisy version of the nearest-neighbour classification algorithm used in the machine-learning community, and structurally equivalent to more general psychological models including the category-learning models described in previous sections and other models such as the one proposed by Bower (1996).

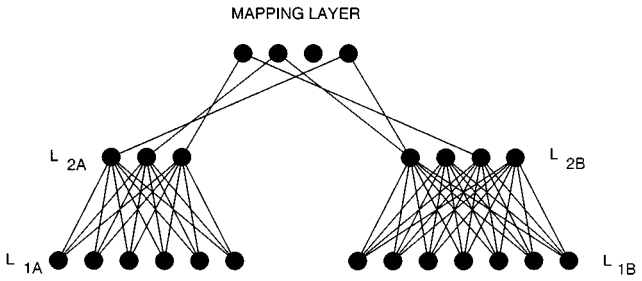

Figure 9. A generic network for supervised learning.

The operation of the network is simple. When the activation of a given node in one of the competitive layers, L_{2A}, hits a race-winning criterion, χ, it can excite one of the nodes in the mapping layer (see Fig. 9). (I assume that once a node in a given layer hits its criterion, other nodes in the layer are prevented from doing so by, for instance, a broadly applied inhibitory signal or raising of the criterion.) Assuming that a similar process occurs in the other competitive layer, L_{2B}, the active map node can then be associated by the L_{2B} winner by simple Hebbian learning. On subsequent presentation of one of the associates, driving, say, L_{2A}, a given classification node will reach criterion, and will activate its map-layer node, in turn activating the L_{2B} node corresponding to its associate. This would allow a relevant response to be made (perhaps by top-down projections from L_{2B} to L_{1B}). The division between the two halves of the network is appropriate when considering cross-modal associations, but will not be so clearly appropriate when associations are required between two items within a modality, for example, between two visually presented words. In this case, processes of selective attention might be employed, so as to classify one word and then the other; they will generally be competitors (cf. Kastner et al. 1998) and hence cannot both win a given race to criterion at the same time. The identity of the first word can be stored by sustained activation at the map layer, while attention is transferred to recognition of the second word. When recognition of the second word is accomplished, associative learning can proceed as before. Alternatively, one might propose a scheme whereby L_2 nodes responding to different objects currently present in the world might be permitted to coactivate to criterion (i.e., not compete; see the earlier discussion of binding), on the basis that they are grouped (or "streamed") separately, with subsequent association being achieved, as before, via the mapping layer.

The provision of the mapping nodes allows a good deal of flexibility in the associations that can be made. The mapping layer can be configured to allow one-to-one, many-to-one, or one-to-many mappings. Moreover, under certain circumstances, in particular when at least one of the associates is learned under low-vigilance (cf. prototype) conditions, remapping of items to alternative associates can be quickly achieved by rapid reconfiguration of connections to and from the mapping layer. The low-vigilance requirement simply acknowledges that flexible remapping of this kind will be difficult to achieve under conditions in which both "sides"

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

of a given set of associations are exhaustively exemplar coded, that is, when each association-learning trial engages two new exemplars (previously uncommitted nodes) linked via a newly established mapping-layer connection.

Such a scheme for supervised learning of pattern associations enjoys a number of advantages over alternative schemes employing distributed representations throughout, such as the BP network.

1. The learning rate can be set to whatever value is deemed appropriate; it can even be set so as to perform fast, one-pass learning of sets of pattern associations. The BP algorithm does not allow fast learning: learning must be incremental and iterative, with the learning rate set slow enough to avoid instabilities. The learning time for backpropagation thus scales very poorly with the size of the training set. By contrast, for the localist model, total learning time scales linearly with the size of any given training set, with subsequent piecewise additions to that training set posing no additional problem.

2. Localist supervised learning is an "on-line" process and is self-organizing, with the degree of learning modulated solely by the variation of global parameter settings for vigilance and learning rate. Typical applications of BP networks require off-line learning with distinct learning sets and separate learning and performance phases (see below).

3. The localist model is, in Grossberg's (1987) terms, both stable and plastic, whereas BP nets are not, exhibiting catastrophic interference in anything resembling "realistic" circumstances (see below).

4. Knowledge can be learned by the localist network in a piecemeal fashion. For instance, it can learn to recognize a particular face and, quite separately, a name, subsequently allowing a fast association to be made between the two when it transpires that they are aspects of the same person. BP nets do not enjoy this facility – they cannot begin the slow process of face-name association until both face and name are presented together.

5. The behaviour of localist nets is easy to explain and interpret. The role of the "hidden" units is essentially to orthogonalize the to-be-associated patterns, thus allowing enduring associations to be made between them. There is none of the murkiness that surrounds the role of hidden units in BP nets performing a particular mapping.

More important, these advantages are enjoyed without sacrificing the ability to perform complex mappings that are not linearly separable (e.g., XOR, see Fig. 10), or the ability to generalize (see earlier). The question arises as to why, given these advantages, there has been resistance to using localist models. This question will be addressed in the next section.

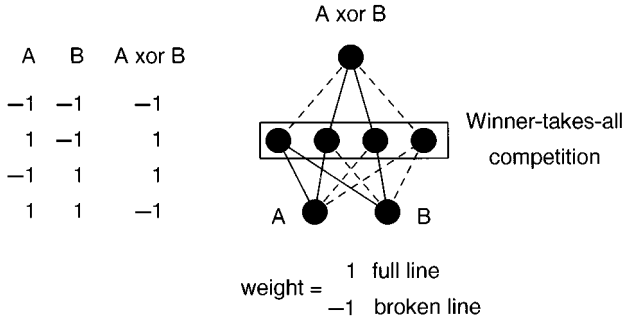

Figure 10. A network that performs the XOR mapping.

5. Some localist models in psychology

In extolling the virtues of localist connectionist models in psychology, I have occasionally encountered the belief that such models are not really connectionist models at all, this title being reserved for "real" connectionist models, such as those employing the backpropagation (BP) learning rule. Indeed, in some quarters it seems as if connectionist modelling and application of the backpropagation learning rule to fully distributed networks are seen as equivalent. I assume that this attitude stems from the great popularity of networks such as the BP network after the release of the PDP volumes with accompanying simulation software in the mid 1980s. Nevertheless, as mentioned earlier, several of the networks discussed in those volumes were localist. This suggests that bias against using localist models, or even against seeing them as connectionist at all, is not based solely on the wide availability of alternative approaches, but also on the assumption that localist models are less capable or less "plausible" than these alternatives. I do not believe either of these is well-founded.

Before addressing this issue further it is worth noting that many successful models in psychology are either localist connectionist models, or, in the light of the preceding discussion, can be readily implemented as such. I do not wish to (and could not) give a complete roll call of such models here, but in the areas in which I have a particular interest, these include Burton et al.'s (1990) model of face perception; Estes's (1986) array model of category learning and Estes's (1972) model of ordered recall (though not necessarily Lee & Estes's 1981 model later development of it); Morton's (1969) logogen model and its variants; Nosofsky's (1986) generalized category model and the mathematical equivalents described above; Kruschke's (1992) ALCOVE model of attentional category learning; Pearce's (1994) configural model of conditioning; Hintzmann's (1986) MINERVA model; models of speech production by Levelt (1989), Dell (1986; 1988), and Hartley and Houghton (1996); Norris's (1994a) model of reading aloud and his SHORTLIST model of spoken word segmentation (Norris 1994b); the DRC model of Coltheart et al. (1993); the TRACE model of word recognition (McClelland & Elman 1986); Usher and McClelland's (1995) model of the time course of perceptual choice; the models of immediate serial recall by Burgess and Hitch (1992; 1999) and Page and Norris (1998); other models of serial recall by Houghton (1990), Nigrin (1993), and Page (1993; 1994); Pickering's (1997) and Gluck and Myers's (1997) models of the hippocampus; Shastri and Ajjanagadde's (1993) model of reasoning; Hummel and Biedermann's (1992) model of object recognition and Hummel and Holyoak's (1997) model of analogy processing; Grainger and Jacobs's (1996) model of orthographic processing; Bower's (1996) model of implicit memory; and those models described in Grainger and Jacobs (1998). Furthermore, not only is the list of distinguished localist models a long one, but in cases where localist and fully distributed approaches have been directly compared with reference to their ability to explain data, the localist models have often proved superior (e.g., Coltheart et al. 1993; López et al. 1998).

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

I should state that not all of these models have stressed their equivalence with localist connectionist models. Indeed, it has become common, in the concluding sections of papers that describe localist models, to apologize for the lack of "distributedness" and to speculate that the same performance could be elicited from a more distributed model. In an attempt to account for this occasional reluctance, I will try, in the next section, to address some of the concerns most commonly voiced in relation to localist models.

First, however, it is worth pausing briefly to ask why some researchers have preferred a localist approach to modelling. I shall take as an example Jacobs and Grainger (1994; also Grainger & Jacobs 1996; 1998), who have been most explicit in their justification of a research programme based, in their case, on the localist interactive activation (IA) model (McClelland & Rumelhart 1981; Rumelhart & McClelland 1982). They see IA as a canonical model, a starting point representing "the simplest model within a given framework that fairly characterizes the qualitative behavior of other models that share its design and system principles with respect to the data at hand" (p. 519).

Despite the simplicity of the underlying model, they have been able to provide detailed simulations of accuracy and reaction time measures from a variety of orthographically driven tasks, contradicting earlier pessimism (e.g., McClelland 1993a) about whether reaction time measures would be susceptible to accurate simulation by such networks (Grainger & Jacobs 1996). They have further identified the IA model as particularly appropriate to a strategy of nested modelling in which, when the model is applied to a new set of data (in their case, data concerning aspects of orthographic processing in visual word recognition), it retains its ability to simulate data sets to which it was earlier applied. The flexibility of the IA model in this regard (as well as with regard to the modelling of functional overlap and scalability – Grainger & Jacobs 1996; 1998) is largely attributable to the technical advantages of localist modelling discussed in section 4.5, thus highlighting an important interaction between the choice of model type and the scientific methodology that is adopted in applying that model. As Jacobs and Grainger (1994) pointed out, not all developments in connectionist modelling have respected this constraint on backwards compatibility. For example, they cite the failure of Seidenberg and McClelland's (1989) model of reading to account explicitly for the word-superiority effect; simulating this effect had been a staple target of the previous generation of models in the area. Although it is possible that networks using thoroughgoing distributed representation could be shown to be capable of flexible, scalable, and nested modelling of functionally overlapping systems, this has not yet been so clearly demonstrated as it has been for the localist competitors to such models.

6. Why some people might be reluctant to use localist models in psychology

This section covers, in a little more detail, many of the issues raised (and countered) by Thorpe (1995) in relation to common arguments used against localist models.

6.1. "Distributed representation is a general principle"

Perhaps the most fundamental reason for choosing a fully distributed modelling approach over a localist one would be the belief that distributed representation is simply a general principle on which the enterprise of connectionist modelling is founded. Such a view was clearly stated by Seidenberg (1993) who gave the following as the first of his general connectionist principles

> Knowledge representations are distributed [distributed representations of orthography and phonology]. (p. 231)

(the bracketed comment refers to the way in which this principle was realized in the Seidenberg and McClelland [1989] model of reading). This enshrinement of distributed representations (assuming it is intended to imply a rejection of localist representation) is not only historically inaccurate – thoroughgoing distributed representation never having been a necessary feature of a connectionist model – but it is also rather ironic. The irony stems from the fact that in the later, improved version of the reading model (Plaut et al. 1996), orthography and phonology (though not the lexicon) were represented locally, as indicated previously.

6.2. "They don't generalize and/or are not efficient"

As noted above, the fact that fully distributed networks can generalize is sometimes taken to imply that localist networks cannot. I hope I have shown above that this is not the case. The wider issue of generalization is discussed in detail in Hinton et al. (1986), in the section entitled "Virtues of Distributed Representations." It is interesting to note that the introduction to this section states that "Several of these virtues are shared by certain local models, such as the interactive activation model of word perception, or McClelland's (1981) model of generalization and retrieval." The virtue of generalization is not confined to fully distributed models.

The chief virtue that Hinton et al. (1986) attribute to fully distributed networks, but deny to localist networks, is that of efficiency. They conclude that certain mappings can be achieved, using fully distributed networks, with far fewer hidden units than are used by the corresponding localist network. This is true and, in this restricted sense, the distributed networks are more efficient. The following three points are noteworthy, however. (1) This notion of efficiency will count for nothing if the process by which the mapping must be learned is not only inefficient but also rather implausible. This point relates both to the disadvantages of "distributed learning" raised above and to the later discussion of catastrophic interference. (2) The localist solution enjoys advantages over the distributed solution quite apart from its ability to perform the mapping. These relate to the comprehensibility of localist models and the manipulability of localist representations and will be discussed later. (3) More generally, efficiency in modelling, particularly when arbitrarily defined, is not necessarily an aim in itself. A lexicon of 100,000 words could be represented by the distinct states of a 17-bit binary vector – very efficient but not very plausible as a psychological model. In terms of modelling neural function, it is at least conceivable that the brain has arrived at computationally effective but representationally "inefficient" solutions to certain problems.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

6.3. "They do not degrade gracefully"

Another advantage often claimed for fully distributed networks is that they continue to perform well after damage, usually considered as loss of nodes or weights. This quality is sometimes termed "graceful degradation"; similar effects are usually tacitly assumed to occur in real brains in response to damage. By contrast, it is implied, localist models do not degrade gracefully, since loss of a given node will render its referent unrepresented. This is true, but only in a strictly limited sense. First, it should be repeated that localist models use distributed/featural representations at "lower" levels – the network will degrade gracefully in response to any loss of nodes at these levels, just as it is able to generalize to new or incomplete input. Second, localist models do not preclude redundancy. There may be many nodes that locally represent a given entity – indeed, in the exemplar models discussed above, this is very likely to be the case. Thus, loss of a given node will not necessarily leave its associated entity unrepresented (although in the model developed earlier, reaction time will increase and accuracy will diminish). By way of example (expanded slightly from Feldman 1988), suppose the brain has 10^{11} neurons and these are being lost at a rate of 10^5 per day; the chance of losing a given cell in a 70-year period is approximately 0.03. If we assume a small amount of redundancy in representation, say, five cells per entity, then the probability of leaving a given entity unrepresented in the same period is, assuming independence, 10^{-8}. I would guess that this is somewhat less than the probability of losing one's entire head in the same period; hence it would not seem an unreasonable risk. In this regard, it is important to note that a number of independent localist representations do not amount to a distributed representation.

It is worth asking whether humans ever seem to have lost their ability to represent highly specific entities (presumably via focal damage rather than by gradual wastage). Howard (1995) describes an aphasic patient who appears to have lost "specific lexical items from a phonological lexicon for speech production" (though see Lambon-Ralph, 1998, for an alternative view, albeit of a different patient). A particularly interesting feature of these data is that the naming accuracy for given words is "not demonstrably related to the availability of either their phonological or their semantic neighbours." While it is unwise to claim that this pattern of results could never be modelled with a fully distributed system, it is certainly more suggestive of a system based on locally represented lexical entries.

6.4. "There are not enough neurons in the brain and/or they are too noisy"

Any assertion to the effect that are too few neurons in the brain to permit localist representations presumes answers to two questions: How many neurons/functional units are there? How many are needed? Assuming that the answer to the second question does not err in requiring the brain locally to represent all possible percepts rather than some actual percepts (an error analogous to requiring a library to have sufficient capacity to store the vast number of possible books as opposed to the comparatively minuscule number of actual books), then perhaps speculating about insufficient capacity underestimates the answer to the first question. Most estimates put the number of cells in the brain at around 10^{11}. Mountcastle (1997) estimates that the number of cells in the neocortex alone is approximately 3×10^{10}. Even if one considers the number of cortical minicolumns rather than cells, the number is in the vicinity of 5×10^8. Similarly, Rolls (1989) cites a figure of 6×10^6 cells in area CA1 of the hippocampus, an area he proposes is responsible for the storage of episodic memories. [See BBS multiple book review of Rolls's "The Brain and Emotion" *BBS* 23(2) 2000.] These are large numbers and they seem to place the burden of proof on those who wish to claim that they are not large enough to allow successful local coding. Furthermore, proponents of a distributed approach would presumably have to allocate not just a node, but rather a whole attractor to each familiar item in memory. Since in most nonlocalist attractor networks the limit on the number of distinct attractor basins is smaller than the number of nodes, it is not clear what is gained in potential capacity by moving from a local to a distributed coding scheme.

With regard to the assertion that neurons (not to mention nodes) might be too noisy to allow small numbers of them to perform significant coding, I follow Thorpe (1995) in citing Newsome et al. (1989) and hence Britten et al. (1992), who measured the activity of relevant MT cortex neurons while a monkey performed a psychophysical discrimination task. They found that the "performance" of certain individual neurons, assessed by placing a discriminant threshold on their activity, was just as good as the performance of the monkey. In other words, the monkey had no more information than could be derived from the activity of single cells. Barlow (1995; and in his seminal paper of 1972) makes similar points and reviews other evidence regarding the sensitivity and reliability of single neurons.

6.5. "No one has ever found a grandmother cell"

The final complaint against localist representations, again taken from Thorpe (1995), concerns whether such representations have ever been found in real brains. I hardly need point out that the assertion in the heading is poorly worded, in that not having found a grandmother cell is not necessarily the same as not finding a localist representation, depending on how one chooses to define the former. Apart from this, the whole question of what would constitute evidence for, or more particularly against, localist representation seems to have become extremely confused. A review of the neuroscientific literature reveals that much of this confusion comes from poor use of terms and model nonspecificity. This review has necessarily been rather cursory, and space restrictions require even more cursory reporting in what follows.

6.5.1. Interpreting cell recordings. First, in one sense, the assertion in heading of section 6.5, even as worded, is not necessarily true. Figure 11 shows a finding of Young and Yamane (1993), who measured the responses of various cells in the anterior inferotemporal gyrus and the superior temporal polysensory area to images of the disembodied heads (!) of Japanese men in full face. The figure shows responses of one of the AIT cells which responded extraordinarily selectively to only one of the twenty faces. This was the only

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

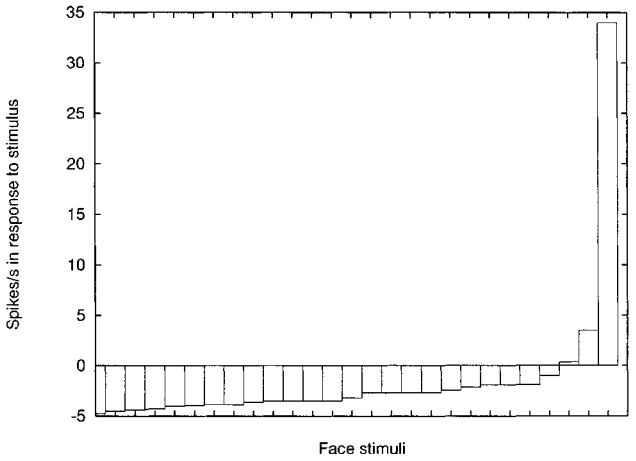

Figure 11. Data taken from Young and Yamane (1993), showing the response of a single cell in the inferotemporal cortex of a macaque monkey to a number of face stimuli. Spiking rates are measured relative to baseline response.

one of the 850 studied cells to respond in this highly selective manner. Nevertheless, the finding is interesting, since this cell is not just a localist representation, but apparently a grandmother cell (or rather a "particular-Japanese-man cell"). Young and Yamane state quite correctly that they cannot conclude that this cell responds to only one stimulus, since only a small number of stimuli (albeit with, to my eye, a high interstimulus similarity) were presented. But this proviso cannot conceal the fact that *no better evidence could have been found* in this experiment for the existence of at least one localist representation sufficiently tightly focussed to be termed a grandmother cell. Of course one might claim that better evidence for grandmother-cell representation in general would have been provided if all 850 cells had responded above baseline for one and only one of the faces. This is true, but such a finding would be enormously unlikely, even if each of the twenty individuals was represented in this extreme manner. Purely as an illustration, suppose that 100,000 cells in the relevant brain region were dedicated to face representation, with five extreme grandmother cells dedicated to each of the twenty stimulus subjects. This would imply a probability of one in a thousand of discovering such a cell on a given recording trial – approximately the probability with which such a cell was in fact found. I do not wish this illustration to be interpreted as indicating my belief in extreme grandmother-cell representation. That is not necessary to my more general defence of localist representation. I simply urge caution in the interpretation of cell-recording data.

The previous paragraph highlights one aspect of a more general problem with the antilocalist interpretations that have been put on some single-cell-recording studies. This involves misconceptions about what to expect if one measures cell responses in a localist network. There seems to be a widespread tendency to assume that if a number of cells activate for several hundred milliseconds following the presentation of any given stimulus, with different degrees of activation for different stimuli, then this speaks against the idea of a localist representation. It does nothing of the sort, although this fact is often obscured in passages such as the following:

> Even the most selective face cells discharge to a variety of individual faces and usually also discharge, although to a lesser degree, to other stimuli as well. Thus, faces are presumably coded in exactly the same way as everything else, namely, by the firing pattern of ensembles of cells with varying selectivity rather than of individual cells acting as complex feature detectors. (Gross 1992, p. 6)

> Neurons responsive to faces exhibited systematically graded responses with respect to the face stimuli. Hence each cell would systematically participate in the representation of many faces, which straightforwardly implies a population code. (Young & Yamane 1992, p. 1330)

Statements such as these are widespread and often used to argue against localist coding. What such arguments seem to miss, however, is the potential compatibility between distributed *processing* and localist *representation* discussed earlier. (They also often miss the compatibility between distributed representation at one level and localist representation at another, but I shall not dwell on that here.) Thinking back to the localist competitive network described earlier, a broad degree of activation (i.e., activation across a potentially large number of competing nodes, particularly if the input threshold, θ, is low) would be expected in response to any given stimulus, even if only one unit were eventually to reach criterion, χ, and/or win a competition for sustained activation. The broad pattern of activation would be different for different stimuli, just as described in the passages quoted above (and in the earlier discussion on sparse distributed representations). That grandmother cells (let alone localist representations) would be "signaling only one face and responding *randomly* to others" (Young & Yamane 1992, p. 1329, my emphasis) is not what would be expected on the basis of any workable localist model. In summary, even if we ignore the occasional tightly tuned cell, the finding of broadly distributed (though often transient) response to a stimulus does not rule out localist representation; indeed it is fully consistent with it.

A similar argument applies to the measurement of the informational content of particular neural firing responses performed by, for instance, Rolls et al. (1996). Among other things, they show that on presentation of a variety of stimuli, the response of a given neuron will convey a lot of information about the identity of an individual stimulus if its firing rate for that stimulus is unusually high or unusually low relative to its responses to the other stimuli. This is perfectly consistent with a localist coding. Suppose there exists a person-A node in the sort of localist network described earlier. Suppose one then presents eight persons to the network for identification, such that most of these persons share some of the features of person-A, only one (person-A herself) shares all of those features, and one person, say person-H, is unusual in sharing no features whatsoever with person-A (e.g., he looks nothing like person-A). On presentation of each of persons A – H, therefore, the person-A node will fire particularly strongly (supercriterion) to person-A, and particularly weakly to person-H, with intermediate responses to the other stimuli. Thus, the response to person-H will contain plenty of information (i.e., "this person looks nothing like person-A"), without any suggestion that the information it contains is of active benefit to the system in its identification task. In this situation it might also be found that the information contained in the firing of a given neuron is low when averaged across stimuli (as has been found experimentally), since this average is dominated by intermediate responses to many stimuli.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

An abiding problem has been that terms such as localist, distributed, population coding, ensemble coding, and so on, have been used without considering the range of models to which they might refer. This has led to interpreting data as supporting or refuting certain types of model without due consideration of the predictions of specific instantiations of each type of model. In many cases, researchers have concluded that some sort of "population coding" exists, but have failed to specify how such population coding operates so as to allow relevant tasks to be performed. For example, it is easy to hypothesize that colour and shape are each population coded, but how does this permit the learning of one response to a green triangle or a red square and another response to a red triangle or a green square, analogous to the classic XOR problem that is a staple of connectionist modelling? Again echoing an earlier point, how does one recall that it was a *yellow* Volkswagen that one witnessed speeding away from the scene of a bank raid? Simply positing population coding is not enough if there is no semipermanent way to tie the individual components of a percept together so as to form a unitized memory.

One answer to these questions by Rolls (1989) illustrates clearly one of the terminological problems that further confuse the literature. In describing the role of the hippocampus in episodic memory, Rolls describes a hierarchical system, culminating in area CA1 of the hippocampus:

> It is suggested that the CA1 cells, which receive these groups of simultaneously active ensembles, can detect the correlations of firing which represent episodic memory. The episodic memory in the CA3 cells would thus consist of groups of active cells, each representing one of the subcomponents of the episodic memory (including context), whereas the whole episodic memory would be represented not by its parts, but as a single collection of active cells, at the CA1 stage. (Rolls 1989, p. 299)

This conclusion is supported by a wealth of data and bolstered by references to the localist connectionist literature (e.g., Grossberg 1982; 1987). Yet when it comes to the paper's conclusion, we have the following:

> Information is represented in neuronal networks in the brain in a distributed manner in which the tuning of neurons is nevertheless not very coarse, as noted for the above hippocampal neurons. (Rolls 1989, p. 305)

This isolated statement is, on one level, true, but it completely deemphasizes the localist character of the conclusions reached throughout the paper. This target article is one attempt to clarify such issues of model taxonomy.

6.5.2. Evidence consistent with localist cortical coding. Positive evidence of localist coding of associations in the cortex, as opposed to hippocampal structures, has been provided by Miyashita and colleagues (e.g., Higuchi & Miyashita 1996; Miyashita 1993; Sakai & Miyashita 1991; Sakai et al. 1994). Sakai and Miyashita trained monkeys in a paired-association task, using twelve computer-generated paired patterns. They found task-related neurons in the anterior inferotemporal (AIT) cortex which responded strongly to one or the other of the patterns in a given associated pair but weakly to any of the twenty-two other patterns. This occurred in spite of the fact that at least some of other patterns were quite similar to one or another of the associated pair, with the two paired patterns being at least as distinct from each other as from the remainder of the set. In a further study, Higuchi and Miyashita showed that lesioning the entorhinal and perirhinal cortex caused a loss both in the knowledge of those associations learned prelesion, and in the postlesion ability of the monkeys to learn new paired associations. The lesion nonetheless had no effect on the response selectivity of cells in AIT cortex to single images from the 24-image set. The authors speculated that projections up to and back from the perirhinal and entorhinal cortex permitted associations to be learned between images which were already selectively represented b in AIT cortex (cf. Buckley & Gaffan 1998). This idea is strikingly similar to learning of associations between locally represented entities through projections to and from a map-layer (e.g., ARTMAP; Carpenter et al. 1991). It is also compatible with Booth and Rolls's (1998) recent discovery of both view-specific and view-invariant representations of familiar objects in the IT cortex and with the more general idea that simple tasks such as item recognition can be mediated by brain areas separate from the hippocampus (Aggleton & Brown 1999).

Perhaps the most crucial part of this series of experiments was carried out by Sakai et al. (1994), concerning what the authors called the fine-form tuning of each of the AIT neurons. They are unusual in making it clear that "one may mistakenly conclude that the most effective form in a screening test, is the optimum form for a recorded cell." In other words, finding that a cell responds most strongly to item D when tested with items A, B, C, and D does not imply that item D is the optimal stimulus for that cell, but only that it is the best of those tested. They circumvented this potential problem by using, in their visual-pattern pair-association task (as above), patterns generated from Fourier descriptors, such that continuous variation of a small number of parameters could generate continuous transformations of each member of the trained pattern set. These transformed patterns were always much closer in parameter space to the original patterns than the original, randomly parameterized patterns were to each other. For each recorded neuron the authors identified the original pattern (from a total of twenty-four on which the monkeys had been trained) that elicited the strongest response. Given the large degree of pattern variation in this screening set, and thus the relatively broad nature of the cell-selection process, there was no guarantee that each cell so selected would respond more strongly to its corresponding trained pattern than to fine-grained transformations of that pattern. Nonetheless, this was exactly what was found. In the majority of cases, presenting the transformed patterns resulted in a weaker response; in no case was the response to the transformed pattern stronger than that to the original learned pattern. This implies that the single-cell response is tuned to, or centred on, the particular visual pattern learned. Such a result is difficult to explain in terms of population coding unless one assumes that individual members of the active population of cells are tuned, by experience, to give a maximum response to a particular learned pattern – but such an account is not just similar to a localist account, it *is* a localist account. I should note that a similar experiment was performed by Amit et al. (1997) and although they report cell-recording results from a single cell that slightly increases its response to a degraded version of a previously trained visual pattern, they indicate that, on average, the IT cells from which recordings were elicited showed a decrease in response to degraded versions of the trained patterns, consistent with the results of Sakai et al. (1994).

7. What's wrong with using distributed representations throughout?

So far my emphasis has been on demonstrating the benefits of an approach to modelling that uses localist representations in addition to featural/distributed representations. This prolocalist, rather than antidistributed stance, has been quite deliberate. Nonetheless, it risks being interpreted as indicating equanimity in the selection of a modelling approach. To counter this interpretation I shall briefly outline some of the reasons the thoroughgoing distributed approach seems less promising. Owing to space limitations, I shall refer to other work for the detail of the some of the arguments. In referring to distributed modelling techniques, I shall take as my example the backprop (BP) network. This is perhaps unfair, for its deficiencies do not necessarily apply to all fully distributed approaches (for a brief discussion of a rather different class of networks, viz. Hopfield-type attractor networks, see sect. 4.3.2). Nevertheless, BP has been a dominant approach in connectionist modelling in psychology over the last decade, and is hence the most obvious candidate for an illustrative example. Before outlining my objections to BP and its relatives I should sound a note of caution. I do not intend my criticisms to be taken as an attempt to devalue the scientific contribution made by the body of work built around BP and distributed representations. In many cases, such as in the fields of reading and past-tense learning, the theorizing of PDP-style connectionists has stimulated considerable debate and has forced a reconsideration of long-held views, not least by elegantly demonstrating, via simulation, the falsity of certain (though not all) opposing claims. The fact that many of these debates are yet to be resolved is testament to the potency and value of the scientific challenge posed by this brand of eliminative connectionism.

7.1. The stability-plasticity dilemma, a.k.a. catastrophic interference

The stability-plasticity dilemma (Grossberg 1987) refers to the need for a learning system to be both stable, in the sense that it protects what it has learned from overwriting, and plastic, in that it remains capable of new learning. Grossberg (1987) offered principled reasons why the BP algorithm, unlike certain localist models, fails to solve the stability-plasticity dilemma. McCloskey and Cohen (1989) identified the same problem in a simulation of association learning and referred to the lack of stability as an exhibition of "catastrophic interference." Essentially the same phenomenon was noted by Ratcliff (1990). There has been a good deal of work on the subject since then (e.g., French 1991; 1994; Lewandowsky 1991; McRae & Hetherington 1993; Murre 1992; Sharkey & Sharkey 1995; Sloman & Rumelhart 1992), most of which has concluded that in order to reduce catastrophic interference one must reduce the overlap between the hidden-unit representations that intervene between particular associated pattern-pairs. This is, of course, exactly what is achieved by using localist representations as intermediates (for a review and a look-up model similar to that proposed here, see Sharkey & Sharkey 1995).

The problem of catastrophic interference occurs in backprop networks as a result of the gradient-descent learning procedure. At any point during learning the network weights are changing to follow a descending trajectory on an error surface in weight-space. The problem occurs because the shape of this error surface depends only on the patterns in the current learning set – indeed, the network can only move appropriately in weight-space by waiting until it has sampled each member of the current training set before making an "amalgamated" move. A consequence is that this error-reducing move does not take into account previously learned training sets. The only way it can do so is by accumulating training sets, so that new training sets are interleaved with all previous training sets. Learning in such networks is therefore "off-line" at two levels. First, any training set must be presented a large number of times, with small weight changes each time, for the relevant mapping to be stably learned. Second, to avoid overwriting, previous training sets must be interleaved with the current set.

Since the problem of catastrophic interference has been well described elsewhere (references above), I shall not describe it further. Rather, I would like to make some observations regarding a proposal advanced by McClelland et al. (1995) that has been taken by some as mitigating the problem of catastrophic interference with reference to brain function, and hence enhancing the plausibility of fully distributed modelling. Their proposal is that the hippocampus permits fast learning of pattern associations on-line, subsequently allowing these associated patterns to be replayed to a fully distributed neocortical learning system off-line, perhaps during sleep. The presentation of this hippocampally generated material to the neocortical system is effectively interleaved with patterns derived from continuing exposure to the environment and other patterns "reactivated" from among those already stored in neocortex. The neocortical system is supposed to be sufficiently slow-learning to avoid catastrophic interference under these conditions.

The idea of such memory consolidation has its roots in proposals by Marr (1970; 1971) and Squire et al. (1984). McClelland et al. add a computational flavour by suggesting that the dual-store system has evolved in this way so as to finesse the interference problems of distributed learning systems. There are several points to be made regarding this account.

1. For McClelland et al.'s proposal to be viable, the hippocampal system must be able to learn pattern associations on-line, with minimal interference. They achieve this by the "use of sparse, conjunctive coding in the hippocampus . . . [such that] representations of situations that differ only slightly may have relatively little overlap." In other words, in order to support a fully distributed system at the neocortex, they assume what is effectively a localist system in the hippocampus. This rather weakens arguments in principle against localist representations.

2. In their description of the function of the dual-store system, McClelland et al. tend to confound the idea and benefits of slow learning with those of slow, *interleaved* learning. Slow off-line consolidation of associations learned by a fast on-line system is appealing, regardless of whether what is learned in the fast system is interleaved with what is already present in the slow system. That is, the benefits of a dual-store system are quite independent of whether *interleaved* transfer is carried out from one to the other, as McClelland et al. propose. A dual-store system, with a fast system learning the individual, contextualized episodes and a slow system maintaining the more enduring, context-free

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

representations (analogous to the exemplar/prototype distinction described earlier), only demands interleaved learning if the slow system is prone to catastrophic interference.

3. Following from the previous point, one must be wary of a superficially tempting train of thought that runs like this: for a fully distributed neocortical system to avoid catastrophic interference, it must be supplemented by a fast, localist system; there exists a fast, localist system, embodied in the hippocampus; therefore the slow neocortical system is fully distributed. This logic is clearly erroneous, since the existence of a localist system in the hippocampus says nothing about whether the neocortical system is localist or fully distributed in nature. Both the fast and the slow systems might be localist, thus eliminating the problem of catastrophic interference in the neocortex without resorting to the complexities of interleaved learning.

4. Last, and perhaps most important, part of the putative mechanism of interleaved consolidation seems to be inadequate. McClelland et al. maintain that new patterns stored in hippocampus are potentially interleaved both with patterns encountered during continuous exposure to the environment and with other patterns previously learned by neocortex. The former (i.e., the items that are continuing to be represented environmentally) will presumably be hippocampally stored and consolidated anyway, so their interleaving can be accomplished either indirectly via the hippocampus or directly from the environment. The problem concerns the source, for interleaving purposes, of those old patterns which are no longer represented in hippocampus, but which are stored solely in neocortex (e.g., those patterns that are hypothesized to survive hippocampal damage in retrograde amnesia). The dilemma is this: How can those patterns associations *stored* in neocortex be used to *train* neocortex? There is a basic problem here: an error-based learning system, such as the one proposed, cannot teach itself. This would be rather like asking someone to mark his or her own homework. First, if the neocortical system is imagined as a trained BP network (or similar), it is unclear how one can extract from the network a representative sample of the input patterns on which it was trained, so that these might be interleaved during training with the new hippocampal patterns. Second, even if one could generate the relevant input patterns, it is unclear how the network could then, given one of those input patterns, generate both an output pattern and a different target pattern, as is required for gradient-descent learning. As these two patterns, if they are both to be generated by the neocortical system, will be the same, there will never be any error term to back-propagate and hence no learning. The old neocortical patterns will remain effectively unconsolidated and hence vulnerable to catastrophic interference.

The only way out of this problem seems to be to find some way of accurately sampling the neocortical store prior to any perturbing inputs from the hippocampus so as to generate a training set of input and target patterns that can be (quickly) stored in some other system and appropriately used for slow, interleaved learning in the neocortex. Such a scheme has not yet been proposed, although Robins (1995) and French (1997a) have suggested similar schemes, whereby a smaller but somehow representative set of pseudopatterns is loaded back from the slow system to the fast system (i.e., presumably, the hippocampus) so that the neocortical training set comprises a hippocampally generated mixture of these pseudopatterns with recently acquired patterns. Disregarding the fact that such schemes seem to provide less than solid protection to old memories (with 20 to 30 percent loss after only 10 to 20 new intervening patterns, often using more pseudopatterns than original pattern pairs), they also seem to imply that all knowledge, old or new, must be effectively located in a fast-learning system (the hippocampus?), with the older knowledge also stored neocortically. Although this could be construed as consistent with evidence from animals and humans with hippocampal damage, it is not consistent with recent data from Graham and Hodges (1997) and Snowden et al. (1996), who show preserved recent memories and impaired distant memories in patients with semantic dementia who have relative sparing of the hippocampal complex.

The previous paragraph illustrates some difficulties in endogenously generating *from* the neocortex patterns for interleaving with hippocampally generated patterns during consolidation *to* the neocortex. If these criticisms are accepted, avoiding catastrophic interference will depend strongly on the assumption that exogenously generated patterns (more particularly, pattern pairs, encountered during ongoing exposure to the environment) will be representative of the older contents of the neocortex. Note that for a localist neocortical system, or indeed for any neocortical system not prone to catastrophic interference, this constraint on the stability of the environment is not required. Hence in McClelland et al.'s approach a fully distributed neocortex demands that the environment be largely stable and the learning rate be very slow. Having arrived at this conclusion one is tempted to ask: why bother, under these conditions, with consolidation from hippocampus to neocortex at all? Evidence for consolidation is an important component of the data evinced by McClelland et al. in support of their model, but, under conditions of a stable environment and a slow-learning neocortex, it is not clear what role consolidation plays. For example, if it is to hasten the incorporation of new knowledge into the neocortex, this will reduce the chances of old knowledge being resampled from the environment during the period over which this incorporation takes place, thus increasing the chances of interference.

7.2. Implausibility of the learning rule

Even if one were to disregard the associated problems of catastrophic interference and interleaved off-line learning, there are still considerable doubts about the validity of the BP learning rule as a brain mechanism. These doubts are readily acknowledged even by those most associated with the use of this technique, and this can lead to some rather curious conclusions:

> As an example, they focus on the back-propagation learning algorithm . . . pointing out that it is very implausible as a model of real learning in the brain. . . . This is, of course, true. . . . But even this glass is a quarter full: in many cases . . . one is not interested in modelling learning per se, and the so-called learning algorithm is used to set the weights in the network so that it will perform the tasks of interest. The term "learning" has irrelevant psychological connotations in these cases and it might be less confusing to call such algorithms "weight setting algorithms." Unless there is some systematic relationship between the way the necessary weights are found and the aspects of model performance under study, which in general we have no reason to expect, it is harmless to use unrealistic learning algorithms. (Farah 1994a, p. 96)

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

Skipping over the fact that the whole localist-distributed debate gives us every reason to expect a systematic relationship between the means of learning and the subsequent performance of the system, it seems that, for Farah at least, one of the major advantages of connectionism over more traditional models – that they could provide some account of how certain mappings are learned by example – is irrelevant. Given this quote, would it be legitimate to use connectionist networks as psychological models even if it could be proved that the weighted connections in those networks could *never* have been acquired by a process consistent with the structure and function of the human brain? And would "lesioning" such networks and comparing their subsequent performance with that of a brain-injured patient still be considered appropriate, as it is now?

The question regarding the correspondence of network models and brain function is a difficult one. It is of course perfectly justified to use networks as cognitive models, disclaiming any necessary connection with actual brains. Having done so, however, one is less justified to use those same networks in simulations of brain damage or in studies involving functional brain imaging. Throughout most of this article, I have been happy to propose localist models as cognitive models; in the latter sections I hope to have conveyed some optimism that they might also be appropriate as models of brain function. The convergence of cognitive and neural models is, I suggest, a good thing, and seems more likely to emerge from a localist modelling approach than any current alternative.

7.3. The dispersion problem

I referred earlier to the so-called "dispersion problem," identified by Plaut et al. (1996) as the problem underlying the poor performance of the Seidenberg and McClelland (1989) model when applied to nonword reading. Their basic observation was that in Seidenberg and McClelland's distributed representation of, say, orthographic onset cluster, the fact that an L is present, with the same associated pronunciation, in the words "Log," "Glad," and "Split" is utterly concealed by the representational scheme used for orthography (in this case so-called Wickelfeatures). As noted earlier, Plaut et al.'s solution was to adopt a completely localist representation of input orthography and output phonology. Is dispersion a general problem with distributed representations? Moving up a hierarchical level, as it were, will we see a similar problem when we wish to represent sentences in relation to their constituent words? Suppose we have a scheme for representing the sentence "John loves Mary" in which neither "John" nor "Mary" nor "loves" are locally represented. Will the similarity between this sentence and the sentences "Mary loves John" or "John loves Ann," or the novel sentence "John is celebrating his 30th birthday today," be concealed by such a representational scheme? It is clearly difficult to answer this question for all such distributed schemes, but the related issues of systematicity and compositionality in connectionist models are identified by Fodor and Pylyshyn (1988) as being of general concern. While local representation on its own is not sufficient to address the problems raised by Fodor and Pylyshyn, the addition of means for dynamic binding and inference (see, e.g., Shastri & Ajjanagadde 1993) might come closer to providing a satisfactory solution.

7.4. Problems deciding "when" and "what"

One problem with fully distributed networks which often gets overlooked in isolated simulations concerns the nature of the decision at the network's output. Let us take as an example a three-layer BP net trained on the mapping between distributed patterns representing faces and others representing their identities. The problem is twofold: how does the network indicate when it has identified a given test face? And how does the network indicate the identity itself? Suppose we activate the input pattern corresponding to a given face. Once activation has percolated through the network, a distributed pattern will be present at the output. Recognition might be signalled the moment this pattern is instated, but, if so, what would be the criterion by which the arrival of an output pattern might be judged? Alternatively, there might be some clean-up process acting on the output pattern which will take time to settle into an equilibrium state. But in this case, how will the network signal that it has arrived at this state? One might speculate that there is some process overseeing the clean-up system which monitors its "energy" and signals when this energy reaches a stable minimum. But what might be the locus of this energy-monitoring system and how might it function? (For a computational implementation of a similar scheme based on settling times, see Plaut et al. 1996.) Even supposing that the system knows *when* a stable state has been reached, how does it surmise *which* state has been reached? It cannot "look at" the states of each of the output nodes individually, since by definition these do not unambiguously identify the referent of the overall pattern. Thus the identification system must consider the states of all the nodes simultaneously and must generate that identity which is maximally consistent with the current output pattern. But such a system is most obviously implemented using just the sort of localist decision-process described earlier. Indeed, Amit (1989) has identified just this kind of localist "read-out" node as an essential adjunct to the distributed attractor networks with which he has been most concerned (Amit 1995, pp. 38–43). Advocates of fully distributed models might claim that all possible actions based on the identity implied by a given output pattern can simply be triggered by that output pattern via subsequent fully distributed networks. I cannot categorically deny this claim, though it seems rather unlikely to prove feasible in general.

This problem is often obscured in actual simulations using distributed systems because the identification process is done by the modeller rather than by the model. A typical approach is to take the distributed output pattern and calculate which of the learned patterns it best matches, sometimes adding a Luce choice-process for good measure. It would be preferable to have this functionality built into the network rather than run as an off-line algorithm. I am not claiming that fully distributed systems *cannot* incorporate such functionality but I have yet to see a specific system that has successfully done so.

7.5. Problems of manipulation

On a related note, it sometimes proves difficult to manipulate distributed representations in the same way that one can manipulate localist representations. As an example, in most models of immediate serial recall (e.g., Burgess & Hitch 1992; 1999; Page & Norris 1998) it proves necessary

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

to suppress the recall of items that have already been recalled. If the items are locally represented, then this can easily be achieved by suppressing the activation of the relevant node. If the items are represented in a distributed fashion, however, such that the representations of different items overlap, it is difficult to suppress one item without partially suppressing others.

7.6. Problems of interpretation

Fully distributed networks are much more difficult to interpret than their localist counterparts. It is often hard to explain how a distributed network performs a given mapping task. This is not necessarily a problem for the model qua simulation, but it is a distinct problem for the model qua explanatory theory. Unfortunately, space does not permit a consideration of this point here but excellent discussions can be found in Forster (1994), Green (1998 and subsequent correspondence), Jacobs and Grainger (1994), Massaro (1988), McCloskey (1991), Ramsey (1997), and Seidenberg (1993).

8. Conclusion

This target article has sought to clarify the differences between localist and fully distributed models. It has emphasized how the difference lies not in their use of distributed representations, which occur in both types of model, but in the additional use of local representations, which are only used by the localist. It has been shown, in general, how localist models might be applied in a variety of domains, noting their close relationship with some classic models of choice behaviour, stimulus generalization, pattern classification, choice reaction-time, and power law speed-up with practice. We have discussed how localist models can exhibit generalization, attractor behaviour, categorical "perception," and effects of age of acquisition. Supervised learning of pattern associations via localist representations was (re)shown to be self-organizing, stable, and plastic.

We have considered a number of powerful cognitive models that are either implemented as, or are transparently implementable with, localist networks, with an attempt to defuse some of the more common criticisms of such networks. Some of the relevant neuroscientific data have been surveyed along with areas in which localist models have been rejected apparently without good cause. Some neuroscientific data supportive of a localist approach have been reviewed, along with some of the reasons a fully distributed modelling stance may be less promising than the localist alternatives, catastrophic interference being the most serious among several enduring problems for the fully distributed approach.

The conclusion is that localist networks are far from being implausible: They are powerful, flexible, implementable, and comprehensible, as well as being indicated in at least some parts of the brain. By contrast, the fully distributed networks most often used by the PDP community underperform in some domains, necessitate complex and implausible learning rules, demand rather baroque learning dynamics, and encourage opacity in modelling. One might even say that if the brain does not use localist representations then evolution has missed an excellent trick.

ACKNOWLEDGMENTS

I would like to thank Dennis Norris, Ian Nimmo-Smith, John Duncan, Rik Henson, Gareth Gaskell, and Andy Young for many useful discussions and for their help in the preparation of this manuscript. I would also like to thank D. Amit, H. Barlow, M. Coltheart, J. Feldman, R. French, J. Murre, P. Thagard, X. Wu, and other, anonymous reviewers for their thoughtful comments.

All correspondence and requests for reprints should be sent to Mike Page, MRC Cognition and Brain Sciences Unit, 15 Chaucer Rd., Cambridge, CB2 2EF, U.K. (mike.page@mrc-cbu.cam.ac.uk)

NOTES

1. See Bundesen (1993) for a review of independent race models that have similar properties with regard to the Luce choice rule. This review came to my attention too late in the preparation of this target article to allow proper discussion within.

2. Indeed, it is possible to cast the optimal, Bayesian approach to the decoding of activation patterns on the coding layer, as discussed and preferred (relative to the weighted vector method described above) by Oram et al. (1998), in the form of a localist classifier of the type discussed earlier. The link relies on formal similarities between the Luce-Shepard choice rule and Bayes's rule as applied to conditional probabilities expressed as exponential functions. Decoding would comprise a winner-take-all competition over a layer of cells, themselves responding to and classifying the patterns of activation found in the coding layer. Because each of the cells in the classification layer would respond best to a given pattern of activation over the coding layer (itself corresponding to a given directional stimulus), and less strongly to more distant patterns, activations in the classification layer would themselves appear to comprise another distributed coding of motion direction, despite being decodable (to give the maximally likely stimulus-direction) by a simple localist competitive process.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Localist representation can improve efficiency for detection and counting

Horace Barlow[a] and Anthony Gardner-Medwin[b]

[a]*Physiological Laboratory, Cambridge CB2 3EG, England;* [b]*Department of Physiology, University College London, London WC1E 6BT, England.*

hbb10@cam.ac.uk **a.gardner-medwin@ucl.ac.uk**

Abstract: Almost all representations have both distributed and localist aspects, depending upon what properties of the data are being considered. With noisy data, features represented in a localist way can be detected very efficiently, and in binary representations they can be counted more efficiently than those represented in a distributed way. Brains operate in noisy environments, so the localist representation of behaviourally important events is advantageous, and fits what has been found experimentally. Distributed representations require more neurons to perform as efficiently, but they do have greater versatility.

In addition to the merits Page argues for, localist representations have quantitative advantages that he does not bring out. The brain operates in an uncertain world where important signals are always

liable to be contaminated and masked by unwanted ones, so it is important to consider how external noise from the environment affects the reliability and effectiveness of different forms of representation. In what follows, we shall adopt Page's definitions of localist and distributed representation, according to which almost any scheme or model has both components. In a scheme emphasising localist representation, the elements can nonetheless be used in combinations, and such combinations represent in a distributed way whatever input events cause them to occur. Similarly in a scheme emphasising distributed representation, each particular element is activated by a particular subset of the possible input patterns, and it represents this subset in a localist way; for example, a single bit in the ASCII code is a localist representation of the somewhat arbitrary collection of ASCII characters for which it is ON. We shall also assume for simplicity that the brain represents noisy data, but does not necessarily add noise; of course this is a simplification, but it is the appropriate starting point for the present problem.

Localist representations and matched filters. The principle of a matched filter is to collect *all* the signal caused by the target that is to be detected, and *only* this signal, excluding as much as possible signals caused by other stimuli. In this way the response from the target is maximised while pollution by noise from nontarget stimuli is minimised, yielding the best possible signal/noise ratio. Localist representations of features or patterns in the input data can be close approximations to matched filters. If the representation's elements are linear and use continuous variables, their outputs will be the weighted sums of their different inputs. If each weight is proportional to the ratio of signal amplitude to noise variance for that part of the input when the desired target is presented, the element will be a matched filter for that target.

Some neurons in sensory areas of the cortex follow this prescription well, and it makes good sense to regard them as members of a vast array of matched filters, each with slightly different parameters for its trigger feature or optimal stimulus. In V5 or MT (an area specialising in coherent motion over small regions of the visual field) the receptive fields of the neurons differ from each other in position, size, direction, and velocity of their preferred motion (Felleman & Kaass 1984; Maunsell & Van Essen 1983a; Raiguel et al. 1995), and preferred depth or disparity of the stimulus (DeAngelis et al. 1998; Maunsell & Van Essen 1983b). It has been shown that many individual neurons can detect coherent motion with as great sensitivity as the entire conscious monkey (Britten et al. 1992; Newsome et al. 1989). Furthermore, human performance in similar tasks varies with stimulus parameters (area, duration, dot density, etc.) as if it was limited by the noise or uncertainty inherent in the stochastic stimuli that are used, so external noise appears to be an important limit (Barlow & Tripathy 1997). On re-examination it also turns out that external noise is important in monkey MT neurons (Mogi & Barlow 1998). For neurons to perform as well as they do, they must have properties close to those of optimum matched filters, which suggests that the whole visual cortex is a localist representation of the visual field using numerous different arrays of filters matched to different classes of feature. This insight may well apply to all sensory areas of the cortex and even to nonsensory parts, in which case the cortex would be a strongly localist representation throughout.

Can efficient detection at higher levels always be done by the weighted combination of inputs from the elements of distributed representations at lower levels? This would require graded signals between the levels, and it is doubtful if those passing between cortical neurons have sufficient dynamic range. With binary signals, and a task of counting occurrences rather than extracting signals from noise, there is an analogous problem of diminishing the effects of overlap in distributed representations.

Counting accuracy in localist and distributed representation. For many of the computations that are important in the brain, such as learning, or detecting that two stimuli are associated, it is necessary to count or estimate how often a specific type of event has occurred. It is easy to see that, because the elements active in the distributed representation of an event that is to be counted also respond to other events, the mean response rates of those elements will be greater than the mean responses due solely to the event to be counted. The average effect of this inflation can readily be allowed for, but in a noisy environment the variance as well as the mean will be increased, and this cannot be corrected. The only way to avoid this problem completely would be to have localist representations for the counted events, though as shown elsewhere (Gardner-Medwin & Barlow, submitted), distributed representations can be efficient at counting if they employ enough elements with sufficient redundancy.

It may be suggested that brains often learn from a single experience and do not need to count accurately, but such an argument would be misleading. Efficient statistics are what an animal needs in order to make correct inferences with the minimum amount of data collection, and this is more, not less, important when the number of available trials is low. A system cannot use inefficient methods of representation if one-shot learning is to occur reliably when it is appropriate and not when it is not.

The relative merits of localist and distributed representations are sometimes finely balanced and are discussed in greater detail elsewhere (Gardner-Medwin & Barlow, submitted). Localist representations have the edge in terms of efficiency, but one must know in advance what needs to be detected and counted, so they are mainly appropriate for frequent, regularly recurring features of the environment. In spite of the large numbers of neurons required, the ability of distributed representations to handle unexpected and unprespecified events without ambiguity makes them better for handling novel experiences.

The principle of local computation. Finally it should be pointed out that the merit of localist representations stems from the fact that computation in the brain is done by local biophysical processes. Every element of a computation requires a locus in the brain where all the necessary factors are collected together so that they can take part in the biophysical process. As an example of the relevance of this principle, consider the Hebbian assumption about the locus of learning. Biophysical processes close to a synapse can readily be influenced by both pre- and postsynaptic activity, since the required information is present there in the way that the principle requires, but it would not be reasonable to assume that distributed patterns of synchronous activity in remote neurons could work in the same way. The implied ban on "action at a distance" may eventually need qualification through better understanding of neuromodulators and dendritic interactions, but localist representations have the advantage that they already collect at one element all the information required for detection and counting; this is what makes it possible for them to perform these jobs efficiently.

Page ends his manifesto by saying "if the brain does not use localist representations then evolution has missed an excellent trick." Plenty of neurophysiological evidence shows that it has not, in fact, missed this trick that is so valuable for achieving sensitive and reliable detection of weak signals in a noisy background, and for the frequency estimations needed for reliable and efficient learning. Doubtless evolution has also exploited the advantages that distributed representation can bring to the handling of the unexpected.

Neurons amongst the symbols?

C. Philip Beaman

Department of Psychology, University of Reading, Reading, RG6 6AL, United Kingdom. **c.p.beaman@reading.ac.uk**
www.rdg.ac.uk/AcaDepts/sx/Psych/People/beaman.html

Abstract: Page's target article presents an argument for the use of localist, connectionist models in future psychological theorising. The "manifesto" marshalls a set of arguments in favour of localist connectionism and against distributed connectionism, but in doing so misses a larger argument concerning the level of psychological explanation that is appropriate to a given domain.

The stance taken by Page in arguing for "localist" representations in psychological modeling has much to recommend it. Page notes that many of the models described in the PDP (Parallel *Distributed* Processing) books are localist in implementation. Indeed, prior to the publication of the PDP volumes there were a great many "localist" psychological theories, although theorists would perhaps have described themselves as "symbolic" rather than localist (e.g., Newell 1980).

"Localist" is defined by Page as a model in which one node responds maximally to a (learned) example of that type. If the quasibiological adherence to the "node" terminology is relaxed (sect. 2.1, para. 1), one could map Page's terminology readily enough onto an example of a production rule responding to a particular condition. On this basis it appears Page is not as isolated in his opinion as he seems to think. Localist is also a term that can be defined by opposition to "distributed" theories, yet Page explicitly states that "localist models almost always use both localist and distributed representations" (sect. 2.6, para. 1). The crux of the argument then is that *fully* distributed models, in which each representational unit responds to multiple instances of a particular type, are not as effective and/or useful as localist models in psychological theorising. A major part of the target article is spent defending the use of localist representations in connectionist network against imagined attacks of those committed to fully distributed representations. However, many of the criticisms Page anticipates are neatly circumvented by his original definition that localist models can use distributed representations. Page thus buys for himself many of the advantages claimed by advocates of the distributed representational approach while also retaining the advantages of localist, symbolic representations. The real question therefore seems to be, not "What are the advantages of this approach compared to its competitors?" but, on viewing the situation overall, does a choice have to be made?

Page's argument is that there is power in diversity. Within the same model there is space at one level for localist representations, and at another level for distributed representations. This is the "excellent trick" that evolution has not missed. However, the arguments are not entirely consistent. Localist representations are put forward as computationally useful devices with the implication that evolution is unlikely to have missed such efficient "tricks" (sect. 8, para. 3) yet at the same time, it is suggested that the brain employs some representationally inefficient solutions (sect. 6.2, para. 2). Although it is possible to reconcile these two arguments (e.g., by assuming that the representational inefficiency is more than made up for by the ease with which the representations can be employed), the apparent inconsistency is noteworthy since it mirrors a more fundamental inconsistency in the manifesto. The fundamental inconsistency is that the power in diversity argument regarding representations in individual connectionist models can also be applied to psychological theorising as a whole. There is no need at the present time to make the decision between localist and fully distributed models, and to make the attempt may be precipitous. Within the computational modeling there is space for fully distributed models and localist models, the trick is to spot which type of model is most appropriate to which situation.

It may be no accident that much of Page's previous research was concerned with models of immediate serial recall that are well-served by architectures in which a local representation of an "item" is an appropriate atomic level (Page & Norris 1998) and manipulations of the item occur. It is instructive to examine this situation in some detail, as it provides an example of circumstances in which prior commitment to localist representations becomes counterproductive.

To model effects of confusion at the phonological level, a more fine-grained level than coded for by localist item nodes in the Page and Norris "primacy" model, Page and Norris found it necessary to posit a two-stage model. In the two-stage model localist item representations are transferred to a further stage in which confusion over similar items can occur (Fig. 1). The second stage tags onto the model – a breakdown of the localist item representations

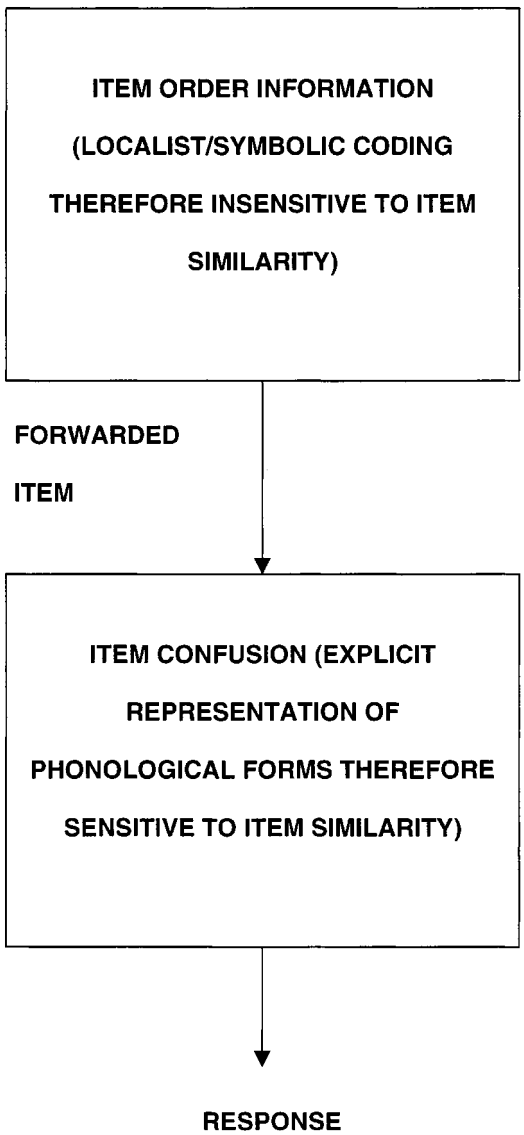

Figure 1 (Beaman). The Page and Norris (1998) account of the phonological similarity effect. Initial choice of localist "item" representations requires an extra stage to be added to the model in which phonological similarity is computed. Phonological confusion follows naturally from distributed phonological representations.

into their constituent phonological parts. This stage is added because the experimental data demand that such a stage must exist if one starts with localist item coding: it is not a natural consequence of the model. If a more distributed approach were taken, one might expect phonological confusions to be an emergent property of the model, as for example shown in Burgess and Hitch (1999). This is not to deny that the Burgess and Hitch model is not also "localist." The point is simply that by breaking down the representations into lower levels initially, the behavior required of the model emerges as a natural consequence of the representations chosen, rather than because of post hoc modifications. The difficulty lies in identifying where the behavior required of a model is contingent upon a certain form of representation, and where issues of representation are incidental.

The central message of the target article – that localist models be taken seriously – is a tenable one. By the same token, however, it is also necessary to take other fully distributed and other computational approaches, not addressed here, into account. It seems

likely that many representational tricks have been exploited by evolution, some as yet undiscovered and some (perhaps) underappreciated.

Some counter-examples to Page's notion of "localist"

Istvan S. N. Berkeley

Department of Philosophy and The Institute of Cognitive Science, The University of Louisiana at Lafayette, Lafayette, LA 70504. **istvan@USL.edu**
www.ucs.usl.edu/~isb9112

Abstract: In his target article Page proposes a definition of the term "localist." In this commentary I argue that his definition does not serve to make a principled distinction, as the inclusion of vague terms make it susceptible to some problematic counterexamples.

In his conclusion (sect. 8), Page tells us that "This target article has sought to clarify the differences between localist and fully distributed [connectionist] models." He makes similar claims in section 6.5.1. In this commentary, I will argue that Page has failed to adequately meet this goal. Page, however, does provide some interesting insight into a novel type of connectionist architecture in the rest of his paper.

Perhaps a useful place to start is to note that, although in current usage, the terms "connectionist" and "PDP" are often used as synonyms, the two terms once had distinct meanings that marked different approaches to network modeling. Originally, so-called "connectionist" models were associated with localist work of Ballard and others, from what Smolensky (1991) calls the "Rochester school." [See Ballard: "Cortical Connections and Parallel Processing" *BBS* 9(1) 1986; Smolensky: "On the proper treatment of connectionism" *BBS* 11(1) 1988.] Indeed, Smolensky (1991) goes on to note that the name "PDP" was intentionally chosen as a name to differentiate the much more distributed work of the "San Diego School," from that of the Rochester "connectionists." The fact that it has become common to use both terms almost interchangeably to describe a variety of approaches to network modeling suggests that maintaining this distinction does not serve any really useful purpose. Moreover, despite Page's attempts at clarifying the distinction, it has still not been shown to be a principled distinction.

Perhaps the clearest statement of Page's position (he himself in sect. 2.6, para. 7, refers to it as a "definition") runs as follows: "a localist model of a particular type of entity . . . is characterized by the presence of (at least) one node which responds maximally to a given familiar . . . example of that type . . . all familiar examples of that type . . . being so represented" (sect. 2.6, para. 6).

In order for any proposed definition to be successful, it must include all objects and entities that intuitively fall within the scope of a term, while ruling out those objects and entities which fall outside the scope of a term. Moreover, a definition should also be such that it appropriately attributes properties in a manner that is intuitive. Unfortunately, Page's definition, as formulated above, fails to meet these conditions of successful definition. This is perhaps best illustrated with a concrete example.

Berkeley et al. (1995) describe a network that was trained to determine the validity and type of a set of 576 logic problems, originally studied by Bechtel and Abrahamsen (1991). Berkeley et al. (1995) subjected the network to a detailed analysis using what has become known as the banding analysis technique (see also Berkeley 2000; Dawson 1998). One particular hidden unit of their network, hidden unit 8, was discovered to adopt three distinct levels or states of activation. These three levels of activation, along with their associated interpretations are illustrated in Table 1.

It appears, intuitively at least, that this unit is localist in nature and may naturally be interpreted as being a "connective detector" (indeed, this is the interpretation suggested by Berkeley et al. (1995). However, following Page's definition of localist above, although the unit appears to satisfy his criteria for being localist, the interpretation that should be applied to this unit is that it is an NOT BOTH . . . AND. . . detector. This is because it is this "familiar example" to which the node "responds maximally." Thus, as formulated, Page's definition seems to attribute properties incorrectly.

Part of the difficulty with Page's notion of a localist model comes from the fact that he intends the term "localist" to be applied at the level of entire models, whereas the properties that distinguish localist from nonlocalist models are specified at the level of the individual processing units. Although Page (sect. 2.6, para. 3) suggests that the term "localist" is used with some care and that it is important to be explicit about the entities under consideration, his notion of a localist model is still problematic. The difficulty arises due to the inherent vagueness of the phrase "familiar example" (see Pelletier & Berkeley 1995).

Consider the case of a simple network trained in the classic XOR problem. Should the inputs be considered as "localist" or not for the purposes of Page's definition? On one interpretation, where the "familiar example" was taken to be something like "single input to a two place function," the network would appear to clearly count as being localist. On the other hand, if our "familiar example" were taken to be something like "inputs to a two place function," the network would clearly not count as being localist. The problematic thing here is that XOR is just a low-order version of the parity problem. Moreover, parity problems are paradigm examples of the distributed problem type (see Rumelhart et al. 1986). This is because every single bit must be considered in order to determine the correct response to any particular input. Thus, it appears, due to the vagueness of the phrase "familiar example," Page's definition of "localist" fails to rule out even this limiting case.

Although Page's main discussion of novel network architectures

Table 1 (Berkeley). *Hidden unit activation bands along with the median activation values and range for each band and the interpretation supplied, for hidden unit 8 of the logic network described in Berkeley et al. (1995; p. 176)*

Band	Number of patterns in band	Median activation	Activation range	Interpretation
A	192	0.03	>0.01	Main Connective is OR
B	192	0.11	0.10–0.13	Main Connective is IF . . . THEN . . .
C	192	0.82	0.80–0.85	Main Connective is NOT BOTH . . . AND . . .

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

is interesting and useful, He appears to have failed to reach one of his intended goals: the goal of clarifying the distinction between localist and fully distributed models. Perhaps he should have used some other terminology?

ACKNOWLEDGMENT

This work was funded in part by a grant from the Artificial Neural Network Analysis Project LEQSF(1998-0)-RD-A-35, awarded to ISNB by the Louisiana State Board of Regents Support Fund.

Further arguments in support of localist coding in connectionist networks

Jeffrey S. Bowers

Department of Experimental Psychology, University of Bristol, BS8-1TN Bristol, United Kingdom. **J.bowers@bris.ac.uk** **eis.bris.ac.uk/~psjxb**

Abstract: Two additional sources of evidence are provided in support of localist coding within connectionist networks. First, only models with localist codes can currently represent multiple pieces of information simultaneously or represent order among a set of items on-line. Second, recent priming data appear problematic for theories that rely on distributed representations. However, a faulty argument advanced by Page is also pointed out.

Page's demonstration of the power, flexibility, and plausibility of connectionist models with local representations is an important contribution and should help bring back this class of models into the mainstream connectionist literature. In the present commentary, I point out an additional computational argument in support of his case, as well as describe some recent priming data that appears to provide further support for localist theories of visual word recognition.

But first, it should be noted that one of Page's key criticisms against distributed representations is mistaken (sect. 7.1). He argues that connectionist models with distributed codes cannot solve the stability-plasticity dilemma (or in other words, suffer from catastrophic interference). Although Page explicitly focuses on backpropagation models when making this criticism, the relevant point concerning the plausibility of distributed representations is that existing models that learn distributed representations do not suffer from this problem. In particular, the original ART network of Carpenter and Grossberg (1987) that solves the stability-plasticity problem with localist representations has been extended to cases in which it learns distributed patterns, in both unsupervised and supervised learning conditions (Carpenter 1997). It seems, therefore, that the stability-plasticity problem is not associated with distributed representations per se but rather the particular learning algorithms that tend to be used in the psychological literature.

Still, there are computational reasons to prefer models with localist representations, at least under some circumstances. A key point not sufficiently emphasized in the target article is that connectionist models that reject localist coding schemes do not do an adequate job of representing multiple pieces of information simultaneously, nor can they adequately represent order amongst a set of items. The reason is straightforward. If a pattern of activation across all the units defines a single concept, then overlapping two patterns on the same set of units will result in nonsense as there is no way of deciding which features go with which representation.

Of course, there are connectionist models that do represent multiple pieces of information simultaneously. But what is interesting to note about these networks is that they rely on localist codes in order accomplish this function. For example, Hummel and Holyoak (1997) developed a model of analogical reasoning that supports the co-activation of multiple semantic representations by relying on a complex interplay between distributed and localist codes, where localist representations function to bind together the semantic features that belong to one representation or another (thus avoiding the nonsense blends that result in networks that avoid localist representations). In the domain of phonology, Page and Norris (1998) and Burgess and Hitch (1999) developed localist connectionist systems of phonological short-term memory in which order was represented in terms of the relative activation level of localist codes (also see Grossberg & Stone 1986; Nigrin 1993). Once again, these models depended on localist coding schemes in order to achieve success.

It might be objected that recurrent networks can store and reproduce sequences of items, and some of these represent information in a distributed fashion. But there are two things to note. First, these models are often more successful in representing order when the inputs are represented in a localist scheme (e.g., Elman 1990; cf. Marcus 1998). But more important for present purposes, the ability to represent sequential order in recurrent networks is the product of slow learning mediated by changes in the connection weights between units, producing a long-term representation of ordered information from the training set. What is needed to model analogical reasoning or phonological short-term memory (among many other capacities), however, is an ability to temporarily activate multiple items in real-time. For example, to model phonological short-term memory, a network must be able to reproduce a random sequence of seven or so inputs following a *single* presentation of the list, with the information quickly decaying from the system after recall. This function has yet to be modeled within a system that relies on distributed codes. Of course, the mere fact that current models rely on localist codes to support these functions does not imply that localist codes are required. But until a model with distributed codes can perform these computations, the common rejection of localist coding seems premature.

Finally, let me briefly note one finding reported by Bowers and Michita (1998) that may pose a problem for models of word identification that rely on distributed coding schemes, but which can readily be accommodated within localist approaches. Employing the lexical decision task, we compared long-term priming for Japanese words written in the same script at study and test (Kanji-Kanji or Hiragana-Hiragana) or in visually unrelated script (Kanji-Hiragana or Hiragana-Kanji). Averaging across the two studies, same script priming was 28 msec and cross-script priming was 24 msec, a difference that did not approach significance. In addition, little priming was obtained when items were spoken at study and written at test (6 msec averaging across studies), suggesting that cross-script priming was mediated by abstract and modality-specific orthographic codes. What is important to note about these scripts is that they do not share any letter-by-letter correspondences, as the individual characters in Hiragana and Kanji represent syllables and morphemes, respectively. Accordingly, the only level at which the two items can map together within the orthographic system is at the whole word *lexical* level, typically rejected by distributed connectionist models of reading (e.g., Plaut et al. 1996; Seidenberg & McClelland 1989). Whether connectionist models with distributed representations could accommodate these lexical priming effect seems doubtful (for similar results, see Brown et al. 1984; Feldman & Moskovljevic 1987).

Neural networks for selection and the Luce choice rule

Claus Bundesen

Department of Psychology, University of Copenhagen, DK-2300 Copenhagen S, Denmark. **bundesen@axp.psl.ku.dk**
axp.psl.ku.dk/~cvc/cb/index.htm

Abstract: Page proposes a simple, localist, lateral inhibitory network for implementing a selection process that approximately conforms to the Luce choice rule. I describe another localist neural mechanism for selection in accordance with the Luce choice rule. The mechanism implements an independent race model. It consists of parallel, independent nerve fibers connected to a winner-take-all cluster, which records the winner of the race.

In the target article, Page proposes a generalized localist model, which is a fairly simple, lateral inhibitory network. The network implements a Thurstone (1927, Case V) selection process by adding (zero-mean, constant-variance, uncorrelated) Gaussian noise to its inputs and selecting the unit that receives the maximal noisy input. Page points out that (a) the choice probabilities of a Thurstone (1927, Case V) selection process are close to probabilities predicted by a Luce (1959) choice rule and (b) the asymptotic choice probabilities of the Thurstone selection process, approached under uniform expansion (Yellott 1977) of the choice sets, are identical to probabilities predicted by a Luce choice rule. In section 4.2.3, Page states that, to his knowledge, no simple "neural" mechanism has previously been suggested for implementing selection in accordance with the Luce choice rule. In note 1, however, Page mentions that "independent race models . . . have similar properties with regard to the Luce choice rule," refers to Bundesen (1993), and suggests that this point should be further considered. I shall do so.

In the terminology of Bundesen (1987), a *race model* of selection is a model in which elements in the choice set are processed in parallel and selection is made of those elements that first finish processing (the winners of the race). If processing times for individual elements in the choice set are independent random variables, the model is called an *independent* race model. There are close relationships between independent race models and the Luce choice rule. Under weak, general assumptions, Bundesen (1993) proved that the Luce choice rule holds for an independent race model if, and only if, the hazard functions of the processing times of the elements in the choice sets are mutually proportional to each other. This implies, for example, that the Luce choice rule holds if the hazard functions are constant over time (i.e., if the processing times are exponentially distributed; cf. Bundesen et al. 1985; Luce & Green 1972). If the hazard functions vary over time, then the Luce choice rule holds if they vary in synchrony so that the ratios among the hazard rates are kept constant over time (cf. Bundesen 1990, Footnote 4; Marley & Colonius 1992). Bundesen (1993) further proved that, regardless of whether the hazard functions are mutually proportional, the Luce choice rule holds for the asymptotic choice probabilities approached under uniform expansion of the choice sets. Apparently, the Luce choice rule holds asymptotically for any plausible independent race model.

To make a neural network implementation of an independent race model, we need a device for recording the winner of a neural race (a firing race among independent neural fibers). Bundesen (1991) noted that a winner-take-all cluster of the general type proposed by Grossberg (1976; 1980) can be used for this purpose. The cluster consists of a set of units such that each unit excites itself and inhibits all other units in the cluster. Suppose that, when the cluster is initialized, a single impulse from the environment to one of the units is sufficient to trigger this unit. Also suppose that, once the unit has been triggered, it keeps on firing because of its self-excitation. Finally suppose that, when the unit is firing, it inhibits the other units in the cluster so strongly that they cannot be triggered by impulses from the environment. If so, one can read from the state of the cluster which unit received the first impulse from the environment after the cluster was initialized. Thus, the cluster serves as a device for recording the winner of a race.

Neural fibers are stochastic latency mechanisms (McGill 1963). A typical neural fiber behaves approximately as a Poisson generator: To a first approximation, the latency measured from an arbitrary point of time (corresponding to the starting time of a race) to the first firing of the fiber is exponentially distributed. A set of parallel, independent nerve fibers (Poisson generators) connected to a winner-take-all cluster for recording the fiber that fires first (the winner of the race) forms a neural mechanism for implementing selection in strict accordance with the Luce choice rule (Bundesen 1991). This implementation of selection by the Luce choice rule is the most simple I can imagine. Approximations to selection by the Luce choice rule are obtained if the network is elaborated so that responses are based on the fiber that first reaches a criterion of having fired γ times, where $\gamma > 1$ (*gamma* race models in the terminology of Bundesen 1987). Other approximations to selection by the Luce choice rule are obtained if the network is elaborated so that responses are based on the fiber that first reaches a criterion of having fired d times more than any other fiber, where $d > 1$ (random-walk models; cf. Bundesen & Harms 1999; Logan 1996).

The many ways to distribute distributed representations

A. Mike Burton

Department of Psychology, University of Glasgow, Glasgow, G12 8QQ, United Kingdom. **mike@psy.gla.ac.uk**
www.psy.gla.ac.uk/~mike/home.html

Abstract: Distributed representations can be distributed in very many ways. The specific choice of representation for a specific model is based on considerations unique to the area of study. General statements about the effectiveness of distributed models are therefore of little value. The popularity of these models is discussed, particularly with respect to reporting conventions.

Page's localist manifesto does great service to the psychology community. He brings together a set of arguments which have previously existed largely as asides or unarticulated hunches accompanying certain modelling research. While agreeing with Page's analysis, I will make two additional comments, the first on a technical aspect of distributed representations, and the second on the reason for their appeal.

If we are told that a particular representation is distributed, then we know almost nothing about it. There is an important sense in which all localist models are alike, but all distributed models are distributed after their own fashion. Properties usually attributed to distributed representations, such as generalisability and graceful degradation, are in fact properties of specific representations, distributed in particular ways, for the purposes of a specific model. Learning algorithms designed to acquire a range of representations are designed to optimise the representational capacity of a particular network, given the range of representations required. Consider two of the influential connectionist models discussed by Page, the model of word recognition by Seidenberg and McClelland (1989) and the model of deficits in person recognition by Farah et al. (1993). These two models have such radically different learning algorithms, resulting in such different representations, that they seem to relate to each other not at all. It is certainly true that both represent information as patterns of activation across a simple set of processing units, and this is a beguiling similarity. However, the similarity is at surface level only. The underlying method by which unit activations are assigned a role in any given representation is so different in the two models, that they cannot be considered to share very much in common.

Despite the heterogeneity of distributed models, they are very often represented as providing general properties, and especially properties which emerge as a direct result of their distributed nature (e.g. see Farah 1994b). Since each distributed model achieves these properties in a different way, there is no general reason to be impressed by the approach. Instead, one must evaluate models individually.

In contrast, localist models are much more severely constrained. If one wants to build a model of (say) familiar person recognition, then it is clear that one must include a specific representational primitive corresponding to (say) Bill Clinton. Given that all localist models will have to include this representation, one is able to choose between models on the basis of how well they are structured, and how well they capture the data, rather than needing to enquire whether model behaviour is an artefact of the particular representational primitives one has chosen to use. Page

correctly points out that many localist models have a distributed component. In the person recognition case, it would be absurd to suggest that one could arrive at a satisfactory model without front-end primitives which are distributed over inputs (e.g., pixels, or retinal activations). Some models of face recognition include these input representations, which are distributed, as well as localist representations of the "Bill Clinton" form (Burton et al. 1999). To my knowledge, no localist modeller has ever denied that big things are made up of little things. What divides localist and distributed modellers is whether we need explicit representations of "things" at all. Developments of these arguments can be found in Young and Burton (1999), and in subsequent discussion by O'Reilly and Farah (1999) and Burton and Young (1999).

Page asks why localist representations have not been popular recently, and lists some common misconceptions. There is a further point to add here. Psychologists have traditionally found it rather hard to evaluate models. There is a marked contrast between presentation of experimental work, and presentation of modelling work. Almost any trained psychologist, on reading an experimental paper, could replicate the experiments reported. There are two reasons for this. First, the psychologist will have been trained in experimental techniques and the analysis of data. Second, publication conventions have evolved which require detailed exposition of methods.

Unfortunately, it is much harder to reproduce simulations from the literature. Very many psychologists have not been trained in simulation techniques. University programmes in psychology emphasise empirical methodology, in many cases exclusively. This leaves researchers poorly equipped to evaluate claims made by computer modellers, or to replicate the models reported. Further, the reporting conventions concerning publication of simulations are not remotely as sophisticated as those for publication of empirical work. This often means that insufficient information is provided to allow replication, even for those with the technical training to do so.

The problem in evaluating models means that one can sometimes be impressed by the behaviour of a model one does not fully understand. Models, when demonstrated, certainly have an immediacy which can make them more compelling than other forms of theoretical development. I believe this is a general problem which results in the evaluation of modelling being much more susceptible to fashion than the evaluation of empirical work, and this may account for the popularity of distributed representations over the past fifteen years. Since this has been rather a critical passage, I should add that the fault has not lain exclusively with the modellers. McClelland and Rumelhart (1988) provided an extremely useful set of software demonstrations, designed explicitly to encourage researchers to learn the detailed workings of these models. It is ironic that some of the most vocal critics of distributed representations first came to modelling through this software.

These problems of reporting and expertise are not unique to psychology. McDermott (1981) provides an insightful analysis of reporting conventions in artificial intelligence, which is still of contemporary interest. His proposals include the notion that we can learn as much from an honest failure as from a working model. Further, he suggests that textual descriptions of models are no substitute for working versions, designed in such a way as to allow detailed inspection. Had these conventions been adopted, it seems unlikely that distributed representations would have been so dominant in recent research.

Combining distributed and localist computations in real-time neural networks

Gail A. Carpenter

Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215. **gail@cns.bu.edu** **cns.bu.edu/~gail/**

Abstract: In order to benefit from the advantages of localist coding, neural models that feature winner-take-all representations at the top level of a network hierarchy must still solve the computational problems inherent in distributed representations at the lower levels.

By carefully defining terms, demonstrating strong links among a variety of seemingly disparate formalisms, and debunking purported shortcomings of winner-take-all systems, Page has made a significant contribution toward the creation of a functional classification of the growing array of neural and cognitive models. One important feature of the target article is a clarification of terminology. For example, a model is here labeled "localist" when the representation at the top level (n) of a network hierarchy is localist (sect. 2.6, para. 1). This definition is based on the logical conclusion that, once a code representation has reached the limit of winner-take-all compression, additional network levels would be redundant. Conversely, any nonredundant localist system would normally have distributed representations at the lower levels $1 \ldots n - 1$. By considering systems in their hierarchical configurations, Page shows that models and related data previously viewed as "distributed" in fact derive essential properties from localist mechanisms.

Page's hierarchical definition of localist networks implies that any such system with more than two levels could inherit the computational drawbacks, as well as the benefits, of distributed networks. As Page points out (sect. 7.1), many distributed models are subject to catastrophic interference and require slow learning and multiple interleaved presentations of the training set. One of my research goals in recent years has been the development of real-time neural network systems that seek to combine the computational advantages of fully distributed systems such as multilayer perceptrons (Rosenblatt 1958; 1962; Rumelhart et al. 1986; Werbos 1974) with the complementary advantages of localist systems such as adaptive resonance theory (ART) networks (Carpenter & Grossberg 1987b; 1993; Carpenter et al. 1991; Carpenter et al. 1992). An initial product of this ongoing project was the distributed ART (dART) family of neural networks (Carpenter 1996; 1997; Carpenter et al. 1998), which permit fast as well as slow learning, and distributed as well as localist code representations, without catastrophic forgetting. Where earlier ART models, in order to help stabilize memories, employed strongly competitive activations to produce winner-take-all coding, dART code representations may be distributed across any number of nodes. In order to achieve its computational goals, the dART model includes a new configuration of the network architecture, and replaces the traditional path weight with a *dynamic weight*, which is a joint function of current coding node activation and long-term memory (LTM). The dART system also employs new learning rules, which generalize the instar (equation [10], sect. 4.4, para. 3) to the case where the target node activation patterns at layer L_2 may be fully distributed. The original instar equation implies that, unless learning is very slow, all weight vectors $\mathbf{w}_j$ would converge to the same input pattern $\mathbf{a}$ at every location where the target L_2 node is active ($a_j > 0$). With the distributed instar learning rule, dynamic weights automatically bound the sum of all LTM changes, even with fast learning. The computational innovations of the dART network would allow distributed representations to be incorporated at levels $1 \ldots n - 1$ in a network hierarchy while retaining the benefits of localist representations at level n.

In contrast to the aim of the dART research program, which is to define a real-time, stand-alone neural network with specified properties, the primary aim of the target article is to unify diverse computational and conceptual themes. In the service of this goal, the corresponding learning module (sect. 4.1) is, by design, skele-

tal. However, such a partially specified model might risk being unduly rejected on the basis of what it seems not to do, and some of the model's properties are subject to misinterpretation if taken at face value. For example, Page's localist model permits learning only at an uncommitted node, which then encodes the current input. The decision whether to activate an uncommitted node depends upon the value of the threshold θ, which is somewhat analogous to the vigilance matching parameter ρ in an ART model. In particular: "If the threshold is set slightly lower [than 1], then only activation patterns sufficiently different from previously presented patterns will provoke learning" (sect. 4.1, para. 2). Page points out that this construction would help solve the problem of catastrophic interference, since coding a new pattern does not affect previous learning at all. On the other hand, this feature might also be the basis for rejecting this model, and by extension other localist models, since each category can be represented only as a single exemplar: there is no opportunity for new exemplars that correctly activate a given category to refine and abstract the initial learned representation. In contrast, a more fully specified localist model could permit controlled learning at committed nodes as well as at uncommitted nodes, hence creating prototype as well as exemplar memories while still retaining the ability to resist catastrophic interference. Even though this capability is not part of Page's simplified model, the possibility of learning at committed nodes is implied later in the article (sect. 4.5, para. 3): "when at least one of the associates is learned under low-vigilance (cf. prototype) conditions, remapping of items to alternative associates can be quickly achieved by rapid reconfiguration of connections to and from the mapping layer."

Similarly, a reader may be misled who takes seriously the assertion: "The extension [of the learning module in the target article] to continuous activations will usually be necessary and is easily achieved" (sect. 4.1, para. 1). This statement is true, but defining an extension of the simplified system is not a matter of straightforward substitution. In particular, the learning module is defined only for the case of binary inputs, and the validity of its computational properties relies implicitly on the assumption that $\mathbf{a}\cdot\mathbf{a} = |\mathbf{a}| = \|\mathbf{a}\|^2$, which is true only when $\mathbf{a}$ is binary.

In summary, the simplified localist learning module defined by Page is a valuable tool for unifying and clarifying diverse formalisms, but a more complete computational development is needed to define stand-alone neural network systems that realize the promise of the localist analysis.

Localist representations and theoretical clarity

Norman D. Cook

Faculty of Informatics, Kansai University, Takatsuki, Osaka, 569 Japan.
cook@res.kutc.kansai-u.ac.jp

Abstract: In the Localist Manifesto, Page enumerated several computational advantages that localist representations have over distributed representations, but the most important difference between such networks concerns their theoretical clarity. Distributed representations are normally closed to theoretical interpretation and, for that reason, contribute little to psychology, whereas the meaning of the information processing in networks using localist representations can be transparent.

Page has clearly demonstrated that localist representations can accomplish as much as or more than fully distributed representations, and obligates us to reconsider the significance of the common, but facile arguments against the localist approach. Insofar as he argues that both local and distributed processes are often needed in neural networks, the manifesto means that modelers should not shy away from declaring that "this group of cells performs function X" or "represents information Y." I have no quarrel with those constructive arguments, but I think that Page has been too gentle on the negative aspects of fully distributed representations. He hints at the problem by referring to the "murkiness" of distributed representations (sect. 4.5), noting that "fully distributed networks are much more difficult to interpret than their localist counterparts" (sect. 7.6) and concludes that they "encourage opacity in modeling" (sect. 8). The relative clarity of local and distributed representation is important because the vast majority of neural network simulations in psychology are devised primarily to facilitate an understanding of brain processes. In other words, they are at some level explanatory theories, and not simply "behavioral" simulations intended to reproduce a particular data set. For this reason, although the technical advantages of localist modeling and learning algorithms that do not suffer from dispersion and catastrophic interference may be real, they are less important than the issue of the clarity of theory. Opacity is the connectionist's pregnancy test: a little bit is already too much. Unlike neural networks used as practical tools – where efficiency can be a relative measure, no level of murkiness can be tolerated in a theoretical model. The specific merit that the neural network methodology, in general, exhibits is that a highly interconnected, multineuron network can be constructed and all murkiness studied and manipulated until the ambiguities are understood and eliminated. Insofar as fully distributed representations allow results to be obtained without clarifying ambiguities, or permit the modeler to avoid directly confronting the nature of the information processing that is taking place, or allow "mysterious" effects that somehow fall out of a highly complex system, then the explanatory model fails to provide understanding and does not exploit the strengths that are uniquely possible in the neural network approach. A model that does not facilitate understanding is a waste of virtual processing and real time.

A still-prevalent pitfall of many neural networks is to produce results that rely crucially on an *inability* to explain what has occurred (Cook 1995a; 1995b; Cook et al. 1995). It is not a problem that the model does not explain everything or contains starting assumptions that are themselves debatable; that is always the case (Cook 1999). It is, however, a problem when unexamined (and, in the case of extremely large distributed-representations, effectively unexaminable) effects play a crucial role in holding the theoretical model together. At the other extreme, a localist model that requires even something as dubious as a unique "grandmother cell" (something that Page would say is unnecessary to postulate within a realistic localist framework) is still to be preferred because the problems, deficiencies and over-simplifications of the model are at least made explicit and placed face up on the table. In contrast, the worst-case distributed representation model will hide its weaknesses in the nonlocalized representation and allow the modeler to pretend that something profound has emerged mysteriously from the complexity of the system.

A simple example shown in Figure 1 illustrates how a localist representation produces a more transparent understanding of network dynamics, despite the fact that the computational results are identical using either local or distributed architectures. Both nets learned the "harmonicity" of simultaneous musical tones using similar learning rules and input stimuli, but in Network B a specific grouping of neurons was implemented and Dale's Law (any given neuron will have exclusively excitatory or inhibitory effects) was enforced. The networks underwent supervised learning such that the strength of activation of the output neuron corresponds to the degree of musical harmonicity (as determined in psychophysical experiments). Network B evolved a set of neurons that have the function of representing various musical concepts. Certain hidden layer neurons represent dissonance between tone pairs, others represent tension in three note chords, still others represent the consonance of tone combinations, and so on. Network B is in effect localist in so far as individual "hidden" neurons come to represent specific functions. Similar functions are in fact also realized in Network A, but in a distributed fashion that masks the cognitive meaning of what the network accomplishes. The architecture of Network B was designed with an explicit model of harmonicity in mind, but even when an explicit model is not the

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

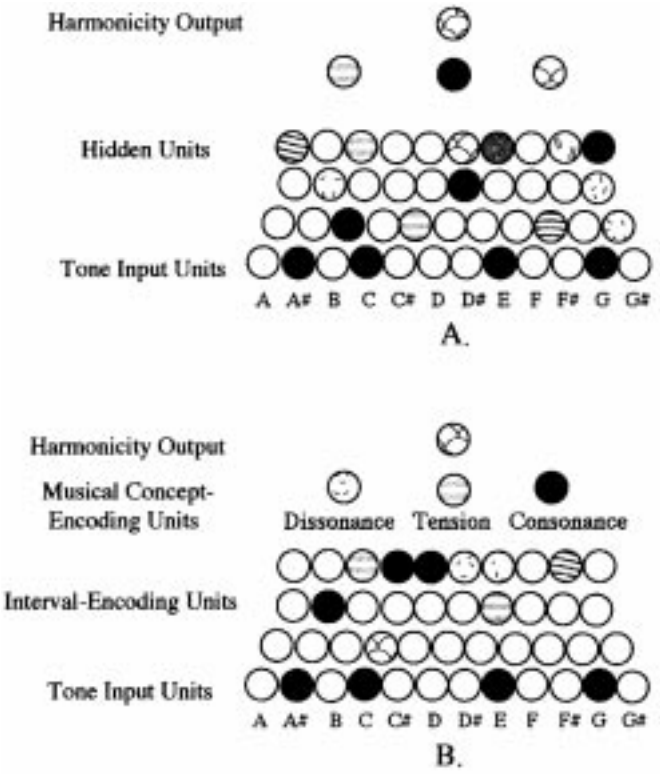

Figure 1 (Cook). Two networks that have been trained to learn the harmonicity of simultaneous tones. Network A has a distributed representation of all musical concepts – the relationships among which remain indecipherable in the network. In contrast, Network B learns the same data, but the qualitative and quantitative relationships among dissonance, chordal tension, and consonance are evident in the trained network. Here, grouping of neurons and enforcing Dale's Law are sufficient to give Network B a localist representation of the musical concepts that presumably Network A also learned. The networks are functionally identical, but the localist version has potential significance for the psychology of music that the distributed version lacks.

starting point of the simulation, enforcement of Dale's Law, the local grouping of neurons, and the usage of physiologically plausible learning rules all work to facilitate local representations and, as a consequence, theoretical transparency.

ACKNOWLEDGMENT

This work was supported by the Research for the Future Program, administered by the Japan Society for the Promotion of Science (Project No. JSPS-RFTF99P0140).

Efficiency, information theory, and neural representations

Joseph T. Devlin,[a] Matt H. Davis,[b] Stuart A. McLelland,[a] and Richard P. Russell[a]

[a]*Centre for Speech and Language, Department of Experimental Psychology, University of Cambridge, Cambridge, CB2 3EB, England;* [b]*Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, CB2 2EF England.*
jtd21@cam.ac.uk csl.psychol.cam.ac.uk/~jdevlin
matt.davis@mrc-cbu.cam.26.ac sam26@cam.ac.uk
rpr23@cam.ac.uk

Abstract: We contend that if efficiency and reliability are important factors in neural information processing then distributed, not localist, representations are "evolution's best bet." We note that distributed codes are the most efficient method for representing information, and that this efficiency minimizes metabolic costs, providing adaptive advantage to an organism.

According to Page, localist representations can be both efficient and reliable. Specifically, he argues that localist encodings require fewer computational resources than certain types of distributed models (e.g., Hopfield networks), they are easier to learn and interpret than distributed representations, and when augmented with repetition coding, they are robust to both noise and damage. In this commentary we challenge these claims and suggest that efficiency and reliability support distributed, rather than localist, representations yielding an adaptive advantage for an organism by reducing metabolic costs.

Information theory distinguishes between *source coding* and *channel coding*. Source coding is the process of representing information efficiently (i.e., with minimal length codes) while channel coding describes methods for introducing redundancy into a representation to make it robust to the effects of noise or damage. Page proposes localist source coding in which individual items are represented by a single binary unit or neuron. In such a system the resources required for representing information scale linearly with the number of items – in other words, the system is $O(N)$ (read "order N"). In contrast, distributed representations require only $O(\log_2 N)$ units to represent N items and therefore scale with the log of the number of units – a dramatic saving in computational resources. Furthermore there are minimal differences between these two coding schemes in terms of reliability. In the localist encoding, an incorrect bit can be detected but not corrected whereas in distributed representations, the error may not be detected or corrected.

Channel coding is used to both detect and correct errors. In general, these methods add systematic redundancy to the original code, improving reliability by increasing the cost of representing information. Page proposes a scheme in which units are repeated multiple times. Errors are corrected by taking a majority vote. For example, if each unit is repeated five times then noise which alters up to two bits per item can be corrected. While this technique is consistent with his "localist manifesto," it is not particularly efficient. More efficient channel coding schemes (e.g., Hamming coding) exist that add new units to a representation based on co-occurrence information (i.e., multiple active units). Of course, for co-occurrences to exist, the original pattern cannot be localist.

To reiterate, efficient source and channel coding techniques involve distributed, not localist, representations. In source coding, the efficiency of the representation is dramatically improved – from $O(N) \rightarrow O(\log_2 N)$. In channel coding, the improvement is less dramatic but the most efficient codes rely on distributed representations.

Page raises three objections to this type of efficiency. First, "this notion of efficiency will count for nothing if the process by which the mapping must be learned is not only inefficient but also rather implausible" (sect. 6.2). Second, localist representations are more comprehensible than distributed codes. Finally, "efficiency in modelling, particularly when arbitrarily defined, is not necessarily an aim in itself" (sect. 6.2). We will focus primarily on this final point and only briefly address the other two issues.

Is efficiency per se beneficial to living organisms or just to satellite communications? The brain has evolved to perform certain functions and thus the efficiency of representations – in terms of number of neurons and robustness in the face of noise and damage – may not be of primary importance from a functional perspective. However, since the brain is one of the most energy-demanding organs of the body, adopting representations that reduce the metabolic demands associated with neural processes will almost certainly have survival value for an organism (Laughlin et al. 1998).

The metabolic costs associated with neural information processing come in two forms, the baseline metabolic costs of maintaining neurons and the additional costs incurred by neural activ-

ity. Localist representations increase the number of neurons required for information processing but decrease the average activity of individual neurons. Distributed representations do the opposite. Since baseline synaptic activity represents up to 75% of resting glucose utilization of the brain (Phelps et al. 1979), it is likely that significant reductions in metabolic cost can be obtained by minimizing the number of neurons. Hence efficient distributed representations will minimize metabolic costs.

Page raises two additional objections to this notion of efficiency: comprehensibility and learnability. Presumably both will be addressed in other commentaries so we will limit our response to two brief comments. First, although localist representations are often transparent and therefore can be interpreted by an outside observer much more readily than distributed representations, the important point to remember here is that this is not the purpose of neural representations. Instead, their purpose is to offer the maximal adaptive advantage to an organism. Second, Page claims that learning distributed representations is both inefficient and implausible. However, if McClelland et al. (1995) theory of complementary learning systems is correct, then the metabolic costs of maintaining a hippocampus must be outweighed by the massive reduction in neocortex which it allows. Furthermore, although backpropagation may be biologically implausible, more realistic algorithms do exist (e.g., Hinton et al. 1995). Thus learning distributed representations need not be an insurmountable problem.

In conclusion, we contend that both efficiency and reliability lead one to adopt distributed, not localist, representations. Distributed codes minimize metabolic costs and therefore provide an adaptive advantage to an organism. Let us be clear. We are not suggesting that the brain uses an efficient coding scheme because it is theoretically optimal. Instead, our claim is that evolution has developed schemes to help minimize the metabolic cost of neural computation. This is achieved through the use of sophisticated encoding schemes resulting in the use of distributed representations. Page (sect. 8) claims "that if the brain doesn't use localist representations then evolution has missed an excellent trick." We would like to suggest, however, that if efficiency and reliability are important factors in neural information processing, then distributed, not localist, representations are evolution's best bet.

ACKNOWLEDGMENT

This commentary was supported by a McDonnell-Pew grant and an MRC Program grant to L. K. Tyler.

The case against distributed representations: Lack of evidence

Simon Farrell and Stephan Lewandowsky

Department of Psychology, University of Western Australia, Nedlands, W.A. 6907, Australia. **{simon; lewan}@psy.uwa.edu.au**
www.psy.uwa.edu.au/user/{simon; lewan}

Abstract: We focus on two components of Page's argument in favour of localist representations in connectionist networks: First, we take issue with the claim that localist representations can give rise to generalisation and show that whenever generalisation occurs, distributed representations are involved. Second, we counter the alleged shortcomings of distributed representations and show that their properties are preferable to those of localist approaches.

Page eloquently extolls the virtues of localist representations and their presumed superiority over distributed representations in connectionist networks. We focus on two aspects of the argument: First, we contend that generalisation cannot occur without involvement of distributed representations. Second, we refute six objections levelled against distributed representations.

Localist representations do not generalise. Page identifies a representation as localist if it is "possible to interpret the state of a given node independent of the states of other nodes" (sect. 2.2., para. 7). For example, the representations {0, 1} and {1, 0} for items A and B would be considered localist, whereas {0, 1} and {1, 1} would be considered distributed. Critically, Page advocates a hybrid approach that "supplements the use of distributed representations . . . with the additional use of localist representations" (sect. 1, para. 3). In support, he presents a generic "localist" network that exhibits a number of desirable properties, among them the ability to generalise a learned response to noisy input. Critics have often questioned whether localist representations are capable of generalisation, so its occurrence in a localist network deserves scrutiny.

We contend that the network's ability to generalise arises entirely from the use of *distributed* representations at the input layer which "reflect, in a graded fashion, the degree of similarity that the current input shares with each of those learned patterns" (sect. 4.3.1, para. 2). Localist representations, as defined by Page, are necessarily orthogonal to each other. Hence, the graded similarity that Page identifies as critical for generalisation is inextricably linked to the presence of distributed representations at the input layer.

Although this supports our claim that generalisation requires distributed representations, other research shows that they need not be confined to the input layer. Hinton (1986) presented a multilayer network in which representations at the input layer were strictly localised whereas the hidden layer used distributed representations. The network was found to exhibit meaningful generalization. Subsequent analysis of the activation profiles of the hidden layer confirmed the crucial role of distributed representations.

Distributed representations resist objections. Page attributes six deficiencies to distributed representations (sects. 7.1–7.6), all of which revolve around the overlap of representations at the hidden layer. We counter these objections as follows.

7.1. Catastrophic interference. We concur with Page that interleaved learning, in particular as instantiated by McClelland et al. (1995), is not a preferred solution to catastrophic interference. We also agree that elimination of catastrophic interference requires minimisation of the overlap between representations at the hidden layer. However, it does not follow that localist representations are therefore preferable. First, as alluded to by Page, *distributed* solutions other than interleaved learning exist that reduce catastrophic interference (for a review, see Lewandowsky 1994). Second, localist solutions to the interference problem, as for example provided by ALCOVE (Kruschke 1992), have been shown to engender impaired generalisation (Lewandowsky 1994). By contrast, all available distributed solutions to interference are known to retain their ability to generalise (Lewandowsky 1994).

A careful consideration of catastrophic interference and generalisation therefore points to an advantage of distributed over localist representations.

7.2. Implausibility of the learning rule. This criticism rests entirely on the biologically dubious nature of the gradient-descent algorithm in back-propagation. However, other distributed learning rules, such as Hebbian learning, have been directly supported by biological research (e.g., Kelso et al. 1986). Moreover, at a psychological level, direct empirical support for distributed representations has been provided by the plethora of studies that have confirmed the predictions of the Rescorla-Wagner theory of learning (e.g., Shanks 1991). An essential element of the Rescorla-Wagner theory is that stimuli (e.g., in a categorisation task) are represented by ensembles of attributes or features.

7.3. The dispersion problem. Can distributed representations capture the similarity between sentences such as "John loves Mary" and "Mary loves John"? (sect. 7.3, para. 1). In agreement with Page, we find this question difficult to answer for all possible distributed schemes. However, we note that distributed linguistic parsers have been implemented that address this problem (e.g., Miikkulainen 1996). It follows that distributed schemes are not at a selective disadvantage in handling the dispersion issue.

7.4. Problems deciding "when" and "what." In many distributed networks, a response is identified by some extraneous process "done by the modeller rather than by the model" (sect. 7.4, para. 2). Page correctly identifies this as a serious problem. However, the solution to the problem need not be localist. Distributed networks that can unambiguously identify a response, without any extraneous mechanism or any of the other objections raised by Page, have been presented by Lewandowsky (1999), Lewandowsky and Farrell (in press), and Lewandowsky and Li (1994).

7.5. Problems of manipulation. Contrary to the claim in the target article, response suppression can demonstrably be accomplished in a distributed network using (Hebbian) "anti-learning" (e.g., Lewandowsky, in press; Lewandowsky & Li 1994). Page is correct in assuming that other items might be affected to the extent that they are similar to the suppressed target, but there is no evidence that this does not occur empirically. Indeed, this suppression of "neighbours" might explain why similar list items suffer more during serial recall than dissimilar ones.

7.6. Problems of interpretation. We agree that distributed models are more difficult to interpret than those with localist representations. This is because distributed models, unlike localist schemes, are capable of restructuring the input in interesting and novel ways that may at first glance escape interpretation.

Consider the distributed network presented by Hinton (1986). The network learned a set of input-output patterns whose semantic structure was not captured by the localist input and output representations. Through supervised learning alone, the network was found to organise its hidden layer into a distributed representation that captured the underlying semantic structure. While it required some analysis to visualise that distributed representation, the very fact that it was not immediately obvious implies that the network learned something novel and interesting.

Conclusion. We concur with Smolensky (1990) that representation is "crucial . . . , for a poor representation will often doom the model to failure, and an excessively generous representation may essentially solve the problem in advance" (p. 161). Unlike Page, we do not believe that localist representations are inherently preferable to distributed approaches. The alleged flaws of distributed schemes cited by Page are in fact desirable properties.

Why localist connectionist models are inadequate for categorization

Robert M. French and Elizabeth Thomas

Psychology Department (B32), University of Liège, 4000 Liège, Belgium; Institut Léon Frédéricq, University of Liège, 4000 Liège, Belgium. **{rfrench; ethomas}@ulg.ac.be www.fapse.ulg.ac.be/Lab/cogsci/rfrench.html**

Abstract: Two categorization arguments pose particular problems for localist connectionist models. The internal representations of localist networks do not reflect the variability within categories in the environment, whereas networks with distributed internal representations do reflect this essential feature of categories. We provide a real biological example of perceptual categorization in the monkey that seems to require population coding (i.e., distributed internal representations).

Despite Page's bold frontal assault on distributed connectionism, we wish to point out what appear to us to be two significant problems with this type of localist network.

The problem of category variability. Consider two categories, "fork" and "chair." The variability within the first category is very low: there just aren't that many different kinds of forks. Chairs, on the other hand, come in all different shapes, sizes and materials: they range from beanbag chairs to barstools, from overstuffed armchairs to rattan chairs, from plastic lawn chairs to that paragon of ergonomic design, the backless computer chair that you kneel on; some have four feet, some three, some none; some have backs, some don't; some are made of metal, others plastic, others wood, others, cloth and Styrofoam pellets, and so on. In other words, the variability within the category of *chair* is enormous.

But in the localist model proposed by Page, and in localist models in general, *this information about category variability is lost.* In distributed models, it takes more hidden nodes to encode a category with high-variability than one with low variability. In other words, the internal representations reflect external category variability. However, the category nodes in localist networks are unable to reflect this differential variability-in-the-environment of various categories. The one-node internal "representation" corresponding to the extremely low-variability category "fork" is precisely the same as the one-node internal representation corresponding to the highly variable category "chair."

Why is this a problem? Most significantly, because of the well-documented fact of category-specific losses: in general, naming of inanimate objects is found to be better preserved than naming of animate objects (Farah et al. 1996; Funnell & Sheridan 1992; Warrington & Shallice 1984). A model with distributed internal representations can handle this problem quite simply: low-variance categories (e.g., many natural kinds categories, like *cat*, *horse*, etc.) are encoded over fewer "units" than high-variance categories (e.g., many artificial kinds categories, like *chair*, *tool*, etc.) Random lesioning of the model will be more likely, on average, to destroy the representation of a category with low-variability (e.g., natural kinds categories) that is coded over a small number of units than a high-variability category (e.g., artificial kinds categories) coded over a large number of units. Localist models in which all the category nodes are the same will have considerable problems explaining category-specific deficits of this kind, especially when the featural inputs to the internal category representations remains intact. If, on the other hand, we assume differing degrees of variance associated with the internal encoding of different categories, these kinds of deficits can be predicted in a straightforward manner, as French (1997b) and French and Mareschal (1998) have shown using a dual-network architecture based on the hippocampal-neocortical separation proposed by McClelland et al. (1995).

As Page points out in his target article, we have argued for the necessity of "semi-distributed" representations in connectionist models for many years. But "semi-distributed" does not mean localist. "Semi-distributed" representations preserve category variance information; localist representations do not. Further, it seems crucial to us that these semi-distributed representations *emerge* as a result of learning.

Biological category representations. Page is right in pointing out that some of what is called population or ensemble coding in biological systems can be viewed as localist. For example, even though broadly tuned, cells of the motor cortex have their maximum activity tuned to a particular direction (Georgopoulos et al. 1993). One should therefore be able to ascertain the direction being represented by looking at the activity of individual neurons (or very small groups of neurons). However, an example of a cognitively relevant task that cannot be achieved in this fashion can be found in the anterior temporal cortex. Vogels (1999) reports on the responses of cells in this area during a tree, non-tree categorization task by a monkey. Most of the cells were stimulus selective, (i.e., they did not respond to all of the presented stimuli) and responded to both trees and non-trees. The maximum response of these neurons was not tuned to either category. Even though it was the case that certain (category-selective) neurons responded to particular subsets of tree exemplars, *no individual neuron (or small set of neurons) responded to all of the presented trees,* while not responding to any non-tree. These category-selective neurons alone did not appear to play an important role in the categorization performance of the monkey (Thomas et al. 1999). In other words, population coding was necessary for the monkey to correctly categorize all exemplars in the test set.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

Some cautionary remarks on the "localist model" concept

Richard M. Golden

Cognition and Neuroscience Program, School of Human Development, GR4.1, University of Texas at Dallas, Richardson, TX 75083-0688.
golden@utdallas.edu **www.utdallas.edu/~golden**

Abstract: The notion of a "familiar example" used in Page's definition of a "localist model" is shown to be meaningful only with respect to the types of tasks faced by the connectionist model. It is also shown that the modeling task ultimately dictates which choice of model: "localist" or "distributed" is most appropriate.

For the most part, I applaud Page on an excellent article. He correctly points out that localist representations have unfairly received a bad reputation. Page also correctly notes that localist representations not only can yield insightful and interpretable psychological theories, but neurally plausible theories as well.

In section 2.6, paragraph 2, Page also emphasizes that a model with a "localist representation" is not necessarily a "localist model" and suggests that "a localist model . . . is characterized by the presence of (at least) one node which responds maximally to a given familiar . . . example of that type, . . . all familiar examples of that type being so represented." I agree that a model with a "localist representation" is not necessarily a "localist model," but the notion of "localist model" as developed in the target article needs to be more carefully clarified before it can be practically used. I will illustrate those aspects of the "localist model" concept which I believe require clarification by considering the classic interactive activation (IA) model of context effects in letter perception (McClelland & Rumelhart 1981).

The terminology "pseudoword" in the following discussion will be used to refer to a string of letters that does not form a word yet has perceptual similarities common to words. A "nonword" will be defined as a string of letters that does not form a word and is not a pseudoword. An important finding in the literature concerned with context effects in letter perception is that letters in pseudowords are perceived more efficiently than letters in nonwords. Page refers to the IA model as a localist model and with respect to processing tasks involving the processing of a letter in the context of word (or nonword), I would agree with him. However, consider the case where a letter is processed in the context of a pseudoword. If the pseudoword letter string THIM is presented to the IA model, the word units THIS and SHIM might both become moderately activated, while the word units COME and THAT might be weakly activated. Here, it seems reasonable to consider the pattern of activation over all of the word units (and possibly the letter units as well) as a distributed representation of the concept THIM. That is, the IA model exploits its familiarity with pseudoword concepts represented, not as localistic activation patterns, but as distributed activation patterns.

This example suggests that, in practice, care must be paid to carefully defining the concept "familiar example." If a word is considered to be a "familiar example," then the IA model is a localist model according to the definition in the target article. If a pseudoword is considered to be a "familiar example," then the IA model is not a localist model since it employs distributed representations of pseudowords as well as local representations of words. One might argue that pseudowords are not technically "familiar examples," but the notion of familiarity really needs to be tied to what knowledge is embedded in the system as opposed to the relative frequency of events in the external environment.

Golden (1986) introduced a version of the IA model which was less localist in the sense that it consisted only of units that were "letter position specific feature detectors." That is, each unit was assigned a semantic interpretation of the form "horizontal line segment present in the third letter of the four letter word" or "vertical line segment present in the first letter of the four letter word." In the initial stages of learning, all units responded independently but as learning progressed, Golden's distributed IA model learned to exploit orthographic regularities in letter strings. Golden's model could account for letters being perceived more efficiently in the context of words and pseudowords relative to the context of nonwords.

Would Golden's model be considered a "localist" or "distributed" model? If one considers the "familiar examples" to be "letter position specific features" then Golden's model would be considered localist, but such a definition seems unsatisfying since the model's behavior is guided by its familiarity with the likelihood of particular spatial configurations of "letter position specific features" which make up "letter position specific letters" and "words." Thus, the essential "familiar examples" really are letters and words and in this sense Golden's model would be considered a distributed model.

Furthermore, one could argue that Golden's distributed model is neurally plausible since evidence of retinotopic cortical maps is consistent with the idea of spatial position specific feature detectors. So by this argument one might conclude that here is an example of a distributed model which provides a nice neurally plausible psychological account of how experience with letters in words could give rise to an explanation of the word superiority effect in a way that the original IA model could not.

In summary, I have tried to make two key points. First, the concepts of a "localist model" and a "distributed model" as introduced in the target article are largely dependent upon the notion of a "familiar example." Therefore, the notion of "familiar example" must be carefully considered within the context of the model's behavior. And second, although I agree "localist modeling" assumptions are generally preferable for theory development in psychology, the modeling task ultimately dictates which choice of model: "localist" or "distributed," is most appropriate.

Localist but distributed representations

Stephen Grossberg

Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215. **steve@bu.edu** **www.cns.bu.edu./Profiles/Grossberg/**

Abstract: A number of examples are given of how localist models may incorporate distributed representations, without the types of nonlocal interactions that often render distributed models implausible. The need to analyze the information that is encoded by these representations is also emphasized as a metatheoretical constraint on model plausibility.

Page presents a much-needed analysis of trade-offs between models such as back propagation (BP) which use purely feedforward yet nonlocal interactions, and models such as Adaptive Resonance Theory (ART) which use both feedforward and feedback interactions that obey local constraints. It needs to be emphasized that "localist" models do not necessarily compute winner-take-all categories, even though such categories have powerful computational properties; for example, Carpenter and Grossberg (1991). A key concern is that distributed models such as BP are defined by mechanisms whose information is not locally computed with respect to the underlying network architecture. This is biologically implausible and also hampers their implementation as VLSI chips.

Masking Fields (Cohen & Grossberg 1986; 1987; Grossberg 1978a; 1986) provided one early example of a competitive "localist" network that does not necessarily compute winner-take-all categories. Rather, it is a multiple-scale network that can "weigh the evidence" for representing multiple parts of an input pattern, with the various part representations activated to different degrees. Masking fields were introduced to explain how, under unsupervised learning conditions, an unfamiliar grouping of familiar items can ever overcome the salience of the familiar item representations so that a new representation of the unfamiliar grouping can be learned. This problem arises in both visual object recogni-

tion and speech recognition. A masking field does this by giving the chunks that represent larger groupings, up to some maximal length, a prewired competitive advantage over those that represent smaller groupings. It was shown how this bias could develop from simple developmental growth laws (Cohen & Grossberg 1986). The network clarifies how the most predictive chunk can be maximally activated, while less predictive chunks are less activated, and chunks with insufficient evidence are merely primed. Such a network naturally explains such data as the Magic Number Seven (Grossberg 1978a; 1986; Miller 1956), and predicted data about the word length effect (Samuel et al. 1982; 1983), which shows that a letter can be progressively better recognized when it is embedded in longer words of lengths from 1 to 4. This is the speech analog of the word superiority effect, which it also explains, unlike the Seidenberg and McClelland (1989) model. Masking fields have recently been used, within an ART framework, to quantitatively explain data about how future word sounds can reorganize conscious percepts of earlier word sounds (Grossberg & Myers 1999; Repp et al. 1978). None of the distributed models mentioned by Page can explain these data. More recent developments of ART continue to analyse how a network can automatically discover, through incremental learning in real time, the optimal level of compression with which to represent different input environments.

Page mentions "binding" as one means of generating distributed representations. One mechanism for this is the horizontal connections that exist in neocortex, notably in layers 2/3. Recent modeling work has clarified how bottom-up, horizontal, and top-down interactions interact within the laminar circuits of neocortex, notably visual cortex, to bind together distributed activations into coherent boundary representations (Grossberg 1999; Grossberg & Raizada 1999). This work opens the way toward the very large task of showing how distributed information may be coherently bound in other parts of sensory and cognitive neocortex.

Page notes that both view-specific and view-invariant representations of familiar objects can be found in IT cortex. Such representations have been shown to self-organize in a number of ART-based models; see Bradski and Grossberg (1995) for one such model and related references. A key issue here is that working memories play a useful role in generating these representations. These working memories are "distributed," yet are also clearly localist.

Page quotes the assertion of McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982) that their Interactive Activation (IA) model is a canonical model "that characterizes the qualitative behavior of other models." Actually, the original IA model had serious defects. These defects illustrate that all localist models are not created equal, and that one must exercise as much caution in choosing among them as one does between localist and nonlocal distributed models. In particular, I early noted that the IA model had unrealistic processing levels (phonemes, letters, words) and bottom-up input pathways (both excitatory and inhibitory). These properties were inconsistent with key data, and prevented the model from being able to stably learn from its inputs-even though the authors did not attempt to make the IA model learn (Grossberg 1984; 1987). Later versions of the model changed these properties to be consistent with previously published ART properties; e.g., those in Grossberg (1978a). In this sense, the IA model is dead, and has been subsumed by ART. Problems within models like IA can lead people who prefer nonlocal distributed models to conclude that their models are better. A more proper conclusion is that IA was not an adequate model, localist or not.

Page provides a useful critique of the McClelland et al. (1995) attempt to explain how interactions between the hippocampus and neocortex may control learning and memory. He leaves out at least one issue that I find devastating to all models of this type. Grossberg and Merrill (1996) provide a critique which builds upon this concern. It involves the issue of representation, which is key to all discussions of localist versus distributed coding. In particular, this model proposes that the hippocampus rapidly encodes information which is then later transferred to neocortex. But there is no evidence of which I am aware that the hippocampus can represent the types of information from vision, audition, and so on, that would need to be represented there for this proposal to be plausible. Saying that the information is represented by hippocampus in compressed form does not help, because then one needs to explain how it gets decompressed in the cortex. I am amazed that authors of such models have not bothered to respond to this critique. I hope that it does not take as long as it took the stability-plasticity issues to get discussed which were introduced with ART in 1976.

The Law of Practice and localist neural network models

Andrew Heathcote and Scott Brown

Department of Psychology, The University of Newcastle, Callaghan, 2308, NSW, Australia. **{heathcote; sbrown}@psychology.newcastle.edu.au psychology.newcastle.edu.au/**

Abstract: An extensive survey by Heathcote et al. (in press) found that the Law of Practice is closer to an exponential than a power form. We show that this result is hard to obtain for models using leaky competitive units when practice affects only the input, but that it can be accommodated when practice affects shunting self-excitation.

In a recent survey, Heathcote et al. (in press) analyzed the form of the Law of Practice in 7,910 practice series from 475 subjects in 24 experiments using a broad range of skill acquisition paradigms. When the practice series were not averaged over subjects or conditions, an exponential function (mean response time, $RT = A + Be^{-\alpha N}$, where A is asymptotic RT, B is the amount that learning decreases RT, and N is practice trials) provided a better fit than a power function ($RT = A + BN^{-\beta}$) for the majority of cases in every paradigm. The defining property of an exponential function is that its relative learning rate, $RLR = -dRT/dN/(RT - A)$ equals a constant (α). In contrast, the power function's RLR decreases hyperbolically to zero, $RLR = \beta/N$. Previous findings in favor of a power function (e.g., Newell & Rosenbloom 1981) used practice series averaged over subjects and/or conditions. When exponential practice series with different rates (α) are averaged, the RLR of the average decreases, because fast learners (with large α) control the rate of change early in practice, while slow learners (with small α) dominate later in practice (see Brown & Heathcote, in preparation, for detailed analyses of averaging effects). As theories of skill acquisition model the behavior of individuals, not averages, Heathcote et al. concluded that the "Law of Practice" is better characterized by an exponential than a power function. Hence, the power function prediction made by Page's model does not accord with recent empirical results.

We believe that an exponential law of practice is extremely difficult to obtain using Page's approach to practice effects in competitive leaky integration networks (Equation 5). To see why, consider the time (t) it takes the activation ($x(t)$) of a leaky integrator ($dx/dt = I - kx$, where I is input and k is leakage rate and $x(0) = 0$) to reach a criterion χ.

$$t = \frac{1}{k}\ln\left(\frac{I}{I - k_\chi}\right) \tag{1}$$

The RLR of (1) with respect to I decreases to zero. If we assume, as Page does, that practice decreases t by increasing I, the RLR of (Eq. 1.) with respect to N will decrease to zero unless $I(N) \geq O(N^2)$ for large N. Such a faster than linear increase in input is difficult to justify. The increase of I with N is slower than linear for Page's "noisy-pick-the-biggest" model. Even if all instances, rather than just the maximally activated instance, were to contribute to

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

I, the increase would be only linear. Page's simulation results (Fig. 6) indicate that the same power-like effects of increasing *I* apply to the time it takes competing leaky integrators to pass an activation criterion.

However, competitive leaky integrators can account for Heathcote et al.'s (in press) findings if practice alters shunting terms, such as the weights of self-excitatory connections.[1] Consider a two-unit system of the type discussed by Usher and McClelland (1995), with normalized inputs *I* and $(1 - I)$ and linear threshold transfer functions:

$$dx_1/dt = I - (k - \epsilon)x_1 - \delta x_2 \quad (2)$$

$$dx_2/dt = 1 - I - (k - \epsilon)x_2 - \delta x_1 \quad (3)$$

A response is made when the activation of one unit exceeds a criterion, χ. Assume that as practice proceeds, the self-excitatory weight, ϵ, approaches the leakage rate *k*, using a weight-learning rule like Page's Equation 2:

$$d\epsilon/dN = \lambda(k - \epsilon) \quad (4)$$

In simulations with Gaussian noise added (Eq. 2, 3) at each step of the integration (Page's N_1 term in his Eq. 5) and larger values of *I* so errors did not occur, learning series were consistently better fit by an exponential than by a power function. Insight into this result can be gained from the analytic result for the one unit case (i.e., Eq. 2 with competitive weight, $\delta = 0$, which was also better fit by the exponential in simulations):

$$t = \frac{1}{k} e^{\lambda N} \ln\left(\frac{I}{I - k_\chi e^{-\lambda N}}\right) \quad (5)$$

For a linear Taylor approximation to (Eq. 5), *RLR* decreases marginally with *N*, but asymptotically approaches λ rather than zero. Heathcote et al. (in press) found that an APEX function ($RT = A + Be^{-\alpha N}N^{-\beta}$), which has a *RLR* that decreases to an asymptote greater than zero, consistently fit slightly better than an exponential function. We found the same pattern of fit to our simulation results for both the one and two-unit models. The parameter estimates for these fits also concurred with the survey results. Estimates of the power function *A* parameter were implausibly small (as *N* increases *t* approaches χ/I for the linear Taylor approximation to [Eq. 5], whereas most power function *A* estimates were zero). Fits of a power function with an extra parameter (*E*) to account for prior practice ($RT = A + B(N + E)^{-\beta}$) produced implausibly large *B* estimates, mirroring Heathcote et al.'s (in press) findings with the survey data.

Given limited space it is not possible to quantitatively examine this type of model further (see Heathcote 1998, for related findings and Heathcote & Brown, in preparation, for a detailed analysis). However, the findings presented are sufficient to demonstrate that Heathcote et al.'s (in press) results are not incompatible with the overall localist neural network approach. Indeed, learning in shunting connections, both self-excitatory and competitive, provides an adaptive mechanism for consolidating and differentiating local response representations (cf. Usher & McClelland 1995, who note that the "units" in such models may correspond to collections of neurons bound together by mutually excitatory connections). Reduced leakage with practice can also explain Jamieson and Petrusik's (1977) finding (cited in Usher & McClelland 1995) that the difference between error and correct RTs decreased with practice. As leakage approaches zero, a leaky integrator approximates a classical diffusion process, for which error and correct RTs are equivalent.

NOTE

1. We also obtained an exact exponential result for inputs that (1) increase with practice according to a learning rule like Page's Equation 2 ($I = M(1 - e^{-\lambda N})$), (2) are nonstationary (decreasing with presentation time *t*, as $I = 1/(t + (b - cI))$, $b/c >$ M), and (3) have a shunting effect on a single unit's activation ($dx/dt = (U - x)I$). We will not pursue this model here, as it is very different from Page's approach (see Heath 1992, and Smith 1995, for more on nonstationary inputs, and Heathcote 1998 for more on shunting inputs).

Localism as a first step toward symbolic representation

John E. Hummel

Department of Psychology, University of California, Los Angeles, CA 90095.
jhummel@lifesci.ucla.edu **www.bol.ucla.edu/~hummel/**

Abstract: Page argues convincingly for several important properties of localist representations in connectionist models of cognition. I argue that another important property of localist representations is that they serve as the starting point for connectionist representations of symbolic (relational) structures because they express meaningful properties independent of one another and their relations.

Page's arguments and demonstrations make a compelling case for the essential role of localist representations in connectionist models of cognition (and cognition itself). One important property of localist representations that Page does not emphasize (although he mentions it at the end of sect. 7.3), concerns the role of localist nodes in the representation of relational structures. I argue that localist representations share a crucial property with the kinds of representations that are necessary for relational representation in connectionist systems – namely, independent representation of meaningful entities – and that they therefore play an essential role in the ability of connectionist models to account for symbolic aspects of cognition.

The notion of a "localist representation" is subtle because localism is not a property of a representation, but of the relationship between a representation and the entities it represents. To borrow Page's example, the activation pattern *−woman, +politician,* and *−actor* is a distributed representation of Tony Blair, but a local representation of *woman, politician,* and *actor.* Every representation is local at some level. Even a "fully distributed" representation is localist with respect to some entities, in that each node has an equivalence class of entities to which it corresponds. The equivalence class may be difficult or impossible for the modeler to understand (as in the case of the hidden nodes in many BP networks), but unless a node is always active (in which case it carries no information), its activity will correspond to some state of affairs in the network's universe: The node is a localist representation of that state of affairs. As such, the important question is not whether a representation is localist or distributed, but whether it is localist *with respect* to a meaningful state of affairs in the network's universe.

In this sense, the question of localist versus distributed maps onto the question of *independence* (a.k.a., *separability;* Garner 1974) versus *nonindependence* (a.k.a., *integrality*) in mental representation. If meaningful concepts, entities or dimensions map onto individual nodes (or in the case of dimensions, nonoverlapping populations of nodes) – that is, if the system is localist with respect to those entities or dimensions – then the system represents those entities as independent of one another. To the system, the entities or dimensions are separable (cf. Cheng & Pachella 1984). If individual nodes respond to *conjunctions* of entities or properties, then the resulting representation is integral with respect to those properties (e.g., nodes that respond to specific conjunctions of shape and color constitute an integral representation of shape and color). One hidden limitation of many "fully distributed" representations (e.g., those that emerge in the hidden layers of BP networks) is not only that they lack individual nodes to respond to individual entities (the limitation Page emphasizes), but also that they typically constitute integral, rather than separable representations of the important entities or properties in the network's universe.

This limitation is important because separability is crucial for generalization (cf. sect. 7.3). A network will generalize with respect to what it represents, so if its units do not represent meaningful entities, the network will not generalize with respect to meaningful entities. Consider the simplified example of a network learning to categorize colored shapes, and let categories be defined by color, with shape irrelevant but free to vary. If the network represents color independent of shape, then learning to place, say, all red objects into category A and all blue objects into category B is a simple matter of learning connections from the node(s) representing "red" to the node representing "category A," and from the node(s) representing "blue" to the node representing "category B." Once these connections are in place, learning will generalize to any new red and blue objects, regardless of their shapes: the network will generalize the functions *red(x)* → *category-A(x)* and *blue(x)* → *category-B(x)* universally. A network that violates color-shape independence will not generalize universally. If nodes represented, for example, conjunctions of shapes and colors, then categorizations learned in the context of one set of shapes would not necessarily generalize at all to novel shapes.

The role of separability in generalization makes localism (or at least a form of localism) a necessary ingredient in connectionist representations of relational structures, such as propositions and variablized rules (cf. Holyoak & Hummel 2000). To appreciate what "John loves Mary" has in common with "Mary loves John" a representational system must be able to represent people independent of their relations. Similarly, to represent a variablized rule (such as *for any x, red(x)* → *category-A(x)*), a network must be able to represent variables independent of their values. Of course, localism by itself is not sufficient for symbolic connectionist representations. Representing variables independent of their values (or roles independent of their fillers) makes it necessary to actively bind them together (e.g., by synchrony of firing; see Hummel & Holyoak 1997). But localism – in the sense of placing units into correspondence with meaningful equivalence classes in the network's universe – is a necessary first step.

ACKNOWLEDGMENT

Preparation of this commentary was supported by NSF Grant 9729023.

Integrating exemplars in category learning: Better late than never, but better early than late

J. Eric Ivancich,[a] David A. Schwartz,[b] Stephen Kaplan[a,b]

[a]*Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109;* [b]*Department of Psychology, The University of Michigan, Ann Arbor, MI 48109.* **ivancich@eecs.umich.edu {david.a.schwartz; skap}@umich.edu**

Abstract: Page's target article makes a good case for the strength of localist models. This can be characterized as an issue of *where* new information is integrated with respect to existing knowledge structures. We extend the analysis by discussing the dimension of *when* this integration takes place, the implications, and how they guide us in the creation of cognitive models.

In years past, these pages witnessed a heroic struggle between symbolists and connectionists. The fact that we now find different schools of neural network modelers challenging each other's assumptions within the same pages certainly signals a growing acceptance of the connectionist perspective. In this context Page does the connectionist community a great service by demonstrating the power of localist models and how they address specific weaknesses in models that rely strictly on distributed representations. And Page is right in reminding us that localist models predate distributed models. We applaud this effort, and wish to suggest another dimension that could help localist models handle cognitive phenomena more effectively.

Dimensions of where and when. In one sense, the distinction between distributed models and localist models centers on *where* new information is integrated into existing knowledge structures, that is, whether new information has distributed effects or localized effects. This distinction in integration will also affect how the network will work. When a given concept is being used, will activity be distributed widely or narrowly?

Another important dimension in the analysis of forms of knowledge integration concerns whether new information is integrated directly into existing units of knowledge, or whether this integration happens later, perhaps not even structurally but in the process of using the network. This dimension focuses on *when* this integration takes place. We will use the terms *early integration* and *late integration* to describe extremes on this continuum.

An extreme example of a late integration model would be one in which the network records each individual exemplar encountered. The very processes of generalization, so fundamental to cognition, would thereby occur late in the game, perhaps not until the knowledge is used. In this extreme case, a category would not exist as an identifiable structure, but instead be represented by prior exemplars. When a sensory pattern is encountered, a process of scanning and aggregating the exemplars would take place in order to determine the category. Other, nonsensory knowledge related to the category would also have to be aggregated through a similar scanning process. Clearly this is an extreme case, if not a caricature, of late integration, and the problems are readily apparent.

By contrast, a model that relies on early integration would start the formation of a category – an identifiable structure in the network – as soon as possible. As new exemplars are encountered the categorical knowledge is modified in place. Rather than growing ever larger, the categorical knowledge would undergo refinement through learning. The sensory features most commonly found in the category's exemplars would be reinforced, while the rare and incidental would be discarded. In other words, not only does the categorical structure come to emphasize the general properties of the category, its signal/noise ratio is improved in the process.

Page's generalized localist model has some flexibility along this early to late integration continuum, given its underlying architecture and parameters. The vigilance parameter affects whether existing knowledge is determined adequate and thus whether any learning needs to take place. The higher the vigilance parameter, the more likely a new exemplar will be recorded, and the later any integration is forced to occur. The learning rate parameter determines whether few or many exemplars contribute to a single node. The fewer exemplars contributing to a node, the later the integration.

While having a generalized model can be useful in showing breadth of possible behaviors, in the challenging enterprise of understanding the brain it is perhaps more important to narrow down the space in which to focus our efforts. And in the process of applying this focus, we may find that the generalized model may not be so easily adapted. We will now discuss three areas of psychological considerations that highlight some advantages of early integration. We conclude by describing how one of the earliest localist models still holds remarkable potential in these respects.

Prototype effects. There is much evidence that human categories display a number of prototype effects (see Rosch 1977 for a nice summary). These effects would seem unlikely to emerge from a late integration process, and it is far easier to account for all of them if a structure is created that encodes the prototype directly.

The prototype refers to the encoding of the central tendency of a category along with some measure of its breadth. It thus requires time and a multitude of exemplars to form. Once it forms, the prototype plays a key role in numerous psychological phenomena.

For example, people easily recognize prototypes even if they have never been encountered (Posner & Keele 1968). People will rate an unencountered prototype as more likely to have been previously encountered than a nonprototypic exemplar actually encountered (Bransford & Franks 1971). People more quickly recognize exemplars closer to the central tendency (Rosch 1975). Moreover, certain psychological phenomena suggest a structural reality underlying prototype effects. People fill in missing details of an exemplar its category's prototype (Bruner 1973). In a similar phenomenon, people, when recalling a specific exemplar, will shift it towards the prototype (Bruner & Minturn 1955).

Differentiation or enrichment. If a prototype is ultimately encoded in the network directly, it then becomes important to understand the process by which integration would create it. Many have argued that perceptual learning of a category is better viewed as a process of enhancing discrimination rather than simply as an additive process (e.g., Gibson & Gibson 1955; Hall 1991).

It is often the case that early attempts at categorization are overly broad and that over time perceptual categorization becomes more discriminatory (Mackintosh et al. 1991). Gibson and Gibson offer the example of developing the ability to distinguish different varieties of wine.

Clearly, features most common to the exemplars must be directly encoded within a prototype structure. It is also important, however, that features that are rare or incidental not hold the prototype hostage; they must be minimized. An additive process, which late integration favors, has greater difficulty in accounting for this process of subtraction.

Category drift. Categories change over time. Take, for example, an individual's face or a specific street corner. With late integration an ever growing number of exemplars will be recorded over an extended period of time. With early integration, however, the category will update and shift in place, through the aforementioned process of refinement. This process allows the prototype to track the environment's changes. Moreover, late integration suffers from the costly impact of perpetually needing to allocate new areas to remember new exemplars.

Hebb's cell assembly. The route we suggest to achieving all this is, perhaps surprisingly, a historical one. As our final point, we would like to remind the connectionist community of a specific localist model of neural structures that has been in existence since the 1940s. Despite its important advantages in several crucial domains, it has largely been ignored.

Donald Hebb (1949), created the notion of the cell assembly, a learned conglomeration of neurons that act in concert as a unit representing a category. While many in the connectionist field use and cite Hebb's learning rule, few (although it appears to be a growing number) refer to or use the cell assembly. The cell assembly is a neural structure that responds to a set of exemplars. It is a unitary structure that encodes the variety of exemplars encountered. And if constructed appropriately, it will display prototype effects.

Hebb's cell assembly is intimately connected with his learning rule. We believe, however, that his learning rule is incomplete, and that it requires a synaptic weakening component (examples are discussed in Hetherington & Shapiro 1993). The synaptic weakening component provides a mechanism for Hebb's fractionation concept. Fractionation makes possible the simplification and differentiation necessary to develop a prototype in the process of early integration. And finally, new information can always shape a cell assembly, allowing it to drift in parallel with its real-world correlate.

Hebb's cell assembly is a localist model that offers a number of distinct advantages. Our discussion, though, is not intended to detract from the central thrust of Page's analysis. Dispensing with fallacies regarding localist models is an important element in the debate within connectionism. And if the debate shifts from being primarily between localist and distributed models towards the possibilities within the space of localist models, much will have been accomplished.

ACKNOWLEDGMENT

We would like to thank Mark Weaver and the other members of the Seminar on Environmentally Sensitive Adaptive Mechanisms for their helpful comments.

Instance-based manifesto?

Márk Jelasity

Research Group of Artificial Intelligence, József Attila University, Szeged H-6720, Hungary. **jelasity@inf.u-szeged.hu**
www.inf.u-szeged.hu/~jelasity/

Abstract: Page's definition of localism is inspired by the instance-based paradigm. However, the locality of *representations* is not necessary for a model to be instance-based and, on the other hand, explicit featural representations are generally considered local. The important distinction is between instance-based and noninstance-based paradigms and not between distributed and local representations as Page claims.

Page's discussion of localist models in section 2.6 and his list of references makes it clear that when giving his definition he had the instance-based paradigm in mind. His localist model supports this interpretation, because its knowledge representation scheme is instance-based (weights of output units), its learning method is a version of vector quantization, and its decision method is in fact the simple first-nearest-neighbor algorithm with some noise added (Mitchell 1997).

Though Page concentrates on psychological modeling where delicate details are important, I would like to comment on his article from the more general standpoint of artificial intelligence and mathematics. The problem that connects these two fields is catastrophic inference. From a mathematical point of view, all learning methods are function approximators. Given the exemplars of some concept (i.e., a subset of a general category), they form a function that can classify all possible inputs. There are differences between the methods but practically every method can be applied to every problem or every modeling task in psychology; there is nothing in principle that could prevent this free applicability and, indeed, this is what we see happening in practice. (For example, Page demonstrates that his instance-based model can indeed explain different phenomena.) However, there is a major borderline between instance-based methods and others: the amount of knowledge stored by the instance-based models can be increased any time while the latter forget everything when taught new examples; this is the price to pay to have fast and effective algorithms. Instance-based methods typically run slower and need much more memory but they are flexible and extendable.

Based on the above observations my (simplified) interpretation of Page's suggestion is that – since in psychological modeling computational efficiency is not very important – we should choose instance-based models. If this interpretation is correct (which I shall assume henceforth), I agree with the suggestion. My problem is that Page handles the concept of local representation and instance-based modeling equally, which is unfortunately not true. The dimension of the distributed/nondistributed nature of representation and the dimension of the instance-based/noninstance-based nature of the paradigm are independent.

To show this independence in a brief technical paragraph, I will sketch the basic idea of a truly distributed implementation of an instance-based memory model. Let us represent our learning samples as vectors of features. Let the values of features be real numbers. Now, let the representation of our knowledge be the sum of these vectors. The recognition operation can simply be taken as the following: if the dot product of the test sample (presented as a vector) and our memory (i.e., the sum of all exemplars) is small then we say the sample is not in the memory; if it is large, then we say it is in the memory (i.e., it is recognized). A mathematical assumption is necessary in order to make the model func-

tion properly: the vectors involved must be pairwise (approximately) orthogonal. The number of vectors stored is limited as a function of their length (i.e., the number of features). This model has most of the properties of usual instance based models, e.g., it does not suffer from the catastrophic inference problem, since one can always add new vectors to the memory (until its capacity is exceeded). In this framework, generalization and other phenomena can be modeled too. The model is for illustration only. More sophisticated models based on linear algebraic properties of vectors and matrices exist, see for example (Kohonen et al. 1981; Murdock 1982; Pike 1984).

To summarize: the two dimensions of modeling (the distributed/nondistributed nature of representation and the chosen paradigm) are independent. The instance-based paradigm does not exclude distributed implementations, whereas – unlike Page – many of us would consider, e.g., explicit featural representations (in models that are not necessarily instance-based) localist. Finally, the debate is certainly not confined to the field of connectionism, although instance-based models can have connectionist implementations (or visualizations) as in Page's model. Psychological models and learning algorithms should be classified at a more abstract level. The properties of the actual units do not necessarily reflect the behavior of a model in certain situations. I think – clarifying the terminology – "Modelling in psychology: An instance-based manifesto," would have been a better title for what Page may have wanted to say.

ACKNOWLEDGMENTS

I would like to thank George Kampis and Csaba Pléh for their valuable contributions to this work.

The elementary units of meaning

Paul J. M. Jorion

Théorie et Praxis, Maison des Sciences de l'Homme, 75270 Cedex 6, Paris, France. **paul_jorion@email.msn.com** **aris.ss.uci.edu/~jorion**

Abstract: Examining the implications of a localist model for linguistic performance, I show the strengths of the P-graph, a network of elementary units of meaning where utterance results from relaxation through the operation of a dynamics of affect values. A unit of meaning is stored in a synaptic connection that brings together two words. Such a model, consistent with the anatomy and physiology of the neural tissue, eschews a number of traditional pitfalls of "semantic networks": (1) ambiguity ceases to be an issue, as similar uses of a word are automatically clustered together; (2) faster retrieval of words acquired early is explained by the larger number of their instances. In addition the P-graph takes advantage of a plausible form of information storage: the local topology of the neural tissue.

The elementary units of meaning. The trouble with localist hypotheses is that one is making oneself vulnerable to embarrassing questions such as "All right, one instance of the word 'cat' is associated with a single neuron. So, how do you do this?" Ten years ago I offered a localist model for knowledge representation (Jorion 1989); I evaded the issue of implementation. I still do today. If on the contrary one claims that "The acoustic imprint 'cat' emerges in a holographic manner from the collaborative effort of a set of neurons associated with the word 'cat'" you are spared the thorny issue of "how?" as all will consent that such complex matters will be dealt with . . . in due course.

The stumbling block remains the physiology/chemistry of the storage process of memory traces and until someone comes up with an explanation, or at least a valid testable model, researchers will remain reluctant to commit themselves to localist models. As shown by Page, those of us who believe in their ultimate vindication need to keep building a body of converging evidence. Part of the process consists of refuting commonly assumed obstacles and other misconceptions.

Because I am taking exception to one of Page's statements in this issue, I will sketch how to flesh out a localist approach to linguistic performance. In my 1989 conference paper (Jorion 1989) and in further developments in my book (Jorion 1990) I put forward a template for linguistic performance (both speech and writing) as a dynamics operating on a network of memory traces made of individual words. This P-graph model differs from the classical "Quillian" semantic network that contains, in addition to words as acoustic or visual imprints, "meta-information" in the shape of speech part categorization or syntactic cues. It assumes nothing more than words being stored and being activated by a dynamics of affect values leading to relaxation through word utterance. Words are represented as vertices of a graph and their relationships as nodes: it is the dual graph of the traditional semantic network where words are nodes and relationships between them, vertices.

The P-graph model has specific strengths: (1) it is consistent with the currently known properties of the anatomy and physiology of the nervous system, (2) because the P-graph is the dual of a classical semantic network, (2.1) a word is automatically distributed between a number of instances of itself; (2.2) these instances are clustered according to individual semantic use, (2.3) the scourge of knowledge representation, ambiguity, is automatically ruled out; for example, kiwi the fruit and kiwi the animal being associated only through one relationship, the "superficial" (meaningless) one of homophony, their confusion does not arise: they reside in distant parts of the P-graph, and (3) the growth process of the graph explains why early word traces are retrieved faster than those acquired later: their number of instances is of necessity large because they have acted repeatedly as an "anchor" for the inscription of new words in the process of language acquisition. This allows us to do without Page's extraneous hypothesis that "a node of high competitive capacity in one time period tends to have high competitive capacity in the next" (sect. 4.4).

The loss of synaptic connections that accompanies knowledge acquisition corresponds to increased organization of the neural tissue (negentropy). This also means that the topology of the surviving connections carries significant information. As long as we remain ignorant of how memory traces are stored in the brain, that is, as long as we are unable to read the information encapsulated within the neural tissue, its very configuration is bound to appear meaningless, as if the pattern of its existing connections were haphazard. I claimed in my 1989 paper that it would be negligent for evolution to ignore a highly economical mechanism for information storage such as the topology of the neural tissue (Jorion 1989). I also showed how a simple rule for neuron colonization, that is, a "has a . . ." relationship establishing a symmetrical connection between two words (undirected graph) and an "is a . . ." relationship, an asymmetrical one (directed graph), inscribes topological information into a P-graph while ensuring redundancy in the representation of any individual word, confirming Page's insight that "localist models do not preclude redundancy" (sect. 6.3).

The "paradigm shift" necessary to implement the P-graph alternative to a classical semantic network is that neuron nuclei cease to be the locus of choice for information storage, synaptic connections between two neighboring neurons emerging as stronger candidates. The elementary structure of meaning is no longer the stand-alone, self-contained concept, but instead the relationship connecting two concepts (rat-rodent; mommy-baby bottle; piano-horrible); a view reminiscent of Aristotle's *categories* (cf. Jorion 1996).

In this new perspective the meaning of a word equates with the composite picture emerging from the set of associative units of two to which it belongs (see Jorion 1990, p. 88). This conception is remarkably close to Wittgenstein's view of the meaning of a word as the resultant of its possible uses: "The meaning is the use." Resemblance is no longer only a question of distances measured over a neural network, it also covers topological similarities. For example, synonyms do not need to be stored physically close to each other in the brain (indeed it is unlikely they would be, as syn-

onyms are typically acquired at different times in life rather than simultaneously) as long as they are part of isomorphic configurations of elementary units of meaning. Topological similarity may suffice for harmonics to develop between homological subnetworks, allowing synonyms to vibrate in unison.

Can we do without distributed models? Not in artificial grammar learning

Annette Kinder

Department of Psychology, Philipps-University, D-35032 Marburg, Germany.
kinder@mailer.uni-marburg.de

Abstract: Page argues that localist models can be applied to a number of problems that are difficult for distributed models. However, it is easy to find examples where the opposite is true. This commentary illustrates the superiority of distributed models in the domain of artificial grammar learning, a paradigm widely used to investigate implicit learning.

In his target article, Page impressively demonstrates the explanatory power of localist models. Given the many problems of distributed models, Page's conclusion that localist models should be preferred over distributed ones seems to be totally justified. However, although it is true that various psychological processes can be successfully simulated by localist networks (even more successfully than by distributed ones), there are other processes for which localist models are not adequate. One example is implicit learning, which recently has become a popular research topic in cognitive psychology. In implicit learning, knowledge about complex sequential material is acquired under incidental training conditions. A paradigm widely used to investigate implicit learning is artificial grammar learning (AGL). This commentary illustrates how the existing localist models fail to account for AGL and why it would be difficult to conceive more adequate localist models.

In AGL, strings of letters (or other symbols) are presented that were generated according to an artificial grammar. This kind of grammar comprises a complex set of rules which constrain the order of the letters in the strings. In an AGL experiment, participants first have to memorize a subset of all strings which can be generated by a particular grammar, the so-called grammatical strings. Only after this training stage is over, they are informed about the existence of the rules. Subsequently, they are asked to categorize new grammatical and nongrammatical strings as following or violating these rules, which they normally can do well above chance level. It has been shown that participants use several different sources of information to accomplish this task (Johnstone & Shanks 1999; Kinder 2000; Knowlton & Squire 1996). AGL is considered as a typical example of implicit learning because the learning conditions are incidental and the acquired knowledge seems to be difficult to verbalize.

Two types of connectionist models of AGL have been proposed, a simple recurrent network (SRN) model and several autoassociator models. Whereas most autoassociator models of AGL can be described in purely localist terms, the SRN model mainly relies on distributed representations. An SRN (Elman 1990) comprises a minimum of four layers of processing units, an input layer, a hidden layer, an output layer, and a context layer. The context layer contains a copy of the hidden layer's activation pattern on the last stimulus presentation. Because of this feature, the network can learn to predict a stimulus in a sequence not only from its immediate predecessor but from several stimuli presented before. Although the SRN is not fully distributed because it also contains localist representations, it is closer to the distributed end than to the localist end of the scale: It learns by backpropagation, and the representations of the stimulus sequence, which are the crucial ones in an SRN, are distributed.

When AGL is simulated in an SRN, the network is trained with the same stimuli as the participants (Dienes et al. 1999; Kinder 2000). The letters of each string are coded one at a time, from left to right in the input layer. The network is trained always to predict the next letter in the string. That way, the network gradually becomes sensitive to the absolute and conditional probabilities of letters within the set of training strings. If a network trained that way is tested subsequently, it is capable of discriminating between grammatical and nongrammatical strings. More important, it makes correct predictions about the kind of information participants use at test. For example, it predicts that participants will endorse test items comprising a higher number of familiar string fragments more readily than those comprising a lower number of familiar fragments (Dienes et al. 1999). Furthermore, it correctly predicts that a test string's similarity to a specific training item will influence its endorsement rate only to a very small extent (Kinder 2000).

The only existing localist models of AGL are autoassociator models (Dienes 1992). These models contain a single layer of units each of which represents a letter in a particular position. Every unit is connected to every other unit except to itself. During training, the connection weights are changed in such a way that the network reproduces its input activation pattern as accurately as possible. Since all units are connected to all other ones, the autoassociator models erroneously predict that participants are capable of learning relations between two distant letters even if the intervening letters are not systematically related to either of them (St. John & Shanks 1997; Redington & Chater, submitted). Furthermore, they falsely predict that information about entire training items will be important in grammaticality judgments (Kinder 2000). Both problems could be solved by allowing only associations between representations of adjacent letters. However, this would lead to the false prediction that participants learn information only about letter bigrams. Although it might be possible to conceive a model in which all of these false predictions are avoided, rather complicated assumptions about the impact of spatial vicinity on weight change would have to be made. By contrast, such assumptions are not necessary in the SRN model. As a result of the network's architecture, its learning mechanism, and the fact that letters are presented one after another, spatial vicinity influences the network's behavior quite naturally.

To summarize, there is neither an acceptable localist alternative to the SRN model of AGL nor could such an alternative be conceived easily. In another type of implicit sequence learning (e.g., Cleeremans & McClelland 1991), SRNs are the only existing models. Thus, when we try to explain the (implicit) acquisition of sequences, we cannot do without distributed models.

Localist network modelling in psychology: Ho-hum or hm-m-m?

Craig Leth-Steensen

Department of Psychology, Northern Michigan University, Marquette, MI 49855-5334. **clethste@nmu.edu**

Abstract: Localist networks represent information in a very simple and straightforward way. However, localist modelling of complex behaviours ultimately entails the use of intricate "hand-designed" connectionist structures. It is, in fact, mainly these two aspects of localist network models that I believe have turned many researchers off them (perhaps wrongly so).

From a cognitive modeller's perspective, localist network modelling makes life easy and makes life hard. It makes life easy because the representational properties of localist network models are so well and easily defined. The modeller needs only to separate the representation underlying a cognitive process into its higher and lower level featural components and then to superimpose those components on the (hierarchically arranged) units of a localist network. The task of modelling then involves determining the appropriate set of connections (in terms of both their structure and

their strengths) between the units that, along with the appropriate set of activation processing assumptions, will result in a successful simulation of the behaviour in question. With respect to actual localist modelling endeavours, these behaviours typically include some form of recognition, recall, identification, or categorization performance within either a memory or a perceptual/semantic classification task; behaviours which Page shows are well served by being modelled within a localist network scheme. It is when the modeller tries to simulate more complex kinds of behavioural phenomena (e.g., see the Shastri & Ajjanagadde 1993 connectionist model of reasoning; and the "back end" of the Leth-Steensen & Marley 2000 connectionist model of symbolic comparison) that life becomes hard. Ultimately, localist attempts to solve the modelling problems invoked by complex behaviours entail the use of intricate "hand-designed" connectionist structures. Although I think that most would (should?) agree that designing such structures is a useful scientific exercise in and of itself, there are likely to be many different ways to design a set of connections between localist representations that will solve such problems, but only a limited number of ways in which they are "solved" by the brain in actuality.

It is my belief that a general feeling of disdain for localist network modelling has arisen within some members of the cognitive community for two basic reasons. First, it is precisely this hand-designed, or *a priori* defined nature of most of the available localist network models that has turned many researchers off them. Cognitive modelling using distributed network representations almost invariable begins from the very strong position that the learning of those representations is always part of the modelling enterprise. One is never in the position of staring at a set of network connections and their weights and wondering how they got to be that way. In this article, Page goes a long way toward presenting a very stimulating, generalized approach to the learning of localist representations; one that one hopes can be appreciated by both novice and expert network modellers alike. However, because this approach is applied here only to fairly simple kinds of two- and three-layer network modules, it is not immediately clear how it can then be *scaled up* in order to generate more complex localist structures such as that of the commonly used interactive activation (IA) model (particularly with respect to the seemingly necessary inclusion of multiple localist instances).

Second, there seems to be a generalized intellectual bias against connectionist models that use strictly localist representations in favour of those using distributed ones precisely because of the simple and straightforward way in which information is represented within localist networks. In other words, because the brain is such a complex organ, the belief (hope?) is that neural representation just can't be that simple. For example, as expressed by M. I. Jordan (1990 – to whom, however, I do not presume to ascribe any such bias) – "[distributed] networks with hidden units are worthy of study . . . because their structural and dynamic properties are more *interesting* [emphasis added] than those of simpler systems" (p. 497). With respect to actual behavioural phenomena, however, networks with hidden units also tend to have so much processing power that they often can substantially outperform the learning and processing capabilities of humans and, hence, must then be scaled down to more localist-like systems in order to capture human performance (e.g., see Lacouture & Marley's model of absolute identification, 1991).

In conclusion, let me observe that as intellectually tantalizing as distributed representational systems are, all of us still owe it to behavioural and brain science to fully look into and *exhaust* the possibility that the majority of representation within the brain might just be as simple and straightforward as localist models presume it to be.

Localist representations are a desirable emergent property of neurologically plausible neural networks

Colin Martindale

Department of Psychology, University of Maine, Orono, ME 04469.
rpy383@maine.maine.edu

Abstract: Page has done connectionist researchers a valuable service in this target article. He points out that connectionist models using localized representations often work as well or better than models using distributed representations. I point out that models using distributed representations are difficult to understand and often lack parsimony and plausibility. In conclusion, I give an example – the case of the missing fundamental in music – that can easily be explained by a model using localist representations but can be explained only with great difficulty and implausibility by a model using distributed representations.

For reasons that I have never understood, localist representations are virtually taboo in neural-network theory even though as Page points out in the target article, they offer some distinct advantages over distributed representations. One very big advantage is that people can understand connectionist theories that employ localist representations. Even if such representations served no purpose, they would be a convenient fiction. As far as I know, I am the author of the only two undergraduate cognition texts that consistently use a neural-network approach. Martindale (1981) was written because I felt that students should be presented with a state-of-the-art text rather than one using the thoroughly discredited computer metaphor of mind. I used localist representations, as they were not frowned upon at that time. In the following years, I tried presenting some models using distributed representations in my lectures and was confronted with a sea of completely confused faces. Martindale (1991) also uses localist representations, although I did say in the preface that this was a simplification and that the "grandmother" node was just a stand-in for a pattern of activation across a field of nodes. (I did not fully believe that but was hoping that instructors would see the comment and adapt the book.)

Students are not the only ones who do not understand models using distributed representations. Unless they are experts, professors do not understand them either. Several years ago, I wrote a chapter in which I set forth a model of creativity using distributed representations in a Hopfield (1984) net and simulated annealing (Hinton & Sejnowski 1986) to avoid local minima (Martindale 1995). A year or so later, I was asked to give an invited address on creativity; however, the invitation was contingent: "don't talk about that neural-network stuff, because no one understands it." I have just finished a chapter on neural networks and aesthetic preference (Martindale, in press). I wanted to use the model that I had used to explain creativity, but I also wanted people to understand me. Thus, I ended up using a neural-network model with localist representations. This model could easily be translated into a Hopfield energy-minimization model, but the latter would be very difficult for most psychologists to understand. Professors are not stupid, but they have a passion for parsimony and an aversion to mathematical gymnastics that complicate rather than simplify matters.

In general, connectionist theorists use a brain *metaphor*. We do not usually put forth models where one node corresponds to one neuron, because we don't know enough about how neurons work. A node is thought of being like a neuron or made up of a group of neurons. Theorists who use distributed representations want to stop one step short of saying that this distributed set of neurons will be referred to as the grandmother node. We all agree that nodes are made up of distributed sets of neurons. Why stop at feature nodes? As Page points out, those who suggest localized representations are always confronted with the fact that no one has ever found a grandmother neuron. This is not a relevant question, as we are talking about nodes rather than neurons. However, cells

fairly close to grandmother neurons have in fact been found. Young and Yamane's (1993) "particular-Japanese-man" neuron is an example. Logothetis et al. (1995) showed monkeys a variety of objects. Single-cell recordings from the inferotemporal cortex were made. Most neurons responded to only a given object shown from a particular viewpoint. However, a small portion responded in a view-invariant fashion to the objects to which they had been tuned. This should not be surprising.

Any neurologically plausible neural network (several layers of nodes, lateral inhibition on each layer, bidirectional vertical excitatory connections among nodes on different layers, massive but not absurd interconnections) is bound to generate some localist representations. Consider hearing the musical note C (66 Hz). It is composed of not only this fundamental pure tone but also of its upper partials, which are integer multiples: C′ (132 Hz), G′ (198 Hz), C″ (264 Hz), E″ (330 Hz), and so on (Helmholtz 1885/1954). Consider a C-major net that consists of nodes for all of the notes in the C-major scale on layer L1. Well, the brain doesn't have just one layer, but several. What will happen? Given the Hebb learning rule and its revisions (the correct citation is actually Thorndike 1911), we would expect the fundamental frequency and the upper partials to form excitatory connections with one or more nodes on level L2. These would be localist nodes coding the musical note C (66 Hz) and its upper partials. If we were to remove the fundamental frequency of 66 Hz, common sense would suggest that we should hear C″. In fact, we still hear C. This is easy enough to explain: the upper partials have excited the L2 C node, and it has activated the missing fundamental. This explanation is analogous to the localist explanation of the word-superiority effect. The reader may think that a distributed representation model of the sort proposed by McClelland and Rumelhart (1985) could handle this problem: we connect all the L1 nodes in an excitatory fashion and use the Hebb rule to train the net. This will give us an $N \times N$ matrix of connection weights. It will also give us a mess that would not restore the missing fundamental. The notes in C-major (or any other scale) share too many upper partials. They are not orthogonal enough, so would interfere with each other too much. It might be possible to solve this problem using only distributed representations, but the solution would be neither parsimonious nor plausible.

A phase transition between localist and distributed representation

Peter C. M. Molenaar and Maartje E. J. Raijmakers

Department of Psychology, University of Amsterdam, 1018 WB Amsterdam, The Netherlands. **op_molenaar@macmail.psy.uva.nl**

Abstract: Bifurcation analysis of a real-time implementation of an ART network, which is functionally similar to the generalized localist model discussed in Page's manifesto shows that it yields a phase transition from local to distributed representation owing to continuous variation of the range of inhibitory connections. Hence there appears to be a qualitative dichotomy between local and distributed representations at the level of connectionistic networks conceived of as instances of nonlinear dynamical systems.

This manifesto presents a clear specification and some very challenging implications of localist representation in connectionist psychological modelling. Unfortunately, within the limited confines of this commentary we will have to restrict ourselves to a single aspect of this excellent manifesto.

Section 4 of the manifesto introduces a generalized localist model, the architecture of which is reminiscent of an ART network (e.g., Grossberg 1980). Like ART, the generalized localist model depicted in Figure 3 has two layers for input representation and coding, respectively. Figure 5 shows an extension of this generalized localist model involving an additional coding level, like in ART 2 networks (e.g., Carpenter & Grossberg 1987). Hence, in conformance with several indications given by the author, the generalized localist model and ART models share the same functional architecture.

Given this equivalence between the generalized localist model and ART, including equivalent local representation at the coding level(s) equipped with competitive interactions, and because the distinctions between localist and distributed representations constitute a basic tenet of the manifesto, it might be an interesting question whether ART can also give rise to distributed representations. The scenario we have in mind is one in which an ART network starts with regular local coding, but under continuous variation of a subset of the network parameters suddenly undergoes a transition to distributed coding. Such a scenario is prototypical of a mathematical bifurcation analysis of ART, aimed at the detection and consequent specification of phase transitions in its performance.

Such a bifurcation analysis of ART has been carried out by Raijmakers et al. (1997), using a real-time implementation of ART as a system of coupled nonlinear differential equations (exact ART, cf. Raijmakers & Molenaar 1996). Based on empirical evidence obtained in the developmental neurosciences (e.g., Purves 1994), the range and strength of excitatory and inhibitory connections, arranged in on-center-off-surround competitive fields, were continuously varied as in actual brain maturation. A number of so-called fold bifurcations were thus detected, one of which would seem to be of particular importance for the present discussion. This concerns the bifurcation marking a phase transition between local coding and distributed coding owing to a continuous decrease of the range of inhibitory connections.

A bifurcation or phase transition is indicative of a qualitative change in the behavior of a nonlinear dynamic system (e.g., van der Maas & Molenaar 1992). That is, the dynamic organization of a system's behavior suddenly shows qualitatively new properties like the emergent new types of attractors in phase space (e.g., Thelen & Smith 1994). Hence the phase transition obtained in exact ART owing to a continuous decrease of the range of inhibitory connections in the competitive field is also indicative of a qualitative change, in this particular case a qualitative change marking the transition from local coding to distributed coding. What could this imply for the distinction between local and distributed representation?

At the level of ART as a system of differential equations, and given the equivalence noted above also at the level of the generalized localist model as a system of differential equations, the bifurcation marking the transition between local and distributed coding involves a qualitative change in the sense indicated above. Consequently, there appears to be discontinuous border between local coding and distributed coding. At least at this level of connectionistic models, conceived of as particular instances of nonlinear dynamic systems, there appears to be some kind of dichotomy between local representation on the one hand and distributed representation on the other hand. Of course this leaves completely open the question whether such a dichotomy also exists at other, for instance, more functionally defined, levels. The manifesto presents strong arguments that no such dichotomy between "distributed" and "localist" may exist at these other levels.

Localist models are already here

Stellan Ohlsson

Department of Psychology, University of Illinois at Chicago, Chicago, IL 60607. **stellan@uic.edu** **www.uic.edu/depts/psch/ohlson-1.html**

Abstract: Localist networks are symbolic models, because their nodes refer to extra-mental objects and events. Hence, localist networks can be combined with symbolic computations to form *hybrid models.* Such models are already familiar and they are likely to represent the dominant type of cognitive model in the next few decades.

Connectionist models are of limited interest to the student of higher-order human cognition for three reasons.

1. Human cognition – particularly as expressed in art, literature, mathematics, technology, and science – is not a reflection of the perceived environment. It is driven by goals and by the imagination, that is, by mental representations that go beyond experience. Human behavior is not adaptive or reactive but *centrally generated.* Hence, it cannot be explained in terms of mappings of input patterns onto output patterns, regardless of the mechanism by which those mappings are obtained.

2. Human behavior is both hierarchical and sequential. Elementary actions are orchestrated both in space and over time. Hierarchical, sequential processes that extend over time are precisely the kind of processes that connectionist networks – of whatever kind – do a poor job of modeling. In contrast, these fundamental features of human behavior are quite naturally and unproblematically represented in symbolic models.

3. Connectionist models represent learning in terms of quantitative calculations over network parameters (weights, strengths, activation levels). Changes in these parameters influence how the relevant nodes and links enter into a given cognitive process, that is, they explain how a piece of knowledge is used. However, the more fundamental question in learning is how knowledge is constructed in the first place. That is, why is this node connected to that node? Network models – of whatever kind – have no answer to this question and finesse it by assuming that every node is connected to every other node and by allowing link strengths to take zero as their value.

However, this subterfuge is untenable. There is not the slightest reason to believe that every neuron in the human brain (or neural column or any other functional unit) is connected to every other neuron. In fact, what we do know about transmission pathways within the brain indicates that this is precisely how the brain is *not* wired. Hence, a fundamental assumption, without which most of the theoretical machinery of connectionist modeling becomes irrelevant or unworkable, is not satisfied.

I take these three observations as sufficient proof that connectionist models cannot succeed as models of higher cognition. It does not follow that such models are without lessons to teach. First, network modeling has brought home, as only actual exploration of such models could have brought home, a realization of exactly how much cognitive work can be accomplished by the manipulation of network parameters. Second, the contrast between symbolic models and network models has served to highlight severe weaknesses in the symbolic approach. Enduring lessons include the need to assume distributed rather than serial processing, the need for robust computational techniques, and the quantitative nature of phenomena such as familiarity effects, recognition, priming, typicality in categorization, metacognitive judgments, and so on. There is no doubt that such *graded phenomena* are better explained in terms of adjustments of quantitative network parameters than in terms of symbolic computations as classically understood.

The obvious response to this situation is to conclude that a model of higher-order cognition must include both symbolic computations and quantitative adjustments of the representational network over which those computations are defined. However, such a *hybrid model* was until recently difficult to conceptualize, due to the connectionists' insistence that network models had – or should have – distributed representations as well as distributed processing, a stance that cannot be reconciled with the symbolic assumption that there are atomic symbols that refer to extra-mental objects and events.

Mike Page's localist manifesto provides a welcome resolution of this dilemma. In a localist network, as defined in the target article, the nodes are symbols – they refer to things – and localist models are therefore symbolic models. It then becomes natural to envision a cognitive system in which a localist network is responsible for graded cognitive phenomena while symbolic computations over that network are responsible for phenomena in which compositionality and systematicity are important (deductive reasoning, heuristic search during problem solving, planning, skilled action, etc.).

However, this move reveals that localist networks are less than revolutionary. Ideas about strength adjustments and changing activation levels have played central roles in semantic networks, production systems and other types of symbolic models from the start. The most recent version of the ACT-R model (Anderson & Lebiere 1998) operates both with subsymbolic, quantitative processes (shaped by a different mathematical rationale than the one behind connectionist models) and symbolic operations on chunks and rules. Theoretical discussions of such *hybrid systems* are already in full swing (Sloman & Rips 1998) and empirical evidence for them is accumulating (e.g., Jones & McLaren 1999). Although Page's article thus comes too late to pioneer the hybrid approach that no doubt will dominate cognitive modeling in the next few decades, his article is nevertheless extremely useful in laying out the arguments why network modelers should abandon the idea of distributed representations and return to symbolic modeling, augmented and strengthened by the lessons taught by connectionism.

A competitive manifesto

R. Hans Phaf[a] and Gezinus Wolters[b]

[a]*Psychonomics Department, University of Amsterdam, 1018 WB Amsterdam, The Netherlands;* [b]*Department of Psychology, Leiden University, 2300 RB Leiden, The Netherlands.* **pn_phaf@mac/mail.psy.uva.nl wolters@fsw.leidenuniv.nl**

Abstract: The distinction made by Page between localist and distributed representations seems confounded by the distinction between competitive and associative learning. His manifesto can also be read as a plea for competitive learning. The power of competitive models can even be extended further, by simulating similarity effects in forced-choice perceptual identification (Ratcliff & McKoon 1997) that have defied explanation by most memory models.

The relationship between input pattern and node activation centers in the discussion of localist versus distributed representations. If for a given input pattern a single node in either a (hidden) layer or subdivision of a layer (i.e., a module; see Murre et al. 1992; Phaf et al. 1990) is activated, the representation is called local. Page forcefully argues for localist models. We believe, however, that the representational issue is not the core of the debate. The representational issue hinges on the specific labeling of input and representational patterns, which does not in a principled way relate to internal network design. The McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982) model for context effects in letter recognition, for instance, has local representations in the third (word level) layer when single words are presented, but distributed representations when exposed to four-letter nonwords. So, localist representations in this model seem to depend on deciding that among all possible four-letter strings only those strings that result in a single activation in the third layer are relevant. By the same operation, many distributed models can, in principle, be made localist. If, for instance, a back-propagation network has been trained on at least as many independent patterns as there are hidden nodes, new input patterns can be constructed that have single node representations. This eventually boils down to solving n (the number of hidden nodes) equations with n unknown variables. We can label the newly constructed patterns (e.g., as particular words), and choose to ignore (e.g., call these nonwords) all patterns with distributed representations on which the network was initially trained. So, by subsequently changing the relevance of the input patterns, a back-propagation network may become localist.

However, which input patterns are relevant and which are not is usually not determined after learning has taken place. The issue

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

then changes to the question what kinds of representations develop for relevant input patterns during learning, and it focuses "thus" on the learning procedure. We agree with Page that a broad distinction can be made between (supervised) associative learning and (largely unsupervised) competitive learning. Associative learning couples two activation patterns (i.e., generally input and output patterns) and does not put restrictions on the number of activated nodes forming the association. Competitive learning functions to classify, or categorize, input patterns in distinct classes. By its very nature, competition works to select single nodes or neighbourhoods of nodes (e.g., Kohonen 1995) for these classes and thus leads to the development of local (or even topologically organized) representations for the preselected input patterns. It is, probably, no coincidence that Page extensively discusses competitive learning under the heading of "A generalized localist model" (sect. 4).

One of the many interesting applications of competitive learning that are dealt with by Page is the effect of similarity on practice (sect. 4.2.5). We suggest here that a similar competitive model can account for similarity effects in forced-choice perceptual identification (Ratcliff & McKoon 1997), which has proved a stumbling block for many computational models of memory. Generally, previous study facilitates identification of a briefly presented word (Jacoby & Dallas 1981). Such facilitation, or repetition priming, is also found when after a brief target presentation, a choice has to be made between the target word and a foil, but only when target and foil are similar. More precisely, when the target is studied beforehand there is facilitation, and when the (similar) foil has been studied there is inhibition in forced-choice identification. Both facilitation and inhibition disappear, however, when target and foil are dissimilar. This is often interpreted as evidence that previous study does not strengthen perceptual fluency (which should result in priming also with dissimilar alternatives), but that it induces a response bias (Ratcliff & McKoon 1997). The fluency versus bias issue may reflect the distinction between associative and competitive learning, especially if it is realized that in connectionist terms there is no need to restrict the bias in competition to a response level.

In models using associative learning, facilitation of target identification owing to previous study should be found, irrespective of similarity to a foil. In competitive models reaction times (and errors) do not depend on the extent of the input per se, but on the time it takes to solve competition (e.g., Phaf et al. 1990). Whereas associative models tend to rely on absolute levels of activation, similarity and comparing activations play a more important role in competition. If in the model of Figure 5, topological ordering is introduced in the second layer according to a competitive learning scheme (e.g., Kohonen 1995; Phaf & van Immerzeel 1997), evidence for both target and foil may accrue there in a topologically organized fashion. The two nodes in the third layer represent the choice of either target or foil. When target and foil are similar, they will compete heavily in the middle layer and an increase in weights to the target will speed up the resolution of competition. Increases in weight to the foil will slow the resolution of competition. The competition between dissimilar targets and foils, however, is solved almost immediately and will not be influenced by changes in weights to either target or foil.

In this manner, the rather elusive phenomenon of similarity effects on forced-choice perceptual identification may be modeled quite directly in competitive models. The relative ease of applying these models to experimental data, as in the previous example, justifies the challenging tone of the paper by Page, but we think the title should be changed in "a competitive manifesto."

Dynamic thresholds for controlling encoding and retrieval operations in localist (or distributed) neural networks: The need for biologically plausible implementations

Alan D. Pickering

Department of Psychology, St. George's Hospital Medical School, University of London, London SW17 0RE, United Kingdom. **a.pickering@sghms.ac.uk**
www.sghms.ac.uk/depts/psychology/aphome.htm

Abstract: A dynamic threshold, which controls the nature and course of learning, is a pivotal concept in Page's general localist framework. This commentary addresses various issues surrounding biologically plausible implementations for such thresholds. Relevant previous research is noted and the particular difficulties relating to the creation of so-called instance representations are highlighted. It is stressed that these issues also apply to distributed models.

The target article by Page is a welcome advertisement for an older, alternative tradition in neural net research, previously championed by Grossberg and others, and which usually employs localist representations. Page shows clearly that one cannot justify ignoring this older work on the grounds that it uses localist representations.

This commentary will focus on the general localist framework developed by Page, and especially on its central construct: a dynamic threshold, θ, which controls the course and nature of the learning and performance of the system. There are a number of problems with the article's description of threshold operations. The model described in section 4.1, where considerable detail is given, will not work. In a later brief section (4.4), an outline of an improved model is given. Nonetheless, specific details, concerning a series of critical issues, are not provided. This commentary will suggest that thinking about the biological implementation of these thresholds can be helpful in solving the psychological and computational problems posed by the target article.

The issues identified here apply equally to the ART architectures (Carpenter & Grossberg 1987a) and other localist models (e.g., Pearce 1994) that are very similar to Page's account. However, it must be stressed that any difficulties encountered in implementing these thresholds will not be specific to localist models. Distributed models face identical issues concerning the timecourse and dynamics of learning (Hasselmo & Schnell 1994; Hasselmo & Wyble 1997; Hasselmo et al. 1995).

What is the threshold? In section 4.1, Page considers how localist representations may be learned in an L2 layer driven by feedforward input from a (distributed) L1 layer. Computationally speaking, the threshold parameter is the minimum feedforward input required for L2 nodes to be activated by input from L1. Biologically speaking, one might think of the threshold as a property of the L2 neurons themselves: for example, it might reflect the minimum synaptic current required before any change in the membrane potential of L2 neurons can occur. It may be more useful, however, if the threshold reflects the action of an inhibitory interneuron or neuromodulatory signal, projecting diffusely across the L2 layer. Furthermore, the activation of the threshold cells needs to be driven by the outputs of the L2 neurons. Examples of neuromodulatory thresholds exist in the literature.

In a recent work (Salum et al. 1999), my coauthors and I have proposed that inhibition of firing in striatal neurons by dopaminergic inputs from ventral tegmental area (VTA) or substantia nigra pars compacta (SNc) may serve a threshold function. The tonic level of dopaminergic output from VTA/SNc is initially low. However, as cortical (L1) inputs activate the striatal (L2) neurons, functionally excitatory feedback pathways allow striatal output to increase VTA/SNc firing and hence to increase the threshold. This mechanism turns off many of the initially active striatal (L2) nodes, with sustained activity remaining in those with strongest inputs from cortex (L1); these are the nodes that undergo L1-L2 learning.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

A similar mechanism can be seen in the (distributed) models of cholinergic suppression of synaptic transmission within the hippocampus (Hasselmo & Schnell 1994; Hasselmo et al. 1995). In these models, diffuse cholinergic projections modulate synapses employing other neurotransmitters. An initially high level of cholinergic cell firing sets a high threshold so that only the strongest inputs can begin to activate the equivalent of the L2 layer. Outputs from this layer then feed back and reduce cholinergic firing, thereby lowering the threshold so that the L2 neurons can activate fully and undergo learning.

Different thresholds for different L2 nodes? Next we consider when an L1 input vector, *S*, has been learned by an L2. Page describes the node (the *S*-node) as being "committed" to the input *S*. The system may subsequently experience a similar vector, *S**, and there will be input to the *S*-node in L2 only if the similarity between *S* and *S** exceeds the threshold setting. In this case, the weights to the *S*-node will be recoded to reflect *S**, interfering with previous learning about *S*. If *S** is sufficiently dissimilar from *S*, so that the threshold is not exceeded, Page argues (sect. 4.1) that *S** will come to be represented by an uncommitted L2 node, thereby preventing interference with prior learning. This is a "crucial point" of the paper, but it begs several questions. The similarity between the input vector *S** and the weight vector projecting to an uncommitted node will usually be much lower than the similarity between *S** and the weight vector projecting to the *S*-node. With equal thresholds across L2, the net input to the committed node will therefore be greater than that to uncommitted nodes and so uncommitted nodes will not activate and undergo learning in the presence of *S**. In section 4.4, Page implies that uncommitted nodes have lower thresholds than committed nodes, which could solve this problem, although no details about how this could be achieved are given.

A paper by Hasselmo and Wyble (1997) suggests how high thresholds for committed nodes may be achieved in the hippocampal system. They employed an inhibitory interneuron, I, projecting to all nodes in the equivalent of the L2 layer. The inhibitory weights from I to active L2 nodes were strengthened at the same time as the excitatory L1-L2 weights were learned. In the current scenario, this ensures that when *S** is subsequently presented, the output from I creates a higher threshold for the committed *S*-node (owing to the strengthening of the inhibitory weight between I and the *S*-node, which occurred when *S* was presented) than for uncommitted nodes. If this increase in threshold offsets the increased excitatory input to the *S*-node (based on the excitatory L1-to-L2 weights learned when *S* was presented), then a new L2 node will be recruited to represent *S**.

How to make the system generalise. The foregoing suggestions lead to a system that can create orthogonal representations, without interference, from very similar stimulus inputs. However, such a stimulus will not generalise its prior learning about *S* in response to the similar stimulus, *S**. This has been a central criticism of localist models (see sect. 6.2).

A solution to this problem seems to require that, on a particular trial, the L2 threshold start at a low level. This permits activation of L2 nodes based on prior learning. These activations then feed forward and enter into competition at other layers, where the response is selected. When L1-to-L2 weight learning subsequently takes place, the L2 threshold needs to be switched to a higher setting so that a new representation can be added in the L2 layer. Once again, this computational solution is briefly mentioned by Page in section 4.4. He appears to suggest that a reinforcement signal following the response (correct or incorrect) may raise the threshold appropriately. Interestingly, there is evidence that dopamine cells in VTA/SNc increase firing in response to primary rewards (Schultz et al. 1995). Page's suggestion would therefore fit with Salum et al.'s (1999) account of VTA/SNc dopamine cell firing as providing a threshold for learning in the striatum. Furthermore, the striatum is clearly implicated in the paradigm task for Page's model: human category learning (Ashby et al. 1998). Unfortunately, these dopamine cells stop responding to reward after a modest degree of training (Schultz et al. 1995) and there is no evidence that their projections to the striatum have the plasticity required to differentiate the thresholds for committed and uncommitted nodes.

Are instance representations biologically plausible? Instance representations (IRs) are central to the models of Logan (1988; 1992), Nosofsky and Palmeri (1997), and to parts of Page's article. IRs are formed when repeated occurrences of the same stimulus are represented by different L2 nodes. In principle, this is just an extension of the case in section 2, but with the threshold raised so high that even the same stimulus fails to undergo further learning at the previously committed L2 node. Even modifying Page's model as suggested in this commentary, I have not been able to achieve IRs in a stable, biologically plausible simulation. The need to show that this can be done is an important challenge for the above theorists.

Stipulating versus discovering representations

David C. Plaut and James L. McClelland

Departments of Psychology and Computer Science and the Center for the Neural Basis of Cognition, Mellon Institute, Carnegie Mellon University, Pittsburgh, PA 15213-2683. **plaut@cmu.edu** **jlm@cnbc.cmu.edu**
www.cnbc.cmu.edu/~plaut
www.cnbc.cmu.edu/people/mcclelland.html

Abstract: Page's proposal to stipulate representations in which individual units correspond to meaningful entities is too unconstrained to support effective theorizing. An approach combining general computational principles with domain-specific assumptions, in which learning is used to discover representations that are effective in solving tasks, provides more insight into why cognitive and neural systems are organized the way they are.

Page sets up a fundamental contrast between localist versus distributed approaches to connectionist modeling. To us there appear to be several dimensions to the actual contrast he has in mind. Perhaps the most fundamental distinction is whether it is stipulated in advance that representational units be assigned to "meaningful entities" or whether, as we believe, it is better to discover useful representations in response to task constraints. We agree with Page that localist connectionist models have made important contributions to our understanding of many different cognitive phenomena. However, we think the choice of representation used in the brain reflects the operation of a set of general principles in conjunction with domain characteristics. It is a program of scientific research to discover what the principles and domain characteristics are and how they give rise to different types of representations. As a starting place in the discovery of the relevant principles, we have suggested (McClelland 1993b; Plaut et al. 1996) that the principles include the following: that the activations and connection weights that support representation and processing are graded in nature; that processing is intrinsically gradual, stochastic, and interactive; and that mechanisms underlying processing adapt to task constraints.

Constraint versus flexibility. Page's suggestion that we stipulate the use of representations in which the units correspond to meaningful entities would appear on the face of it to be constraining, but in practice it appears to confer too much flexibility. Indeed, throughout his target article, Page applauds the power and flexibility of localist modeling, often contrasting it with models in which representations are discovered in response to task constraints. A particularly telling example is his treatment of age-of-acquisition effects (which he considers to be "potentially difficult to model in connectionist terms," sect. 4.4). Page describes a localist system, incorporating three new assumptions, that would be expected to exhibit such effects. However, it would have been

even easier for Page to formulate a model that would not exhibit such effects – a localist model without the additional assumptions might suffice. A critical role of theory is to account not only for what does occur but also for what *does not* (see Roberts & Pashler, in press); the localist modeling framework provides no leverage in this respect. In contrast, the distributed connectionist model, which is more constrained in this regard, is potentially falsifiable by evidence of the presence or absence of age-of-acquisition effects. In fact, Page has it exactly backwards about the relationship between such effects and connectionist models that discover representations via back-propagation. Ellis and Lambon-Ralph (personal communication) have pointed out that age of acquisition effects are actually intrinsic to such models, and their characteristics provide one potential explanation for these effects.

Page is exactly right to point out that "it sometimes proves difficult to manipulate distributed representations in the same way that one can manipulate localist representations" (sect. 7.5). In other words, the learning procedure discovers the representations subject to the principles governing the operation of the network and the task constraints, and the modeler is not free to manipulate them independently. Far from being problematic, however, we consider this characteristic of distributed systems to be critical to their usefulness in providing insight into cognition and behavior. By examining the adequacy of a system that applies a putative set of principles to a model that addresses performance of a particular task, we can evaluate when the principles are sufficient. When they fail, we gain the opportunity to explore how they may need to be adjusted or extended.

These considerations are relevant to Page's analysis of the complementary learning system hypothesis of McClelland et al. (1995). These authors made the observation that connectionist networks trained with back-propagation or other structure-sensitive learning procedures (a) discover useful representations through gradual, interleaved learning and (b) exhibit catastrophic interference when trained sequentially (McCloskey & Cohen 1989). Based on these observations, and on the fact that the gradual discovery of useful representations leads to a progressive differentiation of conceptual knowledge characteristic of human cognitive development, McClelland et al. (1995) suggested that the neocortex embodies the indicated characteristics of these learning procedures. One implication of this would be that rapid acquisition of arbitrary new information would *necessarily* be problematic for such a system, and that a solution to this problem would be provided if the brain also exploited a second, complementary approach to learning, employing sparse, conjunctive representations, that could acquire new arbitrary information quickly. The argument was that the strengths and limitations of structure-sensitive learning explained *why* there are two complementary learning systems in hippocampus and neocortex.

In contrast, Page goes to some length to illustrate how a localist approach to learning could completely avoid the problem of catastrophic interference that arises in connectionist networks trained with back-propagation. Indeed, in his approach, the hippocampal system is redundant with the neocortex as there is no need for cortical learning to be slow. Thus, within the localist framework, the existence of complementary learning systems in the hippocampus and neocortex is completely unnecessary, and hence the existence of such a division of labor in the brain is left unexplained.

Learning representations versus building them by hand. A common approach in the early days of connectionist modeling was to wire up a network by hand, and under these circumstances there seemed to be a strong tendency among researchers to specify individual units that correspond to meaningful entities (see, e.g., Dell 1986; McClelland & Rumelhart 1981). However, learning is a central aspect of many cognitive phenomena, so it is essential that a modeling framework provide a natural means for acquiring and updating knowledge. Once one turns to the possibility that the knowledge embodied in a connectionist network might be learned (or even discovered by natural selection), one immediately has the chance to revisit the question of whether the individual units in a network should be expected to correspond to meaningful entities. It is not obvious that correspondence to meaningful entities per se (or the convenience of this correspondence for modelers) confers any adaptive advantage.

To his credit, Page acknowledges the central role that learning must play in cognitive modeling, and presents a modified version of the ART/competitive learning framework (Carpenter & Grossberg 1987b; Grossberg 1976; Rumelhart & Zipser 1985) as a proposal for learning localist representations. However, there are a number of difficulties with this proposal, all of which point to reasons for continuing to pursue other alternatives. We consider three such difficulties here.

(1) On close examination, most of the positive aspects of the proposal derive from properties of the assumed distributed representations that are input to the localist learning mechanism. For example, Page points out that localist models permit graded, similarity-based activation. It is crucial to note, however, that the pattern of similarity-based activation that results depends entirely on the similarity structure of the representations providing input to the localist units. Unfortunately, nowhere in Page's target article does he indicate how his localist approach could solve the problem of discovering such representations.

In contrast, a key reason for the popularity of back-propagation is that it is effective at discovering internal, distributed representations that capture the underlying structure of a domain. For example, Hinton (1986) showed that a network could discover kinship relationships within two analogous families, even in the absence of similarity structure in the input representations for individuals. Although some runs of the network produce internal representations with "meaningful" units (e.g., nationality, generation, gender, branch-of-family), the more general situation is one in which the meaningful features of the domain are captured by the *principal components* of the learned representations (McClelland 1994; see also Anderson et al. 1977; Elman 1991).

(2) Page notes that localist models are capable of considerable generalization. This again arises from the similarity-based activation due to the distributed patterns of activation that are input to the localist units. We suggest that one reason localist-style models (e.g., the generalized context model, Nosofsky 1986, or ALCOVE, Kruschke 1992) have proven successful in modeling learning in experimental studies is because they apply to learning that occurs within the brief time frame of most psychology experiments (1 hour up to at most about 20 hours spread over a couple of weeks). Within this restricted time frame, we expect relatively little change in the relevant dimensions of the representation, so the generalization abilities of models that learn by adapting only the relative salience of existing dimensions may be sufficient.

What seems more challenging for such approaches is to address changes in the underlying representational dimensions themselves. Such shifts can occur in our task-driven approach through the progressive, incremental process by which learning assigns representations in response to exposure to examples embodying domain knowledge (McClelland 1994). On our view, the establishment of appropriate representations is a developmental process that takes place over extended periods of time (months or years), allowing models that develop such representations to account for developmental changes such as progressive differentiation of conceptual knowledge (Keil 1979) and developmental shifts in the basis of categorization of living things from superficial to a metabolic/reproductive basis (Johnson & Carey 1998).

(3) Within the above limitations, Page's proposed localist learning procedure sounds like it might work on paper, but it is telling that he discusses in detail how the learning process might proceed only for the case in which every presentation of an item in a psychological experiment results in a separate localist representation. This form of localism – the idea that every experience is assigned a separate unit in the representation – seems highly implausible to us. It is difficult to imagine a separate unit for every encounter

with every object or written or spoken word every moment of every day. Such instance-based approaches have led to some interesting accounts of psychological data (by Logan and others, as Page reviews), but in our view it is best to treat this form of localist modeling as an interesting and useful abstraction of an underlyingly distributed, superpositional form of representation. More specifically, we agree there is a trace laid down in the brain resulting from each experience and that localist models can approximate how these traces influence processing. We believe, however, that the traces are actually the adjustments to the connections in a distributed connectionist system rather than stored instances. McClelland and Rumelhart (1985), for example, showed how a simple superpositional system can capture several patterns of data previously taken as supporting instance-based theories, and Cohen et al. (1990) demonstrated that distributed connectionist models trained with back-propagation can capture the power law of practice just as Logan's instance models do.

It seems somewhat more plausible to us that multiple occurrences of a meaningful cognitive entity such as a letter or word might be mapped onto the same unit. However, the ability of models that exploit the kind of procedure Page proposes to actually produce such representations is unclear. In our experience, to obtain satisfactory results with such models it is necessary to tune the "vigilance" parameter very carefully, and often in ways that depend strongly on specifics of the training set. But there is a deeper problem. Whenever there is any tolerance of variation among instances of a particular item, one immediately runs into the fact that the modeler is forced to decide just what the acceptable level of mismatch should be. If, for example, a reader encounters a misspelling of the word ANTARTICA, should we necessarily imagine that the cognitive system must create a separate unit for it? Or if, in a certain Wendy's restaurant, the salad bar is not immediately opposite the ordering queue, should we create a new subcategory of the Wendy's subcategory of restaurants? Within a task-driven learning approach, in which robust patterns of covariation become representationally coherent, and in which subpatterns coexist within the larger patterns of covariation, such otherwise thorny issues become irrelevant (McClelland & Rumelhart 1985; Rumelhart et al. 1986).

Task-driven learning can discover localist-like representations. As we have noted, whereas Page would stipulate localist representations for various types of problems, our approach allows an appropriate representation to be created in response to the constraints built into the learning procedure and the task at hand. At a general level, distributed representations seem most useful in systematic domains in which similar inputs map to similar outputs (e.g., English word reading), whereas localist representations (and here we mean specifically representations involving one unit per entire input pattern) are most useful in unsystematic domains in which similar inputs may map to completely unrelated outputs (e.g., word comprehension, face naming, episodic memory). It is thus interesting (although, to our knowledge, not particularly well documented) that standard connectionist learning procedures tend to produce dense, overlapping internal representations when applied to systematic tasks, whereas they tend to produce much sparser, less overlapping representations when applied to unsystematic tasks. Although Page considers the latter to be functionally equivalent to localist representations, there are at least two reasons to reject this equivalence. First, sparse distributed representations scale far better than strictly localist ones (Marr 1970; McClelland & Goddard 1996; Kanerva 1988). Second, and perhaps more important, sparse distributed representations are on one end of a continuum produced by the same set of computational assumptions that yield more dense, overlapping representations when these are useful to capture the structure in a domain.

Other comments on Page's critique of "distributed" approaches. In rejecting what he calls the distributed approach, Page levels several criticisms that are either incorrect or overstated, partly because he seems to adopt an overly narrow view of the approach. For one thing, Page appears to equate the distributed approach with the application of back-propagation within feed-forward networks. He then raises questions about the biological plausibility of back-propagation but fails to acknowledge that there are a number of other, more plausible procedures for performing gradient descent learning in distributed systems which are functionally equivalent to back-propagation (see, e.g., O'Reilly 1996). Page questions whether distributed systems can appropriately fail to generalize in unsystematic domains (e.g., mapping orthography to semantics for pseudowords [sect. 4.3.1]) when such behavior has already been demonstrated (Plaut 1997; Plaut & Shallice 1993). He also questions how a distributed system can decide when and how to respond without some sort of homuncular energy-monitoring system (although see Botvinick et al. 1999, for recent functional imaging data supporting the hypothesis that the anterior cingulate may, in fact, play such a role). In fact, no such explicit decisions are required; all that is needed is that the motor system be sufficiently damped that it initiates behavior only when driven by strongly activated, stable internal representations (see Kello et al., in press, for a simple demonstration of this idea).

Based on the above, we suggest that the representations used by the brain in solving a particular task are not something we should stipulate in advance. Rather, they are selected by evolution and by learning as solutions to challenges and opportunities posed by the environment. The structure of the problem will determine whether the representation will be localist-like or more distributed in character.

ACKNOWLEDGMENT

We thank the CMU PDP research group for helpful comments and discussion.

What is the operating point? A discourse on perceptual organisation

Simon R. Schultz

Howard Hughes Medical Institute and Center for Neural Science, New York University, New York, NY 10003. **schultz@cns.nyu.edu**
www.cns.nyu.edu/~schultz

Abstract: The standard dogmatism ignores the fact that neural coding is extremely flexible, and the degree of "coarseness" versus "locality" of representation in real brains can be different under different task conditions. The real question that should be asked is: What is the operating point of neural coding under natural behavioural conditions? Several sources of evidence suggest that under natural conditions some degree of distribution of coding pervades the nervous system.

When the task that a monkey is required to perform is reduced to that undertaken by a single neuron, single neurons perform as well as the monkey (Newsome et al. 1989). This is interesting not only in that it tells us that we have managed to find a task that fairly accurately isolates the perceptual decision made by the cell; it also tells us that the coding régime in which neurons operate is adjusted in a task-dependent manner. Clearly there are situations in which whole animals make discriminations based upon the output of more than one cell (see sect. 2.5). Further evidence for the task dependence of coding strategies comes from more recent experiments from the Newsome laboratory (Newsome 1999), which provide evidence that both winner-take-all and weighted output decoding strategies are utilised by rhesus macaques depending upon the task being performed. Information can be sparsely or coarsely distributed across populations of neurons depending upon the stimulation conditions. The critical question that must be asked is: What kind of coding (representational) régime does natural stimulation and behaviour invoke?

It is still not yet understood whether perception is mostly due to small numbers of optimally tuned neurons, or the much larger

number of neurons that must be suboptimally excited by a given stimulus. Where is the natural "operating point" on the tuning curve? This provides a way of quantifying precisely how "localist" the representations are in a naturalistic paradigm. A useful way to study this operating point is in terms of information theory. Each neuron that responds must contribute some amount of information, measured in bits, to the overall percept. Neurons that respond with higher firing rates will, in general, contribute more information to the percept than those that elevate their firing rates only marginally above their spontaneous activity level. Thus along a given stimulus dimension, the information per neuron will be higher at the centre (peak) of the tuning curve than further out. However, substantial discrimination performance is maintained at some distance from the centre of the tuning curve, as is demonstrated by the sensitivity MT neurons show to opposed directions of motion quite far from the preferred-null axis (Britten & Newsome 1999). In general, there will be many more neurons operating further from the centre of the tuning curve, and the sum of many lesser contributions may be enough to make up for – or even overwhelm – the contributions of the high-firing neurons to the total information represented. This is shown schematically in Figure 1. Evidence suggesting that this may indeed be the case comes from inferior temporal cortex recordings by Rolls et al. (1997), in which low firing rates were found to contribute more to the total Shannon information than high firing rates, due to both their greater number and the larger noise at high firing rates.

The preceding discussion largely focused on a single-stimulus dimension; however, the principle should be even more applicable to the more realistic situation of multidimensional stimulation. In this case a given neuron, even if "matched" on one dimension, is extremely likely to be responding submaximally on at least one other dimension. Indeed, computer simulations have indicated that for multidimensional stimuli, a large population of cells is required to code stimulus attributes accurately and to account for known behavioural performance (Zohary 1992). Studies of the correlation between simultaneously recorded neurons suggested that the outputs of hundreds of neurons can be effectively pooled together for perception (Zohary et al. 1994). However, that work may significantly underestimate the number of neurons that can be usefully pooled when naturalistic, multidimensional stimulation is taken into account (Panzeri et al. 1999; Schultz & Panzeri 1999): larger dimensionality decreases the signal correlation and increases drastically the range over which the information increases linearly (i.e., factorially) with the number of neurons pooled.

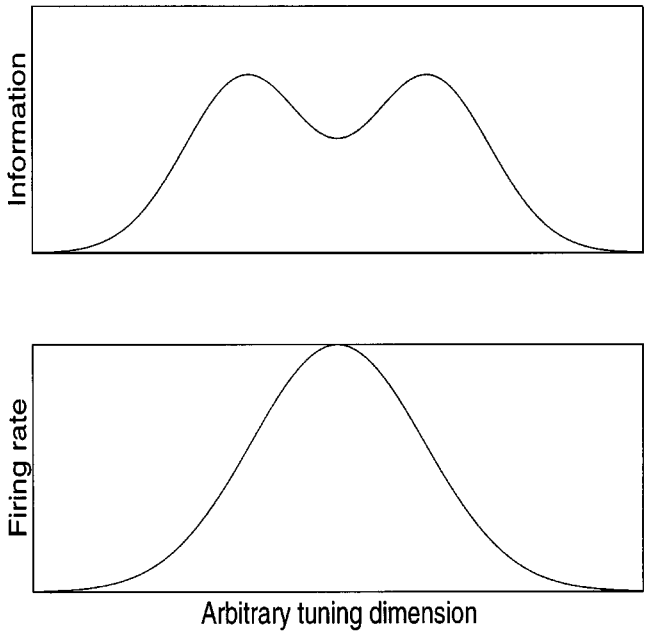

Figure 1 (Schultz). A schematic figure portraying the total information represented in a percept from all cells operating at different portions of their tuning curve, and thus at different firing rates. This could be considered to be a cross-section through a multidimensional tuning curve. Cells with very low firing rates (far from the center of the tuning curve) cannot provide much information; although cells with high firing rates do, there are not many of them, and thus their total contribution cannot be high either. The percept must be mostly contributed to by cells somewhere in between; precisely where is an empirical issue for further study.

It has been suggested that localist coding is particularly important in higher visual areas such as the anterior inferotemporal cortex. However, one should consider that studies of coding receive a strong bias from single unit recording strategies themselves. It has recently been shown that IT cells are in fact arranged in a columnar structure based on stimulus attributes (Wang et al. 1998). However, unlike in early sensory brain areas, our experiments on higher areas do not directly manipulate the dimensions along which this spatial organisation occurs – instead, we may present different faces or objects that change many stimulus dimensions at once. This will lead inevitably to apparent "grandmother-cell-like" encoding, where in fact a very similar coding organisation to early visual areas may exist.

Of course, it is important that we understand what aspects of the neuronal response can be decoded (although we must not neglect the fact that responses at the level of the cortical heterarchy we are considering may "enter perception" in themselves). If only bursting responses were decoded, for instance, this would be likely to bias coding in the direction of localism. If on the other hand every spike is important, as seems reasonable for metabolic reasons, then some degree of distribution of coding must be taken into account. What is important is that we understand where the operating point for natural behaviour lies, which is the task that the system has evolved to undertake. Evidence of the task-dependence of the sparsity of coding and of the flexibility of decoding strategies raises a further intriguing hypothesis: that the degree of distribution of coding is actually dynamic, and may adapt to the nature of the perceptual task at hand.

ACKNOWLEDGMENT

I gratefully acknowledge many useful discussions with Andrew Metha on this topic.

An extended local connectionist manifesto: Embracing relational and procedural knowledge

Lokendra Shastri

International Computer Science Institute, Berkeley, CA 94704.
shastri@icsi.berkeley.edu **www.icsi.berkeley.edu/~shastri**

Abstract: Page has performed an important service by dispelling several myths and misconceptions concerning the localist approach. The localist position and computational model presented in the target article, however, are overly restrictive and do not address the representation of complex conceptual items such as events, situations, actions, and plans. Working toward the representation of such items leads to a more sophisticated and articulated view of the localist approach.

Page has performed an important service to the connectionist and cognitive science community by dispelling several myths and misconceptions concerning the localist approach. Over the years, these myths and misconceptions have acquired the status of self-evident truths and it has become almost obligatory for connectionist models to display the "distributed representations inside" logo in order to lay claim to being the genuine article.

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?

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

Visser et al. (1999) uses hidden Markov models to characterize learning of human subjects. By analyzing a series of responses it is possible to extract a hidden Markov model that is, in its general form, closely related to the sequence of stimuli that were used in the sequence learning experiment. In fact a hidden Markov model is a stochastic version of a finite state automaton, the kind of automaton used by Cleeremans and McClelland (1991) to generate the stimuli for their implicit sequence learning experiment.

Such a procedure can also be used in analyses of a neural network by having the network generate a series of responses or predictions. Using a version of the EM algorithm a hidden Markov model can be extracted from the network (see, e.g., Rabiner 1989). Extraction of a hidden Markov model of the network partitions the state space of the hidden layer of the network in discrete (hidden) states. This model is then interpretable in the sense that the presence or absence of connections between states indicates which sequences of stimuli are admissible and which are not; that is, the states can be regarded as statistically derived proxies of local representations. In addition, the extraction procedure does not rely on inspection of the activities at the hidden layer of the network as is done for example by Cleeremans et al. (1989) and Giles et al. (1992).

Extraction of hidden Markov models, and by implication of finite state machines, can in principle be applied to any neural network but is especially suitable for those that are used for modelling sequential behaviors. This is not to say that those networks should be replaced with HMMs. For example the work of Cleeremans and McClelland (1991) shows that their SRN model is very successful in describing subjects' behavior in implicit sequence learning. Although I strongly support Page's manifesto for localist modelling, it does not solve all problems of interpretation that arise in neural networks. Hidden Markov models are a highly useful tool to gain a better understanding of the internal workings of such networks in terms of proxies of local representations.

A population code with added grandmothers?

Malcolm P. Young, Stefano Panzeri, and Robert Robertson

Neural Systems Group, Department of Psychology, University of Newcastle, Newcastle upon Tyne, NE1 7RU, United Kingdom. **m.p.young@ncl.ac.uk**
www.psychology.ncl.ac.uk/neural_systems_group.html

Abstract: Page's "localist" code, a population code with occasional, maximally firing elements, does not seem to us usefully or testably different from sparse population coding. Some of the evidence adduced by Page for his proposal is not actually evidence for it, and coding by maximal firing is challenged by lower firing observed in neuronal responses to natural stimuli.

Neurophysiologists have for some time entertained the idea that there is graded dimension between grandmother cell codes and very coarse population codes (e.g., Fotheringhame & Young 1996). At one extreme, all information about a stimulus is to be found in the firing of a very small number of neurons, and very little is to be found in the firing of the many cells that are relatively unexcited by it. At the other extreme, all information is to be found in the graded firing of very many cells, and very little in the firing of any single cell. The idea of this "sparseness" dimension has the virtue of suggesting readily testable measures of where neurological reality lies along it, such as the ratio between information in the population at large and that in single unit's responses. Experimental investigation of the various proposals that have populated this dimension has been quite fruitful, and it is not often now contended that there is information carried by the unenthusiastic responses of large numbers of cells. It is also not often contended that sufficient information for most behavioural tasks can be carried by small numbers of cells, some tens of them in several cases (e.g., Shadlen & Newsome 1998; Young & Yamane 1992; 1993). Hence, in visual neuroscience, at any rate, there seems presently not to be great conflict on this particular issue (as once there certainly was): codes appear to be populational, but the populations involved can be very small indeed, particularly toward the top of any sensory system.

Page's thinking seems in one way different from that outlined above. His redefinitions of key terms, such as "localist," are different from those that have exercised experimenters over the last decade or so, and the experiments have tested different ideas than his. Page's idea (hereinafter "localist" with scare quotes) seems to be that a "localist" code possesses both distributed and grandmotherish aspects, rather than supposing that it is one or the other, or some specifiable balance between the two. In this sense he is mistaken that the "localist" position has been "misconstrued," and that there is a widespread tendency to assume that neurophysiological results speak against "localist" representation. His particular avowed position has not previously been in the minds of experimenters. Page's comments about extracts from Young and Yamane (1992) illustrate this mistake particularly clearly. The fact that IT neurons' suboptimal responses are systematically graded, and not random, indicates that information is carried at the population level. The idea ruled out by these results is that of extreme, single-cell coding, in which all of the information is carried by grandmother cells, and which was very much alive at the time, but now is not. Information being carried at the population level (i.e., nonrandom, suboptimal responses) does not bear on Page's new and versatile redefinition of "localist."

In another way, however, we wonder whether his "localist" position is *really* anything usefully different from the sparse population coding that some cortical neurophysiologists think they see, notwithstanding his various comments to the contrary. Page's "localist" code seems to be a population code with occasional, maximally firing, highly informative elements – which appears to us to be very similar to a population code in which a fairly small proportion of the population carries most of the information. The only difference seems to be that "localist" coding commits itself to the claim that there should be at least one maximally firing cell for every familiar stimulus, whereas sparse population coding doesn't care one way or the other whether any cell is firing maximally. If this is the only difference, it is not a very useful or testable one for brain scientists: How could one know whether a cell is firing maximally? For example, the apparently very selective cell reported by Young and Yamane (1993) is the best evidence for "localist" coding that Page can conceive. But "localist" coding involves "the presence of at least one node that responds maximally," and this cell's modest 37 Hz is most unlikely to be its maximum. Many IT cells respond above 100 Hz in the same circumstances and, given sufficient time to characterise its preferences, there is no reason to suppose that this cell could not be driven three times as hard. Hence, this is no evidence at all for "localist" coding. If there are more important differences than the claim of maximal firing, Page should specify in what way "localist" is different from sparse population coding, and in sufficiently clear terms so that we could test which scheme it is (perhaps in IT). We think that being right and being wrong are about equally valuable, while being unspecific or untestable is less valuable. Being right provides a useful foundation for further work to build on; being wrong is useful because in finding out you're wrong, generally you or your rivals find out a lot of interesting things; but being unspecific, untestable or unstimulating doesn't get anyone anywhere. Meeting the challenge of specifying testable distinctions would reassure us that Page's thesis does not reside in the latter category.

Finally, there is increasing evidence that neurons respond at lower rates when processing natural scenes, or indeed any other scenes about which the visual system can apply prior knowledge (e.g., Baddeley et al. 1997; Gallant et al. 1998; Scannell & Young 2000), when compared to the extra-ordinary stimuli traditionally employed by neurophysiologists and psychophysicists. This chal-

lenges Page's idea of coding being undertaken by maximally firing units, as the lower firing rates observed in response to natural stimuli carry more information (Buracas et al. 1998; Rieke et al. 1996), and there are strong theoretical reasons to suspect that generally lower firing accompanies normal vision (Rao & Ballard 1999; Scannell & Young 2000). These reasons are connected to the fact that neurons can take advantage of prior knowledge of the world, and make inferences on it; suggest that visual neurophysiology is again entering an unusually exciting time, in which many of the old certainties are giving way; and that representation and processing of visual information is increasingly unlikely to be mediated in the ways championed by Page.

Author's Response

Sticking to the manifesto

Mike Page

Medical Research Council Cognition and Brain Sciences Unit, Cambridge, CB2 2EF, United Kingdom. **mike.page@mrc-cbu.cam.ac.uk**
www.mrc-cbu.cam.ac.uk/

Abstract: The commentators have raised some interesting issues but none question the viability of a localist approach to connectionist modelling. Once localist models are properly defined they can be seen to exhibit many properties relevant to the modelling of both psychological and brain function. They can be used to implement exemplar models, prototype models and models of sequence memory and they form a foundation upon which symbolic models can be constructed. Localist models are insensitive to interference and have learning rules that are biologically plausible. They have more explanatory value than their distributed counterparts and they relate transparently to a number of classic mathematical models of behaviour.

R1. Introduction

The commentators raise some interesting points regarding the benefits, or otherwise, of a localist modelling approach. I have divided my reply into a number of sections that broadly reflect concerns shared by more than one commentator. There are sections on definitional matters, on distributed processing over localist representations, and on the distinction between prototype and exemplar models. Other topics include models of sequence memory, catastrophic interference, and the interpretability of connectionist models. I briefly discuss the notion of efficiency with regard to biological systems and address the issues raised by those commentators involved in mathematical modelling. On occasion I have stretched the section headings somewhat to allow me to deal in one place with different issues raised in a single commentary. In addition, I have devoted a whole section of my reply to the commentary by **Plaut & McClelland:** these commentators are two of the principal proponents of the fully distributed modelling approach and I felt it necessary to address their arguments in a very direct way. I begin by tackling some problems regarding the definition of terms.

R2. The concept of a localist model has not been redefined, nor is the proposed model a hybrid model

Several commentators (**Beaman; Farrell & Lewandowsky; Plaut & McClelland; T. Shultz; Young et al.**) make one, or both, of two related points. First, that the concept of a localist model has been redefined in the target article. Second, that this new definition permits the inclusion of distributed representations at many (though not all) levels of a localist model and that it is this fact that underlies any advantageous aspects of the "hybrid" model's performance. For example, **Young et al.** find my redefinition of key terms such as "localist" so "new and versatile" that they recommend placing the word within scare quotes. Beaman notes that "many of the criticisms Page anticipates are neatly circumvented by his original definition that localist models can use distributed representations." The impression given by these commentators is that "redefining" localist representations in this way is not quite fair play, allowing the localist modeller to enjoy the benefits of distributed representation without properly subscribing to a fully distributed approach.

This criticism is unsustainable, not least because no *redefinition* of the term "localist" has been either attempted or achieved in the target article. My definitions of localist representation and of a localist model were intended to be clear rather than novel; indeed, it is difficult to see where the above commentators identify the supposed novelty. From their commentaries one might surmise that it is the inclusion of distributed representations in localist models that is novel but, as is stated in the target article, localist models have always included distributed representations of entities at all levels "below" that at which those same entities are locally represented. Taking the same example as I did in section 2.6 of the target article, namely, the interactive activation (IA) model of visual word recognition (McClelland & Rumelhart 1981), this localist model contained distributed word-representations at both letter and visual feature levels. IA was nevertheless a localist model of words because it also contained a level at which words were locally represented. The idea that a model that includes both distributed representations and localist representations of a given class of entity should best be described as a hybrid model misses the point: a localist model of a given entity class is defined (and has been for a long time) by the presence of localist representations of members of that class, not by the absence of their distributed representation.

Young et al. state that "[Page's] particular avowed position has not previously been in the minds of experimenters." If my position is no more than a clear statement of a long-held localist position, then this raises the question of which localist position has been experimentally (i.e., neurophysiologically) tested. If, as **Young et al.** imply, experimenters have been interested solely in ruling out "grandmother-style" localist representation, in which no information is transmitted by suboptimal responses, then there has been a tragic lack of communication between the experimental and the connectionist theoretical work. As far as I can tell, few, if any, localist connectionist modellers in the last thirty years have believed in such a style of representation.

To illustrate again the important point regarding the informational content of "neural" responses, suppose one has a localist representation for each familiar orthographic

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

word form, such that each activates to a degree dependent on its orthographic similarity to a test word. So presented with the word "MUST," the node representing the word "MUSE" (henceforth the MUSE-node) will activate to degree 0.75 ± 0.05; the same node presented with the word "MAKE" will activate to 0.5 ± 0.05 and presented with the word "MUSE" it will, of course, activate optimally at 1.0 ± 0.05. Now suppose we present one of these three words while measuring the activation of this MUSE-node and find that it activates to degree 0.8. Does this nonoptimal response give us any information about which of the three words was presented? It is clear that it does – the test word is most likely to have been "MUST." Now suppose the test word could have been any four letter word – in this circumstance a MUSE-node activation of 0.8 would not help us decide between, say, MULE, FUSE, or MUST, though we would still have enough information effectively to rule out MAKE as the test stimulus. More information will be derived if activation of another node, for example the FUSE-node, were simultaneously measured at, say, 0.98. Such levels of activation are only expected if the orthographic match is perfect. Thus, highly active nodes are informative because high activation is rare. In circumstances where a very low activation is rare, low activations can also be highly informative. In general, though, less active nodes are, individually, rather less informative because their activation can be caused by a larger number of different states of affairs. Remembering that there are likely to be many more partially active nodes than fully active ones for any given test stimulus (see commentary by **S. Schultz**), it is easy to see how any reasonable localist model would be consistent with a finding that a good deal of stimulus information is present in nonoptimal responses.

Young et al. contend that the distinction between localist coding and sparse population coding is not a very useful or testable one. There is one sense in which I agree: if a sparse population code (that remains undefined by **Young et al.** in spite of their recommending clear, testable definitions) is defined as one for which a small number of nodes activate to varying and differing extents to various stimuli, with the more active cells carrying somewhat more (per cell) of the stimulus information, then a localist code is a particular kind of sparse distributed code and it will be difficult to get data preferring one over the other. It is only if a sparse population code is defined exclusively as a code that is not localist, that is, a code that in **Young et al.'s** phrase "doesn't care one way or another whether any cell is firing maximally," that the difference becomes testable. Indeed, it not only becomes testable but also, to some extent, tested. I refer to the work of Sakai et al. (1994), described in section 6.5.2 of the target article, about which **Young et al.** are, perhaps understandably, silent. (Though they both misquote and misapprehend when they claim that the Young & Yamane, 1993, result is the best evidence for localist coding that I can conceive.) Sakai et al. have sought to test whether recently learned and otherwise arbitrary patterns are represented by cells in IT cortex that are maximally active to the trained patterns and thus less active in response to minor variations of those patterns. In **Young et al.'s** view there needn't be any such cells, let alone at least one for each learned pattern as the results indicate. Being testable and specific is indeed of advantage when it comes to hypotheses – making the right predictions is an added bonus. Doing so while stating clearly how a given code might be decoded in vivo (another feature lacking from **Young et al.'s** characterization of sparse population coding) is better still.

A final point on taxonomy. **Molenaar & Raijmakers** observe that variation of a parameter representing inhibitory range in an exact implementation of ART causes a bifurcation in the behaviour of the network. They label this bifurcation as a transition between localist and distributed coding and use this to infer that there is a genuine dichotomy between the two coding styles. While I agree that there is such a dichotomy, I find their argument circular. In the very act of labelling a transition as being one from localist to distributed coding, one is assuming a dichotomy between the two. One cannot then go on to use this characterization as evidence that the dichotomy is, in some sense, real. Perhaps I have missed the point here.

R3. A pattern of activation across localist representations can indicate uncertainty but does not imply a fully distributed model

Another point that arises in a number of commentaries, relates to the fact that winner-takes-all dynamics is not a necessary feature at all levels of a localist model. For example, **Grossberg** refers to the masking field model of Cohen and Grossberg (1987) as one that weighs the evidence for various parts being present in a multipart input pattern, such as that corresponding to a sequence of words. I am less convinced than **Grossberg** that the masking field, as originally implemented, is as powerful as he suggests – my concerns relate particularly to learning issues that are avoided by Cohen and Grossberg in favour of a "developmental prewiring" scheme that is clearly impractical in general and that is abandoned in later, and more implementationally explicit, developments of the masking-field principles (Nigrin 1993; Page 1993; 1994). Nonetheless, it is certainly the case that such evidential weighing can be envisaged as resulting in a pattern of activation across different candidates. This should not conceal the fact that the evidence-based distribution of activation is over localist representations of the candidate words (cf. Shortlist, Norris 1994b). Moreover, when and if the time comes to decide which words are most likely present, those candidates that are mutually inconsistent (i.e., that seek to account for some common portion of the input pattern) must indeed compete in some way before a best-guess parsing can be made. In localist modelling it is usually appropriate to reserve winner-takes-all dynamics for a level(s) and a time at which a categorical response must be made. The level and time at which such dynamics are initiated will be determined by the task demands. Thus, a layer that is capable of winner-takes-all dynamics might not be obliged to employ such dynamics at all times and in all tasks. Under the appropriate task demands, such a layer might process a stimulus as a distributed pattern of activation across localist representations. This is consistent with **S. Schultz's** reference to recent work by Newsome (1999). In tasks such as velocity matching in visual tracking (e.g., Groh et al. 1997), where the response is an analogue quantity rather than a categorical one, generalization by weighted activation would be a more reasonable strategy than one based on winner-takes-all dynamics.

Phaf & Wolters rather confuse the issue when they observe (para. 1) that a localist model of word recognition will

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

exhibit "distributed representations when exposed to four-letter nonwords." Certainly, such exposure will result in a distributed pattern of activation over the word layer, but this is hardly relevant to whether the model is localist at the word level. Being localist at the word level requires that each known (i.e., learned) word results in the maximal activation of at least one word node. Indeed, learning in a localist model comprises the establishment of such a tuned word node. It makes no sense to talk about a model that is localist at the nonword level, since this would require that all known nonwords – a meaningless concept – have a local representation. **Golden** (para. 4) makes the same confusion when he effectively raises the possibility that a pseudoword (word-like nonword) might be considered a "familiar example" of the entity-type "word," which by definition it cannot. The confusion partly springs from **Golden's** omission of the words "of a particular type of entity" from the definition quoted in his second paragraph. A localist model of words will not represent nonwords locally, neither would one expect it to. To answer **Golden's** question regarding his own model, I would consider it distributed both at the letter level and at the word level. As a general model of letter/word recognition, however, it has most of the problems of fully distributed attractor models to which I allude in the target article.

R4. Localist representations must be defined relative to meaningful entities

Berkeley finds my definition of a localist model unsuccessful. He illustrates this with the example of a network trained to validate and categorize logic problems (Berkeley et al. 1995). Subsequent analysis of one of the hidden units revealed that it activated within one of three activation bands (defined post hoc) depending on whether the logical connective was OR, IF . . . THEN, or NOT BOTH . . . AND . . . (in order of increasing activation, the last being assumed to be maximal activation). **Berkeley** claims that by my definition of localist, the node should be deemed a NOT BOTH . . . AND . . . detector, whereas it is most "naturally" interpreted as a connective detector. Although it is true that one can discern the current connective by examining the state of this hidden unit, whether one would really want to dub such a unit a "connective detector" is debatable. Suppose that there were only two types of animals, a dog and a cat, and that you possessed a node that activated strongly when presented with a cat, but weakly when presented with a dog. Would you call the node a "cat detector," an "animal detector" (on the grounds that you can infer the presence of either cat or dog from its level of activation) or perhaps even a "dog detector"? I prefer the former and certainly find it the most "natural" description, but I am not sure it matters much. The "X-detector" terminology is not part of my definition. A localist representation of item X is defined as a representation that activates maximally in the presence of X. A localist model of the class of entities to which X belongs, has a localist representation for each learned entity in the class. Using this definition, Berkeley et al.'s (1995) network does happen to be a localist model of connectives because it has at least one hidden node that is maximally active for each of the three connectives.

As **Hummel** points out, even the hidden units in a fully distributed model are "localist" in the weak (and untenable) sense that there will be some vector, or set of vectors, to which any given hidden unit responds maximally. The model will only be meaningfully localist with respect to a given entity class if this optimal vector, or set of vectors, corresponds to an example of that class, with all learned members of that class being similarly represented. **Barlow & Gardner-Medwin** make a similar point with regard to the ASCII code for computer characters. For any given bit of the 8-bit code, there will be a set of letters for which that bit is set to 1. Unless the members of that set match some other meaningful criterion (e.g., they are all vowels) then it is not useful or appropriate to describe that bit as a localist representation of members of the set.

R5. On exemplars and prototypes

Several commentators (**Carpenter; French & Thomas; Ivancich et al.; T. Shultz**) discuss the benefits of a prototype coding of categories. Ivancich et al. describe a number of so-called "prototype" effects, for example, the fact that prototype patterns are easily recognized even if these have never been previously encountered. My reading of the large literature on prototype versus exemplar coding suggests that the issue is far from as straightforward as these commentators assume. For example, there have been a number of demonstrations that exemplar models are perfectly able to simulate most, perhaps all, of these "prototype" effects (see e.g., Lamberts 1996, and the literature reviewed therein; Dopkins & Gleason 1997). Good recognition of a previously unseen prototype can be explained by its relatively large mean similarity to exemplars in its class. The apparent problem noted by **T. Shultz** (para. 5) would not apply in a high vigilance regime when the localist model described in the target article behaves like a classic exemplar model. The contrast between prototype, exemplar, and decision-bound models is the focus of much current work, both experimental and computational. Although it would be unwise to attempt a synopsis here, it is my impression that exemplar models are more than holding their own.

In recent work, Smith and Minda (1998) and Nosofsky et al. (1994) have sought to moderate the rather strong exclusivity that has characterized the prototype/exemplar debate. The former suggest that both prototype- and exemplar-based mechanisms might be employed depending on factors such as the category structure and the stage at which learning is tested. The latter describe a "rule-plus-exception" model in which patterns are classified using simple logical rules, with exceptions to those rules being individually stored and applied appropriately. Given the mathematical analysis presented in the target article, neither of these models would present too great a challenge with regard to their localist connectionist implementation. Indeed, there is nothing in the localist position generally, nor in the generalized model that I present, that forbids the use of prototype or rule representations (a point also made by **Jelasity**). Some confusion may have been caused in this regard when, in the passage in the target article describing the learning rule (sect. 4.4), I referred to the effective prevention of further learning at a committed node. This requirement is more stringent than is necessary in general and it was suggested specifically with reference to the implementation of an exemplar model, for which it is appropriate. By lowering vigilance and allowing some slow modulation of the weights

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

to a committed node, something more akin to prototype learning is made possible within the localist framework (see **Carpenter; T. Shultz**). In section 7.1 of the target article, I suggest that the distinction between exemplar and prototype might be mirrored by a dual-store (hippocampal-cortical) memory system – the slower store would be dedicated to the evolution of context-free and hence more prototypical representations, as is commonly assumed.

Of course, prototype or rule establishment are not necessarily slow processes. For example, Smith and Minda (1998) suggest that prototype coding dominates in the *early* learning of more highly differentiated categories. Models such as those in the ART family, very similar to the one I describe, are able to establish and refine localist representations of category prototypes (see **Carpenter's** commentary). One principle that has been employed in the ART2A (Carpenter et al. 1991) network, is that of fast commitment and slow recoding. This involves a new pattern, sufficiently different from any established category, being able to become represented quickly by a previously uncommitted node. Subsequent modification of the incoming weights to this node during the presentation of sufficiently well-matched patterns, can be slowed so as allow the weights to track a running average (or prototype) of the patterns that excite it. Implementation of fast learning at an uncommitted node together with slow learning once committed might involve no more than modulating the learning rate, λ, in Equation 10 by some decreasing function of the sum of its bottom-up weights (see sect. R9). Such a mechanism would allow the category drift referred to by **Ivancich et al.,** but is not the only way to do so. Category drift might also be modelled using an exemplar model in which older exemplars drop out while new exemplars are learned. This is because, contrary to **Carpenter's** interpretation (**Carpenter** para. 3), every category in an exemplar model is represented by the many exemplars or instances to which that category node is connected. Such a mechanism would cause the "implied prototype" (not itself represented) to drift in the direction of more recent exemplars of the relevant class.

Ivancich et al. are right that localist category learning can be difficult in circumstances where multidimensional input stimuli include many dimensions that are either irrelevant, or even positively unhelpful, to the categorical distinction that is being learned. Bishop (1995) makes a similar point (with regard to the application of radial basis function networks, p. 184) and **Plaut & McClelland** raise related concerns in their commentary. To be more specific, while an exemplar or instance model is guaranteed to be able to learn the members of any given training set, simply by learning the instance and connecting the relevant node to the corresponding category, if many of the dimensions across which this instance is defined carry information related to some other aspect of the stimulus, then generalization to new instances is likely to be poor. Kruschke (1992), in his ALCOVE model, implemented a system that learned to focus "attention" only on dimensions relevant to the task in hand. In that implementation, the attentional learning was carried out using a backprop-like learning rule. I am currently working on a similar attentional learning system that uses a learning rule more consonant with the localist networks espoused in the target article.

French & Thomas have a different criticism of the localist model of category learning. They seem to assume the opposite view to that of the commentators described above, in that they assume that the only hardware available for representing, say, the categories "fork" or "chair" is a single, appropriately labelled node for each – what one might call two ultraprototypes. Given this assumption, they reflect that such a localist model and, indeed, localist models in general must lose information relating to category variability. I find their position somewhat puzzling. The model I describe in the target article is capable of being run as an exemplar, or even an instance model with each instance being learned before the relevant node (a unitized representation of that instance) is connected to another node representing the category label. Assuming the model is run as a high-vigilance exemplar model, a large proportion of all chairs so far experienced will be represented in memory and associated with the chair category. Even in more realistic circumstances, when vigilance is relaxed somewhat, one would expect many different subcategories of chair to be so represented. Not only that, but one would further expect the number of memorized subcategories of the category "fork" to be rather fewer, given the presumed relative lack of variability in that category. This difference in the number of nodes devoted to each category is consistent with **French & Thomas's** explanation for category specific anomias, though the fact that the reverse deficit (inanimate named worse than animate; e.g., Lambon-Ralph et al. 1998; Tippett et al. 1996) can also be found suggests that this issue is not as straightforward as they maintain. In the limit, the exemplar model described in the target article is capable of representing in memory exactly as much variability as exists among the learned examples of a given labelled category. Not only that, but it is capable (with a little extra mechanism) of doing as **French & Thomas** do, and generating (i.e., listing) a good number of specific instances of different types of chair. One wonders how the generic (though rather ill-specified) distributed network that **French & Thomas** prefer might manage that.

French & Thomas also summarize some data from Vogels (1999) and make much of the fact that "no individual neuron (or small set of neurons) responded to all of the presented trees, while not responding to any non-tree." I would say only this: when single-cell recording, it is extremely easy *not* to find the cells participating, perhaps in a localist fashion, in a given task. It would continue to be difficult even if one knew those cells to exist.

R6. Models of sequence memory

Both **Visser** and **Kinder** raise questions about the capabilities of localist models with respect to memory for sequences. Visser describes the simple recurrent network (SRN) used by Elman (1990) and Cleeremans and McClelland (1991) in the modelling of sequence learning. **Visser** points out that both used localist representations of the sequence items, but further notes that the representation at the hidden units, presumed to encode the learned sequential dependencies, is a distributed one and hence rather resistant to interpretation. **Visser** suggests the extraction of a hidden Markov model of the network that permits its description in terms of transitions between hidden states, each state corresponding to a configuration of the hidden units. **Kinder** also refers to the SRN in the context of artificial grammar learning (AGL). She too notes the dis-

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

tributed representation of sequence information at its hidden layer.

With regard to the simple recurrent network I agree with both **Visser** and **Kinder** that the sequence information is represented in a distributed fashion. As a consequence, the SRN is subject to many of the criticisms of fully distributed networks evinced in the target article. Most notably, the SRN has a serious problem with catastrophic interference and with the plausibility of its learning rule. Other connectionist models of general sequence learning (e.g., Nigrin 1993; Page 1993; 1994 – both based on the masking field principles cited by **Grossberg**) allow for localist representations of lists, or more particularly list-chunks, in addition to the localist representation of list items. A sequence of list chunks can itself be learned, with the lists of lists equivalently represented by higher-order chunk nodes. Page (1994) applied such networks to the learning of simple melodies and to the musical expectations that can be evoked by the incomplete presentation of familiar melodic phrases. These networks have none of the problems of interpretation or interference of their SRN counterparts. Some of the principles of sequence representation used in this work, and in the earlier work of Grossberg (1978b), have been used in the modelling of short-term memory (STM) for serial order (e.g., Page & Norris 1998). It is interesting that in this field there is no experimental evidence that supports the "chaining-style" sequential representation implicit in work using the SRN (Henson et al. 1996). While the localist networks described above have not, to my knowledge, been applied to the AGL task, it would seem premature to be as confident as is **Kinder** that modelling this task requires distributed networks. This is particularly so when one considers that much of the experimental work that has been used to support the "grammar learning" interpretation does not clearly favour such an interpretation over one based on the learning and subsequent recognition of familiar lists/list-chunks (see e.g., Johnstone & Shanks 1999 for a recent review). This is precisely the sort of function for which the localist alternatives were developed.

Bowers points out that several of the more successful connectionist models of short-term memory for serial order are localist models. He also notes the requirements of the task, namely, the repetition of a list immediately after one presentation, are not particularly suited to simulation using distributed representations and networks such as that used by Elman (1990). **Beaman,** however, seems rather confused about the use of localist representations in models such as those found in Burgess and Hitch (1999) and Page and Norris (1998). These models assume that order is held over unitized representations of list items (e.g., words). Obviously this requires that these unitized items must be "unpacked" during the generation of, say, a spoken response. Such phonological unpacking is commonplace in models of speech production, where the intended utterance is represented initially across localist representations of lexical items (e.g., Dell 1986; 1988; Levelt 1989). Indeed, in Page and Norris (1997) we used a Dell-like speech production stage as the output stage for our memory model – hardly an optional extra, as **Beaman** implies. It does not contain "distributed phonological representations" (see **Beaman,** figure caption) but localist representations of output phonemes driven by localist representations of the words of which they form part. Both our and Burgess and Hitch's (1999) short-term memory models assume such a two-stage process of recall, with order being initially determined over unitized items and with possible additional componential (e.g., phonemic) errors occurring during the unpacking process. This is to allow the models to simulate some otherwise difficult data relating to the phonological similarity effect (PSE; Baddeley 1968; Henson et al. 1996). Other models that do not adopt this strategy have so far been unable to simulate these data. For example, **Farrell & Lewandowsky's** suggestion that interference caused by "anti-Hebbian" suppression might underlie the PSE, fails to acknowledge that the PSE has its locus primarily on order errors rather than item errors.

R7. Catastrophic interference

Carpenter illustrates by reference to the dART model (Carpenter 1997; Carpenter et al. 1998) that catastrophic interference (CI) is not an inevitable consequence of using fast learning to associate distributed patterns (see also **Bowers**). The dART framework contains several features novel to connectionist modelling and is sufficiently complex that I cannot honestly claim to be in a position to judge here its strengths and its weaknesses (if any). Nonetheless, as **Carpenter** makes clear in her commentary, there are benefits in having additional localist representations in a model beyond the avoidance of catastrophic interference. Some of these benefits were discussed in sections 7.2–7.6 of the target article.

Farrell & Lewandowsky claim that there are more conventional distributed solutions to the problem of CI and cite Lewandowsky (1994) in support. In fact, most of the networks investigated by Lewandowsky, which employed versions of the BP learning algorithm modified in various ways to reduce overlap in hidden unit representations, were not particularly good at avoiding CI. The best used Kortge's novelty rule, which can be applied only to autoassociative networks (or approximations thereof), and Lewandowsky showed that this functioned by transforming the input patterns so that their hidden-layer representations were mutually near-orthogonal. **Jelasity** also describes a distributed model that, nonetheless, requires that its representations are mutually orthogonal if interference is to be minimized. This functional orthogonality is what is achieved in localist networks such as those described in the target article. The localist solution is more general than that using the novelty rule in that it is applicable to tasks other than autoassociation. Moreover, although **Farrell & Lewandowsky** showed that ability to generalize is impaired in certain localist models (such as ALCOVE implemented with sharply tuned "hidden" units), the same problem does not occur in models like that suggested in the target article for which the functional gradient of generalization can be varied. There is no justification, therefore, for **Farrell & Lewandowsky's** conclusion that consideration of CI and generalization favours distributed models over localist.

Farrell & Lewandowsky make a number of other points that I shall address here for convenience. They point out that my criticism of the biological plausibility of the learning rule is focussed on the BP learning rule. I make the same point in the target article, noting that the BP learning rule has, nonetheless, been the one most often employed for fully distributed models. **Farrell & Lewandowsky** favour a Hebbian learning rule, again echoing the target ar-

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

ticle. They do not acknowledge properly that the Hebbian learning rule applied to, say, the association of distributed representations, can lead to implausible levels of crosspattern interference. **Farrell & Lewandowsky's** point about the Rescorla-Wagner (RW) learning rule is invalid: not only does the RW theory often apply to situations in which target patterns are locally represented, but it also suffers from catastrophic interference (López et al. 1998). More recent developments of similar models (e.g., Pearce 1994 – a localist model) do not share this problem.

Farrell & Lewandowsky refer to Miikkulainen's (1996) "distributed linguistic parser" as an existence proof that distributed schemes are not at a disadvantage in this area. Miikkulainen's work is indeed impressive but I would hesitate to draw **Farrell & Lewandowsky's** conclusion. The sphere of application of Miikkulainen's network is quite limited, and a good deal of special purpose control structure is built in. Miikkulainen acknowledges (p. 63) that the way in which the model learns is nothing like the way in which the equivalent facility is acquired in vivo. Moreover, the structure of the resulting network suggests that it succeeds in its task by acquiring the ability to copy word representations into role-defined slots, with the relevant channelling closely specified by short word sequences (e.g., "who the") in its highly stereotyped input sequences. It is repeatedly asserted by Miikkulainen that the network is not an implementation of a symbolic system but the very opacity of the learned network leaves the reader unable to assess the justifiability of this claim. Because we are not told in sufficient detail how the network actually functions in carrying out its role-assignments, we are left wondering whether, on close analysis, it would prove isomorphic to a symbolic system, in which copy operations were conditionally triggered by well-specified word sequences. The fact that parsing is helped somewhat by semantic attributes encoded in the distributed input patterns to the network is used by Miikkulainen to emphasize the subsymbolic nature of the resulting processing. It would be a mistake, however, to believe that such gradedness is the sole preserve of the distributed subsymbolic model (see commentaries by **Hummel, Shastri,** and **Ohlsson**). It would also be instructive to confirm that the network was not over-reliant on semantics to the extent that the case-role parsing of sentences with either unusual meaning (man bites dog) or no meaning (the plig wugged the hab) remained possible.

R8. The interpretability of localist models

Barlow & Gardner-Medwin, Cook, Leth-Steensen, and **Martindale** all identify the interpretability of localist models as one of their major advantages. Martindale gives as an example a model of restoration of the missing fundamental to a musical tone, his suggested model being very similar to those proposed by, for instance, Taylor and Greenhough (1994) and Cohen et al. (1995). Cook even accuses me of having been a little gentle on the negative aspects of fully distributed representations. If I have done so then it is because I feel uncomfortable in defending the localist position simply because it is easier to understand. To give an example from physics, the quantum theory of matter is notoriously hard to understand, certainly more difficult than the classical theories that preceded it. Nobody would conclude from this that classical theories are actually better than quantum theory as a description of reality. Quantum theory is held to be a better theory, despite its additional complexity, because it accounts for more data in the realm of its application. For this reason, I am reluctant to overplay the interpretability argument. It is at least possible that the brain functions, like fully distributed networks, in a way that is difficult to interpret (**Farrell & Lewandowsky** make a similar point). It is also possible, however, that the brain turns out to be rather more interpretable, and perhaps for Jordan 1990 (cited by **Leth-Steensen,** para. 3) less "interesting," than has been widely imagined. For the more complex and less explanatory accounts to be attractive, one would like them to be bolstered either by sound arguments from the point of view of biological plausibility, or (as for quantum physics) by clearly favourable experimental data, or preferably by both. At the moment we have neither.

Leth-Steensen also suggests that a general disdain for localist modelling among some members of the cognitive community has emerged because of the "hand wiring" of some of the better known localist models. His claim is that models using distributed representations start from the "very strong position" that learning of those representations is a fundamental part of the learning process. This claim is weakened, however, if one considers that most applications of fully distributed modelling techniques are set up extremely carefully. Sometimes they employ featural (i.e., localist) representations (whose existence their learning theory cannot begin to explain) in their input and output representations (e.g., Elman 1990; Plaut et al. 1996). Sometimes, as in the passage quoted from Farah (1994a) in section 7.2 of the target article, the learning rule used is explicitly disavowed as a plausible biological mechanism. This leaves us very much, as **Leth-Steensen** puts it (para. 2), "in the position of staring at a set of network connections and their weights wondering how they got to be that way." In the target article I attempted to lay out some general strategies for the learning of localist representations: this review was in no way exhaustive. So, while I agree with **Leth-Steensen** that the scaling up of localist models is a challenging exercise (though somewhat less challenging than the scaling-up of fully distributed approaches), I am optimistic that it can be achieved.

On the question of interpretation, I concur with **Barlow & Gardner-Medwin's** view of localist representations as closely analogous to matched filters. This analogy sharpens the discussion in section 7.4 of the target article, regarding the establishment of *what* stimulus is present and *when* that decision can be made. For an array of matched filters overlooking a given stimulus pattern at a lower layer, "what" can be determined by looking for the best-matched filter, and "when" can be determined by the time at which the activation of that filter meets some criterion (see e.g., Hanes & Schall 1996; and Carpenter & Williams 1995 for experimental evidence supporting similar models of reaction time in vivo).

Finally, **Burton** makes a number of points relating to definition and interpretability. I agree to such an extent that I have nothing useful to add to his comments.

R9. The implementation of symbolic models

Hummel, Shastri, Ohlsson, and **Bowers** all see localist connectionist models as a useful first step in the building

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

of symbolic and relational models that "form the grist of human cognition" (**Shastri,** para. 2). I agree with their analysis. Clearly, the modelling of "higher cognition" introduces complexities that could only be mentioned briefly in the target article (e.g., sects. 2.2 and 7.3). I am therefore grateful to these commentators for making more explicit the nature of this added complexity. What they all agree is that localist models form a natural and convenient foundation on which to build models of higher cognition. **Hummel** attributes this advantage to the separability of localist representations that permits generalization across symbolically defined classes. He gives the example of a network applied to the classification of red shapes differently from blue shapes and notes that a localist (or otherwise separable) solution is readily achieved by connecting a red-node and a blue-node to nodes corresponding to their respective categories. It is my feeling that getting a localist network to learn to ignore the shape dimension (among others) and to base its categorization solely on colour, is an attentional learning task that is quite challenging for the sort of localist models I describe, at least in the general case. The challenge can be seen clearly in the case where the input stimulus is a vector in which those dimensions relevant to colour are heavily outnumbered by, say, those relevant to shape. Unless attentional weighting can be learned, the "closest" stimulus to a (novel) red star, will be likely be a (learned) blue star, leading to an incorrect classification. As noted in section R5, this attentional learning is the subject of current work.

Shastri also gives more references to "recruitment learning" and explores the overlap with the learning mechanisms suggested in the target article. In particular he identifies more clearly than I had, the fact that recruited, or what I called "committed" nodes, can be distinguished from free nodes on the basis of the strength of their incoming weights. **Pickering's** commentary adds a detailed neurobiological perspective to this discussion and he suggests various ways in which nodes might be marked as committed in a way that allows a threshold mechanism to act differentially on committed and uncommitted cells. In the target article, I suppose that uncommitted nodes should have an effective activation threshold of zero (sect. 4.4). This is to allow them to activate and learn under circumstances in which no committed node generates superthreshold activation in response to a given stimulus. **Pickering** is justified in thinking that I did not give enough prominence to this aspect of the model's functioning in the target article and I am grateful to him for proposing several biologically plausible means by which this functionality might be effected. The one most closely related to my suggested mechanism involves the idea that as the bottom-up weights increase to a previously uncommitted node, an inhibitory connection to a "threshold-controlling" node is also strengthened. (In fact, it would be helpful if the learning on the inhibitory connection lagged the bottom-up excitatory learning slightly, so as to prevent its interfering with this learning.) Uncommitted cells will have a weak connection to the inhibitory threshold node and so will be able to activate in the absence of active committed nodes. As **Pickering** notes, a very similar mechanism was proposed by Hasselmo and Wyble (1997), although I was unaware of this when I made my suggestion. (The fact that the learning scheme I suggested converges with one closely based on neurophysiological observation suggests that **Jorion** is overly pessimistic in his desire to avoid implementational issues – **Jorion,** para. 1–2.) The mechanism underlying the committed/uncommitted distinction might also be usefully co-opted in the modulation of learning rate on the bottom-up weights. If a strong connection to the threshold-controlling node were able to decrease the rate of bottom up learning, this would provide a mechanism for the fast-commitment, slow-recoding strategy described earlier in section R5.

Last, I must question **Ohlsson's** assumption (para. 4) that the sort of networks I propose (or indeed other networks) assume that every functional unit in the brain is connected to every other. In fact, in the networks I describe it is traditionally assumed that every node is connected to all nodes in adjacent layers – this is a very different assumption from **Ohlsson's.** Indeed, even this assumption is overly prescriptive. There only need be sufficient connections between adjacent layers to make it likely that novel patterns at lower layers can be well represented by previously uncommitted nodes. This does not require full connection though the lower limit for viable connectivity will likely vary depending upon the details of the task in hand.

R10. Efficiency

Devlin et al. make some points about the efficiency of distributed representations. I agree with much of what they say but feel their commentary is missing five crucial words, namely, "all other things being equal." To parody somewhat, all the events that occur at a rate of 10 per second over a lifespan of 100 years, can be uniquely labelled with a binary, distributed representation across 35 nodes. All the words in an average vocabulary would need 16 nodes to be so represented. No one is suggesting that these levels of compression are approached, let alone achieved, in the brain. So issues other than simply abstract efficiency of representation must be germane – the target article suggests what some of those issues might be. Put simply, evolution will only prefer a compact, efficient and, according to **Devlin et al.,** metabolically cheap code if that code is capable of doing what is demanded of it. No matter how theoretically elegant a code might be, if it is not capable of supporting arbitrary mappings, or does not afford good discrimination in noisy environments, or is not learnable (perhaps at speed), and so on, then it will not last long in the evolutionary mix. Efficiency can only be used to choose between models that work.

R11. A reply to Plaut & McClelland

R11.1. Stipulating representations or stipulating properties? It is suggested by **Plaut & McClelland** that I am overstipulative in postulating that representational units be assigned to meaningful entities. I think they misapprehend the motivations that drive a localist approach. At core, localists are interested in ensuring that the networks they develop are at least consistent with some general precepts of biological plausibility: that they use learning rules that use only locally available information; that they can perform unsupervised and supervised learning; that they can learn quickly if necessary; that their learning rule does not result in catastrophic levels of retroactive interference. It is these things that inform the modelling approach; the postulation

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

of localist representations is not itself stipulated, rather it is a proposed mechanism consistent with these constraints. Localist representation should not be ruled out a priori in favour of other less constrained approaches.

I also find **Plaut & McClelland's** views on stipulation somewhat inconsistent. In their continued adherence to a learning rule that, by its very nature, makes the emergence of localist representations unlikely, they are in effect stipulating that representational units should not be assigned to meaningful units (except when convenient, e.g., Plaut et al. 1996). **Plaut & McClelland** (para. 1) give a list of the principles underlying their approach to modelling but strangely omit distributed representation from this list. The principles that remain (viz., graded activations and weights, gradual, stochastic and interactive processing, and adaptation to task constraints) are perfectly compatible with localist modelling and, indeed, have been features of various localist models for many years.

I described above some of the motivations behind a localist modelling approach. What might the motivation be underlying a fully distributed approach that all but forbids localist representation, even in circumstances where this might be useful and appropriate? It cannot be a concern for biological plausibility because such considerations weigh against this approach. It cannot be study of the brain in vivo because what evidence there is seems more consistent with a localist strategy. It cannot be explanatory value because fully distributed networks are notoriously poor explanatory devices. Could it be that for **Plaut & McClelland** and others (e.g., Seidenberg 1993), distributed representation is simply axiomatic? If so, it is hard to know what evidence or argument would force a reappraisal.

R11.2. Over-dependence on task constraints. Advocated by **Plaut & McClelland** is the idea of representations that are discovered in response to task constraints. This seems to characterize all representations as intermediaries in the performance of some task, presumably, given the nature of their preferred learning rule, a supervised learning task. But this highlights a major limitation of their approach. Any intermediate (e.g., hidden layer) representation that emerges by slow gradient descent in the context of a given task is likely to be of use only in the performance of that task. In other words, the representations are optimized to the task. Were they to be applied in the context of some other task, they would either be ineffective or, if learning of the new task were permitted, vulnerable to catastrophic interference. Unsupervised learning, resulting in localist or featural representations that can be flexibly applied to novel tasks, is simply not a feature of **Plaut & McClelland's** preferred models. This makes it all the more surprising that later in their commentary they say of the input patterns to a localist network "nowhere in Page's article does he indicate how his localist approach could solve the problem of discovering such representations." If the structured input patterns to which they refer are featural (i.e., localist representations at a lower level), as they were for Plaut et al. (1996), then I go to some lengths to describe how such representations might arise. They, however, do not. How do they explain the existence of localist letter/phoneme nodes in the input/output layer of their own model? And if their answer involves, as I suspect it might, some idea of unsupervised localist learning then exactly what is it about localist learning that they are against?

R11.3. Flexibility in modelling. With regard to **Plaut & McClelland's** point concerning flexibility, they highlight my suggested account of age-of-acquisition (AOA) effects, noting that I make three modelling assumptions. They, like **Jorion,** seem to exaggerate the arbitrariness of these assumptions. The first assumption is that there is some enduring variability in the competitive capacities of the nodes in a given layer. The alternative assumption, either that there is no variability or that there is no enduring component in what variability exists, seems even more unlikely to be the case. The second assumption is that word acquisition is a competitive process – this is not an additional assumption at all, and is fundamental to the whole localist model that I propose. The third assumption is that word naming/recognition is also a process involving competition – I am hardly alone in this proposal, competition forming an important part of many, perhaps most, models of the naming/recognition process. So **Plaut & McClelland's** claim that it would have been much easier to have a model without these assumptions is very questionable, particularly when one seeks consistency with existing models. And what of their own preferred model? Any distributed model of the age-of-acquisition effect must solve the problem of catastrophic interference. As I note in the target article, this involves additional assumptions galore (and even these are insufficient). Applying **Plaut & McClelland's** test, would it not have been easier to have a model that did not make these assumptions and thus failed to avoid catastrophic interference and to show AOA effects?

R11.4. Complementary learning systems do not imply completely different learning rules. The complementary learning system hypothesis of McClelland et al. (1995) is defended by **Plaut & McClelland,** who suggest that the deficiencies of distributed learning techniques give a reason for having two complementary learning systems in the neocortex and the hippocampus. They do not seem to consider that there might be other reasons for such a distinction, one that boils down to there being a system for fast, one-shot episodic learning and another for more long-term averaging. It may well be that the hippocampus is specialized towards fast learning of individual events – it is well placed in the brain for the learning of such multimodal configurations and it appears to implement a sparse conjunctive coding appropriate for such a task. This does not mean that all other parts of the brain that use a localist learning rule have to specialize in exactly the same way. Indeed if they did, one could not really call it specialization. The neocortex doesn't do the same thing as the hippocampus – that really would be redundant – but that doesn't mean that it must use a completely different type of learning rule. If **Plaut & McClelland** really believed that parts of the brain with the same type of learning rule would necessarily have to perform the same task, then presumably they would deny different specializations across different areas of neocortex – an untenable position.

Plaut & McClelland state that rapid acquisition of arbitrary new information would *necessarily* be problematic for their conception of neocortex. Yet recent in vivo studies of lateral prefrontal cortex (see, e.g., Asaad et al. 1998) show that this region of neocortex has exactly this specialization. Whether it can be viably maintained that the rapid changes seen in prefrontal cortex are all due to fast learning in the hippocampus is, as far as I am aware, an open question.

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

R11.5. The plausibility of exemplar memory. Localist modelling is not necessarily exemplar-based (see sect. R5), but it can be. **Plaut & McClelland** ask what is an acceptable level of mismatch before a new node is formed. In the target article I make it clear that the level of vigilance is likely to be set by task requirements. If in reading a text we come across a word that is mildly misspelled, so that the intended word is clear, the task demands (i.e., comprehension) are not likely to be such that a high vigilance regime is effected so as to ensure learning of the "new" word. This said, it would be interesting to know how many people reading this response who also read **Plaut & McClelland's** commentary, can remember which word (misspelled) they gave as an example, and how that word was misspelled. If some readers can remember these details, as I anticipate, we might ask how a slow learning system, or indeed one that cannot store exemplars, could have performed such one-shot learning and retained the detail relevant to specifying the misspelling. One might claim that this is episodic memory (perhaps implemented with a localist system), but if this is the claim then the encounter with the misspelled word has indeed created a separate memory. It is inconsistent to admit the possibility of episodic memory but find exemplar memory (for perhaps a proportion, rather than the entirety, of what is encountered) "highly implausible."

R11.6. Biological plausibility. It is suggested by **Plaut & McClelland** that in using the back propagation (BP) algorithm as my example in section 7 of the target article, I appear to equate the distributed approach with the application of this algorithm. They seem to have missed the point made in the introduction to that section where I explicitly state that all the deficiencies of BP do not necessarily apply to other distributed learning rules. As an example **Plaut & McClelland** cite O'Reilly's (1996) GeneRec algorithm that avoids the use of nonlocal information in its implementation of gradient descent by running its network in two phases and by mirroring forward going connections with backward going connections of (ideally) the same weight. O'Reilly's suggested mechanism is convincing enough as a means of avoiding nonlocality in the learning rule. But even if we assume that the more complicated network is biologically plausible, the five other problems outlined in section 7 remain. Most important, O'Reilly's network has exactly the same problem with catastrophic interference as does the standard BP network. **Plaut & McClelland** do not respond to the argument in section 7.1 of the target article, in which I question the plausibility of supplying different output and training patterns for a given (old) input pattern presented to a neocortical learning network. I assume, therefore, that catastrophic interference remains a problem for such networks.

R11.7. Inappropriate generalization and the "what" and "when" problems. I agree that Plaut (1997) offered a means by which the semantic representations generated by nonwords were differentiable from the semantic representations generated by words. Making the word/nonword distinction involved the introduction of a new variable called the "stress" of a pattern; this is the degree to which the semantic pattern is far from being a binary pattern. It is not that a nonword gives no semantic pattern, just one that is less binary. Such a mechanism requires not just an extra mechanism for calculation of stress values but also that all genuine semantic patterns be binary, which places quite a severe constraint on featural models of semantic memory. If, alternatively, the semantics of nonwords are considered weak if they are a long way from being what they should be, or a long way from the nearest learned semantic vector, then we're back to the question of how such measurements are implemented in the absence of localist representations.

In response to section 7.4 of the target article ("Deciding what and when"), **Plaut & McClelland** make two points. First, there is some recent functional imaging data supporting an "energy-monitoring" role for the anterior cingulate. Second, no such explicit decisions are required. Taking the first point, the paper referred to (Bottvinick et al. 1999) does not give any more evidence of an energy-monitoring system (which Plaut 1997, considers biologically implausible, p. 791), than of any number of alternative interpretations. On the second point, **Plaut & McClelland** claim that "all that is needed is that the motor system be sufficiently damped that it initiates behaviour only when driven by strongly activated, stable internal representations." But the notion of a strongly activated internal representation controlling motor response is only practical, in the context of distributed representation, if the network can also act differently in response to different stable states. For example, a calculation of the length of a vector might tell you when to make a response (e.g., when the length exceeds a criterion, although this assumes that the learned patterns are longer than other nonlearned patterns) but it cannot tell you which response to make. The response itself is determined by the pattern across the whole vector, and the best way to signal the presence of a given pattern is to use a localist representation (cf. a matched filter – **Barlow & Gardner-Medwin**). Similar problems apply to the attractor networks favoured by **Farrell & Lewandowsky.** Of course, one could extensively train a network to respond to each of a number of distributed stimulus patterns with a given distributed response pattern, but for a fully distributed network this learning is necessarily a slow process. Yet if I ask you to perform the rather Pythonesque task of raising your left foot every time you hear the word "mattress," you can learn the task in one trial. Being intrigued as to how Kello et al. (in press) had implemented their "simple demonstration," I referred to the simulations presented therein. They model a Stroop response using localist output units, one for each colour – reaction time is simulated as the time at which the activation of a unit passes a threshold. Their reason for using such a scheme was "to make the measurement of response latency and duration straightforward." Quite.

R12. Mathematical considerations

R12.1. The power law of practice. I am grateful to **Heathcote & Brown** for drawing my attention to the ongoing debate regarding the power law of practice. Being a law-abiding citizen I had taken for granted that the ability of the generalized localist models to show power-law effects was a good thing. **Heathcote & Brown** (and, in more detail, Heathcote et al., in press) maintain that power-law fits have been artefactually improved by fitting to an average subject rather than individually to each subject. When individual fits are performed they find an exponential model is more appropriate. If they are correct, then the model's treatment

https://doi.org/10.1017/S0140525X00403353 Published online by Cambridge University Press

of practice effects will have to be modified. I am doubly grateful to these commentators for giving an account of the sort of modifications that might be necessary. They demonstrate the important point that an exponential improvement with practice can be obtained with a localist model similar to the one I suggest. Of course, I am not claiming that other learning rules cannot model either a power law or an exponential law of practice. **Plaut & McClelland** cite Cohen et al. (1990) who use a BP learning rule to achieve power-law behaviour, albeit by using what they call "simple and interpretable" localist output units, presumably to circumvent the "what and when" problems of section 7.4 of the target article.

The fact that **Heathcote & Brown** demonstrate a localist model capable of an exponential speed-up with practice means that my defence of localist models does not need to extend to a defence of the power law of practice. Nonetheless, it is worth making one point. For many tasks, instance theory suggests that performance is actually due to a mixture of algorithmic and memorial processes. Algorithmic processes dominate early in practice, before a sufficient stock of memorized responses has been accumulated. Logan (1988; 1990; 1992) has suggested that algorithmic and memorial processes race to respond. Rickard (1997) has suggested that both algorithmic and memorial processes speed-up as a power function of practice, but that a decision is made on each trial as to which type of process is used. A variety of other models are possible. In either case, the combined curve can differ in its form from the component curves from which it is made. Heathcote et al. (in press) are, of course, well aware of this. They present evidence from a number of studies in which participants were asked to specify how they dealt, or intended to deal, with a given trial, either by algorithm or by memory. In some studies this judgement was given retrospectively, after the relevant response had been given. In other cases participants were asked to make a rapid strategy assessment before responding. When Heathcote et al. (in press) fitted the algorithm and memory trials separately, they still found a considerable advantage for the exponential over the power-law practice function. There are a number of assumptions underlying their analysis: that participants can accurately gauge their strategy, which might be difficult if both strategies were operative and racing; or that the algorithmic path does not itself contain memory-based components, as might be imagined in the case of, say, multidigit arithmetic. Nonetheless, they make their case for an exponential function convincingly. It will be interesting to see how the proponents of a power-law respond to this legal challenge. In the context of my target article, Heathcote et al.'s (in press) work does not challenge a localist position (any more than it challenges a distributed one) though it would require modifications to my suggested model, perhaps of the sort they suggest.

Heathcote (personal communication) also pointed out an error in the target article where I used a mathematical argument to suggest that the Thurstone model is asymptotically identical to the Luce choice rule under uniform expansion of the set of instances. In fact, as Heathcote (personal communication and Colonius 1995) point out, the extreme value distribution (e.g., the distribution of the maximum of N values drawn from a given normal distribution) is degenerate as N approaches infinity. A more accurate statement of my point is that the Thurstone model becomes increasingly difficult to distinguish from the Luce choice model as the number of instances of each exemplar N increases.

R12.2. The Luce choice rule. An alternative "neural" formulation of the Luce choice rule in terms of independent race models is described by **Bundesen.** The analysis presented in Bundesen (1993) is extremely powerful and suggests a Luce implementation that is even simpler than, and just as localist as, the one described in the target article. Let us assume that nodes are Poisson generators and that, on presentation of a test stimulus in a classification or identification task, each fires with a rate proportional to a Gaussian function of the distance between its learned weight vector and the test vector, that is each has a Gaussian tuning curve with distance. Under these circumstances **Bundesen's** suggested mechanism for registering the first spike will choose a given identification/classification according to the Shepard-Luce-Nosofsky (SLF) formula given in Equation 8 with d representing a squared Euclidean distance. I have chosen to describe the model with a squared Euclidean distance here, because it implies a Gaussian tuning curve for the node: the Gaussian distribution is often used to model neural tuning curves. For an SLF model using a linear (rather than squared) distance measure, **Bundesen's** model would need to have nodes whose tuning curves approximated a negative exponential of the distance. This tuning function was used in Kruschke's (1992) ALCOVE model, but is rather less neurally plausible than the Gaussian, due to its being very sharply peaked at zero distance.

In his final paragraph, **Bundesen** notes a number of elaborations of the first-to-fire criterion. These models, which require integration over several spikes and which give approximations to the Luce choice rule, are more like the model that I present in the target article. It is clear that a variety of localist networks are able accurately and naturally to model the Luce choice rule.

R13. Conclusion

Having set out my case in the target article, and having read the commentaries that it elicited, I am still firmly of the opinion that localist models have much to recommend them in the field of psychological modelling. Connectionist modelling received a boost in popularity with the appearance (or reappearance) of the BP learning algorithm, but this has tended to lead to the overshadowing of older, localist approaches. With the growing prominence of other investigative techniques applied to the understanding of brain function, a localist modelling perspective will, I believe, prove increasingly valuable.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Aggleton, J. P. & Brown, M. W. (1999) Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences* 22(3):425–89. [aMP]

Ajjanagadde, V. & Shastri, L. (1991) Rules and variables in neural nets. *Neural Computation* 3:121–34. [LS]

Amit, D. J. (1989) *Modeling brain function: The world of attractor neural networks.* Cambridge University Press. [aMP]

(1995) The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences* 18:617–57. [aMP]
Amit, D. J., Fusi, S. & Yakovlev, V. (1997) Paradigmatic working memory (attractor) cell in its cortex. *Neural Computation* 9:1071–92. [aMP]
Anderson, J. A., Silverstein, J. W., Ritz, S. A. & Jones, R. S. (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* 84:413–51. [DCP]
Anderson, J. R. & Lebiere, C. (1998) *The atomic components of thought.* Erlbaum. [SO]
Asaad, W. F., Rainer, G. & Miller, E. K. (1998) Neural activity in the primate prefrontal cortex during associative learning. *Neuron* 21:1399–407. [rMP]
Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U. & Waldron, E. M. (1998) A neuropsychological theory of multiple systems in category learning. *Psychological Review* 105:442–81. [ADP]
Baddeley, A. D. (1968) How does acoustic similarity influence short-term memory? *Quarterly Journal of Experimental Psychology* 20:249–63. [rMP]
Baddeley, R. J., Abbot, I. F., Booth, M. C. A., Sengpiel, F., Freeman, T., Wakeman, E. A. & Rolls, E. T. (1997) Responses to neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London Series B* 264:1775–83. [MPY]
Bailey, D. (1997) When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs. Ph. D. dissertation, Computer Science Division, University of California, Berkeley. [LS]
Barlow, H. (1972) Single units and sensation: A neuron doctrine for perceptual psychology. *Perception* 1:371–94. [aMP]
(1995) The neuron doctrine in perception. In: *The cognitive neurosciences,* ed. M. S. Gazzaniga. MIT Press. [aMP]
Barlow, H. B. & Tripathy, S. P. (1997) Correspondence noise and signal pooling as factors determining the detectability of coherent visual motion. *Journal of Neuroscience* 17:7954–66. [HB]
Bechtel, W. & Abrahamsen, A. (1991) *Connectionism and the mind.* Blackwell. [ISNB]
Berkeley, I. (1999) What the #$*%! is a subsymbol? *Minds and Machines.* (forthcoming). http://www.ucs.usl.edu/~isb9112/dept/phi1341/subsymbol/subsymbol.html [ISNB]
Berkeley, I., Dawson, M., Medler, D., Schpflocher, D. & Hornsby, L. (1995) Density plots of hidden unit activations reveal interpretable bands. *Connection Science* 7(2):167–86. [ISNB, rMP]
Bishop, C. M. (1995) *Neural networks for pattern recognition.* Oxford University Press. [rMP]
Bliss, T. V. P. & G. L. Collingridge (1993) A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* 361:31–39. [LS]
Booth, M. C. A. & E. T. Rolls (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex* 8:510–23. [aMP]
Botvinick, M., Nystrom, L., Fissell, K., Carter, C. S. & Cohen, J. D. (1999) Conflict monitoring versus selection-for-action in anterior cingular cortex. *Nature* 402:149. [DCP, rMP]
Bower, G. H. (1996) Reactivating a reactivation theory of implicit memory. *Consciousness and Cognition* 5:27–72. [aMP]
Bowers, J. S. & Michita, Y. (1998) An investigation into the structure and acquisition of orthographic knowledge: Evidence from cross-script Kanji-Hiragana priming. *Psychonomic Bulletin and Review* 5:259–64. [JSB]
Bradski, G. & Grossberg, S. (1995) Fast-learning VIEWNET architecture for recognizing three-dimensional objects from multiple two-dimensional views. *Neural Networks* 8:1053–80. [SG]
Bransford, J. D. & Franks, J. J. (1971) The abstraction of linguistic ideas. *Cognitive Psychology* 2:331–50. [JEI]
Britten, K. H. & Newsome, W. T. (1999) Tuning bandwidths for near-threshold stimuli in area MT. *Journal of Neurophysiology* 80:762–70. [SRS]
Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience* 12(12):4745–65. [HB, aMP]
Brown, H. L., Sharma, N. K. & Kirsner, K. (1984) The role of script and phonology in lexical representation. *Quarterly Journal of Experimental Psychology* 36A:491–505. [JSB]
Brown, S. & Heathcote, A. (in preparation) The effects of averaging on the determination of curve form. [AH]
Bruner, J. S. (1973) Going beyond the information given. In: *Beyond the information given,* ed. J. S. Bruner & J. M. Anglin. Norton. [JEI]
Bruner, J. S. & Minturn, A. L. (1955) Perceptual identification and perceptual organization. *Journal of General Psychology* 53:21–28. [JEI]
Buckley, M. J. & Gaffan, D. (1998) Perirhinal cortex ablation impairs configural learning and paired-associate learning equally. *Neuropsychologia* 36:535–46. [aMP]
Bundesen, C. (1987) Visual attention: Race models for selection from multielement displays. *Psychological Research* 49:113–21. [CB]
(1990) A theory of visual attention. *Psychological Review* 97:523–47. [CB]
(1991) Towards a neural network implementation of TVA. Paper presented at the First Meeting of the HFSP Research Group on Brain Mechanisms of Visual Selection, School of Psychology, University of Birmingham, UK. [CB]
(1993) The relationship between independent race models and Luce's choice axiom. *Journal of Mathematical Psychology* 37:446–71. [CB, arMP]
Bundesen, C. & Harms, L. (1999) Single-letter recognition as a function of exposure duration. *Psychological Research/Psychologische Forschung* 62:275–79. [CB]
Bundesen, C., Shibuya, H. & Larsen, A. (1985) Visual selection from multielement displays: A model for partial report. In: *Attention and performance XI,* ed. M. I. Posner & O. S. M. Marin. Erlbaum. [CB]
Buracas, G. F., Zador, A., DeWeese, M. & Albright, T. (1998) Efficient discrimination of temporal patterns by motion-sensitive neurons in the primate visual cortex. *Neuron* 20:959–69. [MPY]
Burgess, N. & Hitch, G. J. (1992) Towards a network model of the articulatory loop. *Journal of Memory and Language* 31:429–60. [aMP]
(1999) Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review* 106:551–81. [CPB, JSB, arMP]
Burton, A. M. (1994) Learning new faces in an interactive activation and competition model. *Visual Cognition* 1(2/3):313–48. [aMP]
Burton, A. M., Bruce, V. & Hancock, P. J. B. (1999) From pixels to people: A model of familiar face recognition. *Cognitive Science* 23:1–31. [AMB]
Burton, A. M., Bruce, V. & Johnson, R. A. (1990) Understanding face recognition with an interactive activation model. *British Journal of Psychology* 81:361–80. [aMP]
Burton, A. M. & Young, A. W. (1999) Simulation and explanation: Some harmony and some discord. *Cognitive Neuropychology* 16:73–79. [AMB]
Carpenter, G. A. (1996) Distributed activation, search, and learning by ART and ARTMAP neural networks. *Proceedings of the International Conference on Neural Networks (ICNN '96): Plenary, Panel and Special Sessions,* 244–49. IEEE Press. [GAC]
(1997) Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks* 10:1473–94. [JSB, GAC, rMP]
Carpenter, G. A. & Grossberg, S. (1987a) ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics* 26:4919–30. [PCMM, ADP, aMP]
(1987b) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing* 37:54–115. [JSB, GAC, aMP, DCP]
(1991) *Pattern recognition by self-organizing neural networks.* MIT Press. [SG]
(1993) Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neuroscience* 16:131–37. [GAC]
Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H. & Rosen, D. B. (1992) Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* 3:698–713. [GAC]
Carpenter, G. A., Grossberg, S. & Reynolds, J. H. (1991) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks* 4:565–88. [GAC]
Carpenter, G. A., Grossberg, S. & Rosen, D. B. (1991) ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks* 4(4):493–504. [arMP]
Carpenter, G. A., Milenova, B. L. & Noeske, B. W. (1998) Distributed ARTMAP: A neural network for fast distributed supervised learning. *Neural Networks* 11:793–813. [GAC, rMP]
Carpenter, R. H. S. & Williams, M. L. L. (1995) Neural computation of log likelihood in the control of saccadic eye movements. *Nature* 377:59–62. [rMP]
Cheng, P. W. & Pachella, R. G. (1984) A psychophysical approach to dimensional separability. *Cognitive Psychology* 16:279–304. [JEH]
Cleeremans, A. & McClelland, J. L. (1991) Learning the structure of event sequences. *Journal of Experimental Psychology: General* 120:235–53. [AK, rMP, IV]
Cleeremans, A., Servan-Schreiber, D. & McClelland, J. M. (1989) Finite state automata and simple recurrent networks. *Neural Computation* 1:372–81. [IV]
Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990) On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review* 97(3):332–61. [rMP, DCP]
Cohen, M. A. & Grossberg, S. (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics* 13:815–26. [aMP]
(1986) Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short-term memory. *Human Neurobiology* 5:1–22. [SG]
(1987) Masking fields: A massively parallel neural architecture for learning,

recognizing, and predicting multiple groupings of patterned data. *Applied Optics* 26:1866–91. [SG, rMP]
Cohen, M. A., Grossberg, S. & Wyse, L. (1995) A spectral network model of pitch perception. *Journal of the Acoustical Society of America* 98:862–79. [rMP]
Colonius, H. (1995) The instance theory of automaticity: Why the Weibull? *Psychological Review* 102:744–50. [rMP]
Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993) Models of reading aloud: Dual route and parallel-distributed-processing approaches. *Psychological Review* 100(4):580–608. [aMP]
Cook, N. D. (1995a) Artefact or network evolution? *Nature* 374:313. [NDC]
(1995b) Correlations between input and output units in neural networks. *Cognitive Science* 19:563–74. [NDC]
(1999) Simulating consciousness in a bilateral neural network: "Nuclear" and "fringe" awareness. *Consciousness and Cognition* 8:62–93. [NDC]
Cook, N. D., Früh, H. & Landis, T. (1995) The cerebral hemispheres and neural network simulations: Design considerations. *Journal of Experimental Psychology: Human Perception and Performance* 21:410–22. [NDC]
Dawson, M. (1998) *Understanding cognitive science.* Blackwell. [ISNB]
DeAngelis, G. C., Cummins, B. G. & Newsome, W. T. (1998) Cortical area MT and the perception of stereoscopic depth. *Nature* 394:677–80. [HB]
Dell, G. S. (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3):283–321. [arMP, DCP]
(1988) The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language* 27:124–42. [arMP]
Diederich, J. (1989) Instruction and high-level learning in connectionist networks. *Connection Science* 1:161–80. [LS]
Dienes, Z. (1992) Connectionist and memory array models of artificial grammar learning. *Cognitive Science* 16:41–79. [AK]
Dienes, Z., Altmann, G. T. M. & Gao, S. (1999) Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science* 23:53–82. [AK]
Dopkins, S. & Gleason, T. (1997) Comparing exemplar and prototype models of categorization. *Canadian Journal of Experimental Psychology* 511(3):212–30. [rMP]
Elman, J. L. (1990) Finding structure in time. *Cognitive Science* 14(2):179–211. [JSB, AK, IV, rMP]
(1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195–225. [DCP]
Estes, W. K. (1972) An associative basis for coding and organization in memory. In: *Coding processes in human memory,* ed. A. W. Melton & E. Martin. V. H. Winston. [aMP]
(1986) Array models for category learning. *Cognitive Psychology* 18:500–49. [aMP]
Fahlman, S. E. & Lebiere, C. (1990) The Cascade-correlation learning architecture. In: *Advances in neural information processing systems 2,* ed. D. S. Touretsky. Morgan Kaufmann. [TRS]
Farah, M. J. (1994a) Interactions on the interactive brain. *Behavioral and Brain Sciences* 17(1):90–104. [arMP]
(1994b) Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences* 17:43–104. [AMB]
Farah, M. J., Meyer, M. & McMullen, P. (1996) The living/non-living distinction is not an artefact: Giving an a priori implausible hypothesis a strong test. *Cognitive Neuropsychology* 13:137–54. [RMF]
Farah, M. J., O'Reilly, R. C. & Vecera, S. P. (1993) Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review* 100:571–88. [AMB]
Feldman, J. A. (1982) Dynamic connections in neural networks. *Biological Cybernetics* 46:27–39. [LS]
(1988) Connectionist representation of concepts. In: *Connectionist models and their implications,* ed. D. Waltz & J. A. Feldman. Ablex. [aMP]
(1989) Neural representation of conceptual knowledge. In: *Neural connections, mental computation,* ed. L. Nadel, L. A. Cooper, P. Culicover & R. M. Harnish. MIT Press. [LS]
Feldman, L. B. & Moskovljevic, J. (1987) Repetition priming is not purely episodic in origin. *Journal of Experimental Psychology: Learning, Memory and Cognition* 13:573–81. [JSB]
Felleman, D. J. & Kaas, J. H. (1984) Receptive field properties of neurons in middle temporal visual area (MT) of owl monkeys. *Journal of Neurophysiology* 52:488–513. [HB]
Fodor, J. & Pylyshyn, Z. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71. [aMP]
Foldiak, P. (1991) Models of sensory coding. Technical Report CUED/F-INFENG/TR 91, Physiological Laboratory, University of Cambridge. [aMP]
Forster, K. I. (1994) Computational modeling and elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 20(6):192–310. [aMP]
Fotheringhame, D. K. & Young, M. P. (1996) Neural coding and sensory representations: Theoretical approaches and empirical evidence. In: *Cognitive neuroscience,* ed. M. D. Rugg; *Studies of Cognition* series, ed. G. Humphreys. Psychology Press. [MPY]
French, R. M. (1991) Using semi-distributed representations to overcome catastrophic interference in connectionist networks. In: *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society,* 173–78. Erlbaum. [aMP]
(1992) Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science* 4:365–77. [aMP]
(1994) Dynamically constraining connectionist networks to produce orthogonal, distributed representations to reduce catastrophic interference. In: *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society,* 335–40. [aMP]
(1997a) Pseudo-recurrent connectionist networks and the problem of sequential learning. Paper presented at the Fourth Neural Computation and Psychology Workshop, London, April 1997. [aMP]
(1997b) Pseudo-recurrent connectionist networks: An approach to the "sensitivity-stability" dilemma. *Connection Science* 9(4):353–79. [RMF]
French, R. M. & Mareschal, D. (1998) Could category-specific semantic deficits reflect differences in the distributions of features within a unified semantic memory? In: *Proceedings of the 20th Annual Conference of the Cognitive Science Society, 1998, New Jersey,* 374–79. LEA. [RMF]
Fried, L. S. & Holyoak, K. J. (1984) Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* 10:234–57. [aMP]
Funnell, E. & Sheridan, J. (1992) Categories of knowledge? Unfamiliar aspects of living and non-living things. *Cognitive Neuropsychology* 9:135–53. [RMF]
Gallant, J. L., Connor, C. E. & Van Essen, D. C. (1998) Neural activity in areas V1, V2, and V4 during free viewing of natural scenes compared to controlled viewing. *NeuroReport* 9:85–90. [MPY]
Gardner-Medwin, A. R. & Barlow, H. B. (submitted) The limits of counting accuracy in distributed neural representations. *Neural Computation.* [HB]
Garner, W. R. (1974) *The processing of information and structure.* Erlbaum. [JEH]
Georgopoulos, A. P., Kettner, R. E. & Schwartz, A. B. (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. *The Journal of Neuroscience* 8:298–. [aMP]
Georgopoulos, G., Masato, T. & Lukashin, A. (1993) Cognitive neurophysiology of the motor cortex. *Science* 260:47–51. [RMF]
Gibson, J. J. & Gibson, E. J. (1955) Perceptual learning: Differentiation or enrichment? *Psychological Review* 62:32–41. [JEI]
Giles, C. L., Miller, C. B., Chen, D., Chen, H. H., Sun, G. Z. & Lee, Y. C. (1992) Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation* 4:393–405. [IV]
Gluck, M. A. & Myers, C. E. (1997) Extending models of hippocampal function in animal conditioning to human amnesia. *Memory* 5(1/2):179–212. [aMP]
Golden, R. M. (1986) A developmental neural model of visual word perception. *Cognitive Science* 10:241–76. [RMG]
Graham, K. S. & Hodges, J. R. (1997) Differentiating the roles of the hippocampal complex and the neocortex in long-term memory storage: Evidence from the study of semantic dementia and Alzheimer's disease. *Neuropsychology* 11(1):77–89. [aMP]
Grainger, J. & Jacobs, A. M. (1996) Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review* 103(3):518–65. [aMP]
(1998) On localist connectionism and psychological science. In: *Localist connectionist approaches to human cognition,* ed. J. Grainger & A. M. Jacobs. Erlbaum. [aMP]
Green, C. D. (1998) Are connectionist models theories of cognition? *Psycoloquy* 9. ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/psyc.98.9.04.connec tionist-explanation.1.green [aMP]
Groh, J. M., Born, R. T. & Newsome, W. T. (1997) How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *Journal of Neuroscience* 17:4312–30. [rMP]
Gross, C. G. (1992) Representation of visual stimuli. *Philosophical Transactions of the Royal Society London, Series B* 335:3–10. [aMP]
Grossberg, S. (1972) Neural expectation: Cerebellar and retinal analogy of cells fired by learnable or unlearned pattern classes. *Kybernetik* 10:49–57. [aMP]
(1976) Adaptive pattern classification and universal recoding. I: Parallel development and coding of neural feature detectors. *Biological Cybernetics* 23:121–34. [CB, DCP]
(1978a) A theory of human memory: Self-organization and performance of sensory-motor codes, maps and plans. In: *Progress in theoretical biology, vol. 5,* ed. R. Rosen & F. Snell. Academic Press. Reprinted in: Grossberg, S. (1982) *Studies of mind and brain.* Kluwer/Reidel. [SG]
(1978b) Behavioral contrast in short-term memory: Serial binary memory models or parallel continuous memory models. *Journal of Mathematical Psychology* 17:199–219. [rMP]

(1980) How does a brain build a cognitive code? *Psychological Review* 87:1–51. [CB, PCMM]
(1982) *Studies of mind and brain.* Reidel. [aMP]
(1984) Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory* 7:263–83. [SG]
(1986) The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In: *Pattern recognition by humans and machines, vol. 1: Speech perception,* ed. E. C. Schwab & H. C. Nusbaum. Academic Press. [SG]
(1987) Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11:23–63. [SG, aMP]
(1997) Neural models of development and learning. *Behavioral and Brain Sciences* 20:566. [aMP]
(1999) How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision* 12:163–86. [SG]
Grossberg, S. & Merill, J. W. L. (1996) The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience* 8:257–77. [SG]
Grossberg, S. & Myers, C. W. (1999) The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review.* (in press). [SG]
Grossberg, S. & Raizada, R. D. S. (1999) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research.* (in press). [SG]
Grossberg, S. & Stone, G. (1986) Neural dynamics of attention switching and temporal-order information in short-term memory. *Memory and Cognition* 14(6):451–68. [JSB]
Hall, G. (1991) *Perceptual and associative learning.* Oxford University Press. [JEI]
Hanes, D. P. & Shall, J. D. (1996) Neural control of voluntary movement initiation. *Science* 274:427–30. [rMP]
Harnad, S., ed. (1987) *Categorical perception: The groundwork of cognition.* Cambridge University Press. [aMP]
Harris, C. S. (1980) Insight or out of sight: Two examples of perceptual plasticity in the human adult. In: *Visual coding and adaptability,* ed. C. S. Harris. Erlbaum. [aMP]
Hartley, T. & Houghton, G. (1996) A linguistically constrained model of short-term memory for nonwords. *Journal of Memory and Language* 35:1–31. [aMP]
Hasselmo, M. E. & Schnell, E. (1994) Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modelling and brain slice physiology. *The Journal of Neuroscience* 14:3898–914. [ADP]
Hasselmo, M. E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampus region CA3. *The Journal of Neuroscience* 15:5249–62. [ADP]
Hasselmo, M. E. & Wyble, B. (1997) Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research* 89:1–34. [ADP, rMP]
Hayes-Roth, B. & Hayes-Roth, F. (1977) Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior* 16:321–38. [TRS]
Heath, R. A. (1992) A general nonstationary diffusion model for two-choice decision-making. *Mathematical Social Sciences* 23:283–309. [AH]
Heathcote, A. (1998) Neuromorphic models of response time. *Australian Journal of Psychology* 50:157–64. [AH]
Heathcote, A. & Brown, S. (in preparation) Neuromorphic models of the law of practice. [AH]
Heathcote, A., Brown, S. & Mewhort, D. J. K. (in press) The Power Law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review.* [AH, rMP]
Hecht-Nielsen, R. (1987) Counterpropagation networks. *Applied Optics* 26:4979–84. [aMP]
Henson, R. N. A., Norris, D. G., Page, M. P. A. & Baddeley, A. D. (1996) Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology* 49A:80–115. [rMP]
Hetherington, P. A. & Shapiro, M. L. (1993) Simulating Hebb cell assemblies: The necessity for partitioned dendritic trees and a post-not-pre LTD rule. *Network* 4:135–53. [JEI]
Hinton, G. E. (1986) Learning distributed representations of concepts. In: *Proceedings of the Eight Annual Conference of the Cognitive Science Society,* 1–12. Erlbaum. [SF, DCP]
Hinton, G. E., Dayan, P., Frey, B. & Neal, R. (1995) The "wake-sleep" algorithm for unsupervised neural networks. *Science* 268:1158–61. [JTD]
Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986) Distributed representations. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations,* ed. D. E. Rumelhart, J. L. McClelland and the PDP Research Group. MIT Press. [aMP]
Hinton, G. E. & Sejnowski, T. J. (1986) Learning and relearning in Boltzman machines. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations,* ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [CM]
Hintzman, D. L. (1986) Schema abstraction in a multiple-trace memory model. *Psychological Review* 93(4):411–28. [aMP]
Hoguchi, S.-I. & Miyashita, Y. (1996) Formation of mnemonic neural responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. *Proceedings of the National Academy of Sciences USA* 93:739–43. [aMP]
Holyoak, K. J. & Hummel, J. E. (in press) The proper treatment of symbols in a connectionist architecture. In: *Cognitive dynamics: Conceptual change in humans and machines,* ed. E. Dietrich & A. Markman. Erlbaum. [JEH]
Holyoak, K. J. & Thagard, P. (1989) Analogical mapping by constraint satisfaction. *Cognitive Science* 13:295–355. [TRS]
Hopfield, J. (1982) Neuronal networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554–58. [aMP]
(1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA* 81:3058–92. [CM, aMP]
Houghton, G. (1990) The problem of serial order: A neural network memory of sequence learning and recall. In: *Current research in natural language generation,* ed. R. Dale, C. Mellish & M. Zock. Academic Press. [aMP]
Howard, D. (1995) Lexical anomia: Or the case of the missing lexical entries. *Quarterly Journal of Experimental Psychology* 48A(4):999–1023. [aMP]
Hummel, J. E. & Biederman, I. (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* 99(3):480–517. [aMP]
Hummel, J. E. & Holyoak, K. J. (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104(3):427–66. [JSB, JEH, aMP]
Jacobs, A. M. & Grainger, J. (1994) Models of visual word recognition–sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance* 20(6):1311–34. [aMP]
Jacoby, L. L. & Dallas, M. (1981) On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General* 3:306–40. [RHP]
Jamieson, D. G. & Petrusik, W. M. (1977) Preference time to choose. *Organizational Behavior and Human Performance* 19:56–57. [AH]
Johnson, S. J. & Carey, S. (1998) Knowledge, enrichment and conceptual change in folk biology: Evidence from Williams syndrome. *Cognitive Psychology* 38:156–200. [DCP]
Johnstone, T. & Shanks, D. R. (1999) Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997). *Journal of Experimental Psychology: Learning, Memory and Cognition* 25:524–31. [rMP]
Jones, F. W. & McLaren, I. P. L. (1999) Rules and associations. In: *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society,* ed. M. Hahn & S. Stoness. Erlbaum. [SO]
Jordan, M. I. (1990) A non-empiricist perspective on learning in layered networks. *Behavioral and Brain Sciences* 13:497–98. [CL-S, rMP]
Jorion, P. (1989) An alternative neural network representation for conceptual knowledge. Paper presented at the British TELECOM, CONNEX Conference, Martlesham Heath, January 1990. http://cogprints.soton.ac.uk/abs/comp/199806036 [PJMJ]
(1990) *Principes des systèmes intelligents.* Collection "Sciences cognitives." Masson. Summary in English: http://cogprints.soton.ac.uk/abs/comp/199806039 [PJMJ]
(1996) La linguistique d'Aristote. In: *Penser l'esprit: Des sciences de la cognition à une philosophie cognitive,* ed. V. Rialle & D. Fisette. Presses Universitaires de Grenoble. http://cogprints.soton.ac.uk/abs/phil/199807012 [PJMJ]
Kanerva, P. (1988) *Sparse distributed memory.* MIT Press. [aMP, DCP]
Karmiloff-Smith, A. (1992) *Beyond modularity: A developmental perspective on cognitive science.* MIT Press. [TRS]
Kastner, S., Weerd, P. D., Desimone, R. & Ungerleider, L. G. (1998) Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science* 282:108–11. [aMP]
Keeler, J. D. (1988) Comparison between Kanerva's SDM and Hopfield-type neural networks. *Cognitive Science* 12:299–329. [aMP]
Keil, F. C. (1979) *Semantic and conceptual development: An ontological perspective.* Harvard University Press. [DCP]
Kello, C. T., Plaut, D. C. & MacWhinney, B. (in press) The task-dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference on speech production. *Journal of Experimental Psychology: General.* [DCP, rMP]
Kelso, S. R., Ganong, A. H. & Brown, T. H. (1986) Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences USA* 83:5326–30. [SF]

Kinder, A. (2000) The knowledge acquired during artificial grammar learning: Testing the predictions of two connectionist networks. *Psychological Research* 63. [AK]

Knowlton, B. J. & Squire, L. R. (1996) Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:169–81. [AK]

Kohonen, T. (1984) *Self-organization and associative memory.* Springer-Verlag. [aMP]

(1995) *Self-organizing maps.* Springer. [RHP]

Kohonen, T., Oja, E. & Lehtiö, P. (1981) Storage and processing of information in distributed associative memory systems. In: *Parallel models of associative memory,* ed. G. Hinton & J. A. Anderson. Erlbaum. [MJ]

Kornbrot, D. E. (1978) Theoretical and empirical comparison of Luce's choice model and logistic Thurstone model of categorical judgment. *Perception and Psychophysics* 37(1):89–91. [aMP]

Kruschke, J. K. (1992) ALCOVE: An examplar-based connectionist model of category learning. *Psychological Review* 99(1):22–44. [SF, arMP, DCP]

Kunda, Z. & Thagard, P. (1996) Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint satisfaction theory. *Psychological Review* 103:284–308. [TRS]

Lacouture, Y. & Marley, A. A. J. (1991) A connectionist model of choice and reaction time in absolute identification. *Connection Science* 3:401–33. [CL-S]

Lamberts, K. (1996) Exemplar models and prototype effects in similarity-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:1503–507.

Lambon Ralph, M. A. (1998) Distributed versus localist representations: Evidence from a study of item consistency in a case of classical anomia. *Brain and Language* 64:339–60. [aMP]

Lambon-Ralph, M. A., Howard, D., Nightingale, G. & Ellis, A. W. (1998) Are living and non-living category-specific deficits causally linked to impaired perceptual or associative knowledge? Evidence from a category-specific double dissociation. *Neurocase* 4:311–38. [rMP]

Laughlin, S. B., de Ruyter van Stevenick, R. R. & Anderson, J. C. (1998) The metabolic cost of neural information. *Nature Neuroscience* 1(1):36–41. [JTD]

Lee, C. L. & Estes, W. K. (1981) Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Human Learning and Memory* 7:149–69. [aMP]

Leth-Steensen, C. & Marley, A. A. J. (2000) A model of response time effects in symbolic comparison. *Psychological Review* 104:62–100. [CL-S]

Levelt, W. J. M. (1989) *Speaking: From intention to articulation.* MIT Press. [arMP]

Lewandowsky, S. (1991) Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In: *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock,* ed. W. E. Hockley & S. Lewandowsky. Erlbaum. [aMP]

(1994) On the relation between catastrophic interference and generalization in connectionist networks. *Journal of Biological Systems* 2:307–33. [SF, rMP]

(1999) Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology* 34:434–46. [SF]

Lewandowsky, S. & Farrell, S. (in press) A redintegration account of the effects of speech rate, lexicality, and word frequency in immediate serial recall. *Psychological Research.* [SF]

Lewandowsky, S. & Li, S.-C. (1994) Memory for serial order revisited. *Psychological Review* 101:539–43. [SF]

Lewis, J. E. & Kristan, W. B., Jr. (1998) A neuronal network for computing population vectors in the leech. *Nature* 391:77–79. [aMP]

Linden, D. J. (1994) Long-term synaptic depression in the mammalian brain. *Neuron* 12:457–72. [LS]

Logan, G. D. (1988) Towards an instance theory of automatization. *Psychological Review* 95:492–527. [arMP, ADP]

(1990) Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology* 22:1–35. [arMP]

(1992) Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition* 18(5):883–914. [arMP, ADP]

(1996) The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review* 103:603–49. [CB]

Logothesis, N. K., Pauls, J. & Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys. *Current Biology* 5:552–63. [CM]

López, F. J., Shanks, D. R., Almaraz, J. & Fernández, P. (1998) Effects of trial order on contingency judgements: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology: Learning, Memory and Cognition* 24:672–94. [arMP]

Luce, R. D. (1959) *Individual choice behaviour: A theoretical analysis.* Wiley. [CB, aMP]

Luce, R. D. & Green, D. M. (1972) A neural timing theory for response times and the psychophysics of intensity. *Psychological Review* 79:14–57. [CB]

Mackintosh, N. J., Kaye, H. & Bennett, C. H. (1991) Perceptual learning in flavour aversion conditioning. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology* 43B:297–322. [JEI]

Marcus, G. F. (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 37(3):243–82. [JSB]

Marley, A. A. J. & Colonius, H. (1992) The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology* 36:1–20. [CB]

Marr, D. (1970) A theory for cerebral neocortex. *Proceedings of the Royal Society of London B* 176:161–234. [aMP, DCP]

(1971) Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B* 262:23–81. [aMP]

Marshall, J. A. (1990) A self-organizing scale-sensitive neural network. In: *Proceedings of the International Joint Conference on Neural Networks,* vol. 3, June 1990. IEEE Publication. [aMP]

Martindale, C. (1981) *Cognition and consciousness.* Dorsey. [CM]

(1991) *Cognitive psychology: A neural-network approach.* Brooks/Cole. [CM]

(1995) Creativity and connectionism. In: *The creative cognition approach,* ed. S. Smith, T. Ward & R. Finke. MIT Press. [CM]

(in press) Cognition and aesthetics. In: *The visual arts: Education, psychology, and aesthetics,* ed. J. P. Frois. Gulbenkian. [CM]

Massaro, D. W. (1987) Categorical partition: A fuzzy logical model of categorization behavior. In: *Categorical perception: The groundwork of cognition,* ed. S. Harnad. Cambridge University Press. [aMP]

(1988) Some criticisms of connectionist models of human performance. *Journal of Memory and Language* 27:213–34. [aMP]

Maunsell, J. H. R. & Van Essen, D. C. (1983a) Functional properties of neurons in middle temporal visual area of the macaque. I. Selectivity for stimulus direction, speed and orientation. *Journal of Neurophysiology* 49:1127–47. [HB]

(1983b) Functional properties of neurons in the middle temporal area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity. *Journal of Neurophysiology* 49:1148–67. [HB]

McClelland, J. L. (1979) On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* 86(4):287–330. [aMP]

(1981) Retrieving general and specific information from stored knowledge of specifics. In: *Proceedings of the Third Annual Meeting of the Cognitive Science Society,* August 19–21. [aMP]

(1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* 23:1–44. [aMP]

(1993a) Toward a theory of information processing in graded, random and interactive networks. In: *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience,* ed. D. E. Meyer & S. Kornblum. MIT Press. [aMP]

(1993b) The GRAIN model: A framework for modeling the dynamics of information processing. In: *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience,* ed. D. E. Meyer & S. Kornblum. Erlbaum. [DCP]

(1994) The interaction of nature and nurture in development: A parallel distributed processing perspective. In: *International perspectives on psychological science, vol. 1: Leading themes,* ed. P. Bertelson, P. Eelen & G. d'Ydewalle. Erlbaum. [DCP]

McClelland, J. L. & Elman, J. (1986) The TRACE model of speech perception. *Cognitive Psychology* 18:11–86. [aMP]

McClelland, J. L. & Goddard, N. H. (1996) Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* 6:654–65. [DCP]

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995) Why are there complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102(3):419–57. [JTD, SF, RMF, arMP, DCP]

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. [RMG, SG, arMP, RHP, DCP]

(1985) Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* 114:159–88. [CM, DCP]

(1986) A distributed model of human learning and memory. In: *Parallel distributed processing, vol. 2,* ed. J. L. McClelland & D. E. Rumelhart. MIT Press. [TRS]

(1988) *Explorations in parallel distributed processing.* Bradford Books. [AMB]

McCloskey, M. (1991) Networks and theories. *Psychological Science* 2(6):387–95. [aMP]

McCloskey, M. & Cohen, N. J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In: *The psychology of learning and motivation, vol. 24,* ed. G. H. Bower. Academic Press. [aMP, DCP]
McCollough, C. (1965) Color adaptation of edge detectors in the human visual system. *Science* 149:1115–16. [aMP]
McDermott, D. (1981) Artificial Intelligence meets natural stupidity. In: *Mind design,* ed. J. Hagueland. MIT Press. [AMB]
McGill, W. J. (1963) Stochastic latency mechanisms. In: *Handbook of mathematical psychology, vol. 1,* ed. R. D. Luce, R. R. Bush & E. Galanter. Wiley. [CB]
McLaren, I. P. L. (1993a) APECS: A solution to the sequential learning problem. In: *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society,* 717–22. Erlbaum. [aMP]
(1993b) Catastrophic interference is eliminated in pretrained networks. In: *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society,* 723–28. Erlbaum. [aMP]
McRae, K., de Sa, V. R. & Seidenberg, M. S. (1997) On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General* 126:99–130. [aMP]
Medin, D. L. & Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review* 85:207–38. [aMP]
Miikkulainen, R. (1996) Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science* 20:47–73. [SF, rMP]
Miller, G. A. (1956) The magic number seven plus or minus two. *Psychological Review* 63:81–97. [SG]
Minsky, M. & Papert, S. (1969) *Perceptrons.* MIT Press. [aMP]
Mitchell, T. M. (1997) *Machine learning.* McGraw-Hill. [MJ]
Miyashita, Y. (1993) Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience* 16:245–63. [aMP]
Mogi, K. & Barlow, H. B. (1998) The variability in neural responses from MT. *Journal of Physiology (London)* 515:101P–102P. [HB]
Morrison, C. M. & Ellis, A. W. (1995) Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory and Cognition* 21(1):116–33. [aMP]
Morton, J. (1969) The interaction of information in word recognition. *Psychological Review* 76:165–78. [aMP]
Mountcastle, V. B. (1997) The columnar organization of the neocortex. *Brain* 120:701–22. [aMP]
Murdock, B. B., Jr. (1982) A theory for the storage and retrieval of item and associative information. *Psychological Review* 89:609–26. [MJ]
Murre, J. M. J. (1992) *Learning and categorization in modular neural networks.* Harvester Wheatsheaf. [aMP]
Murre, J. M. J., Phaf, R. H. & Wolters, G. (1992) Calm: Categorizing and learning module. *Neural Networks* 5:55–82. [aMP, RHP]
Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [CPB]
Newell, A. & Rosenbloom, P. S. (1981) Mechanisms of skill acquisition and the law of practice. In: *Cognitive skills and their acquisition,* ed. J. R. Anderson. Erlbaum. [AH]
Newsome, W. T. (1999) Algorithms for reading out a cortical map. *Proceedings of the Australian Neuroscience Society* 10:13. [rMP, SRS]
Newsome, W. T., Britten, K. H. & Movshon, J. A. (1989) Neuronal correlates of a perceptual decision. *Nature* 341:52–54. [HB, aMP, SRS]
Nigrin, A. L. (1993) *Neural networks for pattern recognition.* MIT Press. [JSB, arMP]
Norris, D. (1994a) A quantitative multiple-levels model of reading aloud. *Journal of Experimental Psychology: Human Perception and Performance* 20:1212–32. [aMP]
(1994b) SHORTLIST: A connectionist model of continuous speech recognition. *Cognition* 52:189–234. [arMP]
Nosofsky, R. M. (1985) Luce's choice model and Thurstone's categorical judgement model compared. *Perception and Psychophysics* 37(1):89–91. [aMP]
(1986) Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: Learning, Memory and Cognition* 115(1):39–57. [aMP, DCP]
(1987) Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition* 13(1):87–108. [aMP]
(1990) Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34:393–418. [aMP]
Nosofsky, R. M. & Palmeri, T. J. (1997) An exemplar-based random walk model of speeded classification. *Psychological Review* 104(2):266–300. [aMP, ADP]
Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. (1994) Rule-plus-exception model of classification learning. *Psychological Review* 101:53–79. [rMP]
Oden, G. C. & Massaro, D. W. (1978) Integration of featural information in speech perception. *Psychological Review* 85:172–91. [aMP]
Oram, M. W., Földiák, P., Perrett, D. I. & Sengpiel, F. (1998) The "ideal homunculus": Decoding neural population signals. *Trends in Neurosciences* 21:259–65. [aMP]
O'Reilly, R. C. (1996) Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation* 8(5):895–938. [rMP, DCP]
O'Reilly, R. C. & Farah, M. J. (1999) Simulation and explanation in neuropsychology and beyond. *Cognitive Neuropsychology* 16:49–72. [AMB]
Page, M. P. A. (1993) *Modelling aspects of music perception using self-organizing neural networks.* Ph. D. thesis, University of Wales, College of Cardiff, Cardiff, U. K. [arMP]
(1994) Modelling the perception of musical sequences with self-organizing neural networks. *Connection Science* 6(2/3):223–46. [arMP]
Page, M. P. A. & Nimmo-Smith, I. (in preparation) Properties of a localist, connectionist, Thurstonian model. [aMP]
Page, M. P. A. & Norris, D. (1997) A localist implementation of the primacy model of immediate serial recall. In: *Localist connectionist approaches to human cognition,* ed. J. Grainger & A. M. Jacobs. Erlbaum. [arMP]
(1998) The primacy model: A new model of immediate serial recall. *Psychological Review* 105:761–81. [CPB, JSB, arMP]
Palmeri, T. J. (1997) Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23(2):324–54. [aMP]
Panzeri, S., Schultz, S. R., Treves, A. & Rolls, E. T. (1999) Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London B* 266:1001–12. [SRS]
Pearce, J. M. (1994) Similarity and discrimination. *Psychological Review* 101:587–607. [arMP, ADP]
Pelletier, F. J. & Berkeley, I. (1995) Vagueness. In: *The Cambridge dictionary of philosophy,* ed. R. Audi. Cambridge University Press. [ISNB]
Phaf, R. H., Van der Heijden, A. H. C. & Hudson, P. T. W. (1990) SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology* 22:273–341. [RHP]
Phaf, R. H. & van Immerzeel, M. A. (1997) Simulations with a connectionist model for implicit and explicit memory tasks. In: *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society,* ed. M. G. Shafto & P. Langley. Erlbaum. [RHP]
Phelps, M. E., Huang, S. C., Hoffman, E. J., Selin, C., Sokoloff, L. & Kuhl, D. E. (1979) Tomographic measurement of local cerebral glucose metabolic rate in humans with (F-18)2-fluoro-2-deoxy-D-glucose: Validation of method. *Annals of Neurology* 6:371–88. [JTD]
Pickering, A. D. (1997) New approaches to the studying of amnesic patients: What can a neurofunctional philosophy and neural network methods offer? *Memory* 5(1/2):255–300. [aMP, ADP]
Pike, R. (1984) A comparison of convolution and matrix-distributed memory systems. *Psychological Review* 91:281–94. [MJ]
Plaut, D. C. (1997) Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes* 12:767–808. [rMP, DCP]
Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103:56–115. [JSB, arMP, DCP]
Posner, M. I. & Keele, S. W. (1968) On the genesis of abstract ideas. *Journal of Experimental Psychology* 77:353–63. [JEI]
Purves, D. (1994) *Neural activity and the growth of the brain.* Cambridge University Press. [PCMM]
Quartz, S. R. & Sejnowski, T. J. (1997) The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences* 20:537–96. [aMP]
Quillian, M. R. (1968) Semantic memory. In: *Semantic information processing,* ed. M. Minsky. MIT Press. [aMP]
Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–86. [IV]
Raiguel, S., Van Hulle, M. M., Xiao, D.-K., Marcar, V. L. & Orban, G. A. (1995) Shape and spatial distribution of receptive fields and antagonistic motion surrounds in the middle temporal area (V5) of the macaque. *European Journal of Neuroscience* 7:2064–82. [HB]
Raijmakers, M. E. J. & Molenaar, P. C. M. (1996) Exact ART: A complete implementation of an ART network in real time. *Neural Networks* 10:649–69. [PCMM]
Raijmakers, M. E. J., van der Maas, H. L. J. & Molenaar, P. C. M. (1997) Numerical bifurcation analysis of distance-dependent on-center off-surround shunting neural networks with external input. *Biological Cybernetics* 75:495–507. [PCMM]
Ramsey, W. (1997) Do connectionist representations earn their explanatory keep. *Mind and Language* 12(1):34–66. [aMP]
Rao, R. P. N. & Ballard, D. H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* 2:79–87. [MPY]
Ratcliff, R. (1978) A theory of memory retrieval. *Psychological Review* 85(2):59–108. [aMP]

(1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review* 97:285–308. [aMP]
Ratcliff, R. & McKoon, G. (1997) A counter model for implicit priming in perceptual word identification. *Psychological Review* 104:319–43. [RHP]
Read, S. J. & Marcus-Newhall, A. (1993) Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology* 65:429–47. [TRS]
Read, S. J. & Miller, L. C. (1998) On the dynamic construction of meaning: An interactive activation and competition model of social perception. In: *Connectionist models of social reasoning and social behavior,* ed. S. J. Read & L. C. Miller. Erlbaum. [TRS]
Redington, M. & Chater, N. (submitted) Computational models of artificial grammar learning. [AK]
Reike, F., Warland, O., Ruyter van Stevenick, R. & Bialek, W. (1996) *Spikes: Exploring the neural code.* MIT Press. [MPY]
Repp, B. H., Liberman, A. M., Eccardt, T. & Pesetsky, D. (1978) Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance* 4:621–37. [SG]
Rickard, T. C. (1997) Bending the power-law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General* 126:288–311. [rMP]
Roberts, S. & Pashler, H. (in press) How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review.* [DCP]
Robins, A. (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science* 7:123–46. [aMP]
Roelfsema, P. R., Engel, A. K., König, P. & Singer, W. (1996) The role of neuronal synchronization in response selection: A biologically plausible theory of structured representations in the visual cortex. *Journal of Cognitive Neuroscience* 8(6):603–25. [aMP]
Rolls, E. T. (1989) Parallel distributed processing in the brain: Implications of the functional architecture of neuronal networks in the hippocampus. In: *Parallel distributed processing: Implications for psychology and neurobiology,* ed. R. G. M. Morris. Oxford University Press. [aMP]
Rolls, E. T., Critchley, H. D. & Treves, A. (1996) Representation of olfactory information in the primate orbitofrontal cortex. *Journal of Neurophysiology* 75(5):1982–96. [aMP]
Rolls, E. T., Treves, A., Tovee, M. J. & Panzeri, S. (1997) Information in the neuronal representations of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience* 4:309–33. [SRS]
Rosch, E. (1975) Cognitive representation of semantic categories. *Journal of Experimental Psychology: General* 104:192–233. [JEI]
(1977) Principles of categorization. In: *Cognition and categorization,* ed. E. Rosch & B. Lloyd. Erlbaum. [JEI]
Rosenblatt, F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65:386–408. [GAC]
(1962) *Principles of neurodynamics.* Spartan Books. [GAC]
Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning internal representations by error propagation. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. Foundations,* ed. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. MIT Press. [ISNB, GAC, aMP]
Rumelhart, D. E. & McClelland, J. L. (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review* 89:60–94. [SG, aMP, RHP]
Rumelhart, D. E., McClelland, J. L. & PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. Foundations.* MIT Press. [aMP]
Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models,* J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press. [DCP]
Rumelhart, D. E. & Zipser, D. (1985/1986) Feature discovery by competitive learning. (1985) in: *Cognitive Science* 9:75–112. [DCP]. (1986) in: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. Foundations,* ed. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. MIT Press. [aMP]
Sakai, K. & Miyashita, Y. (1991) Neural organization for the long-term memory of paired associates. *Nature* 354:152–55. [aMP]
Sakai, K., Naya, Y. & Miyashita, Y. (1994) Neuronal tuning and associative mechanisms in form representation. *Learning and Memory* 1:83–105. [arMP]
Salum, C., Roque da Silva, A. & Pickering, A. (1999) Striatal dopamine in attentional learning: A computational model. *Neurocomputing* 26–27:845–54. [ADP]
Salzman, C. D. & Newsome, W. T. (1994) Neural mechanisms for forming a perceptual decision. *Science* 264:231–36. [aMP]
Samuel, A. G., van Santen, J. P. H. & Johnston, J. C. (1982) Length effects in word perception: We is better than I but worse than you or them. *Journal of Experimental Psychology: Human Perception and Performance* 8:91–105. [SG]
(1983) Reply to Matthei: We really is worse than you or them, and so are ma and pa. *Journal of Experimental Psychology: Human Perception and Performance* 9:321–22. [SG]
Scannell, J. W. & Young, M. P. (2000) Primary visual cortex within the cortico-cortico-thalamic network. In: *Cerebral cortex, vol. 15: Cat primary visual cortex,* ed. A. Peters, E. G. Jones & B. R. Payne. Plenum Press. [MPY]
Schultz, S. R. & Panzeri, S. (1999) Redundancy, synergy and the coding of visual information. *Proceedings of the Australian Neuroscience Society* 10:25. [SRS]
Schulz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R. & Dickinson, A. (1995) Reward-related signals carried by dopamine neurons. In: *Models of information processing in the basal ganglia,* ed. J. C. Houk, J. L. Davis & D. G. Beiser. MIT Press. [ADP]
Seidenberg, M. S. (1993) Connectionist models and cognitive theory. *Psychological Science* 4:228–35. [arMP]
Seidenberg, M. S. & McClelland, J. L. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96:523–68. [JSB, AMB, SG, aMP]
Shadlen, M. N. & Newsome, W. (1998) The variable discharge of cortical neurons: Implications for connectivity, computations and coding. *Journal of Neuroscience* 18:3870–96. [MPY]
Shanks, D. R. (1991) Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17:433–43. [SF]
Sharkey, N. E. & Sharkey, A. J. C. (1995) An analysis of catastrophic interference. *Connection Science* 7(3/4):301–29. [aMP]
Shastri, L. (1988) *Semantic networks: An evidential formalization and its connectionist realization.* Morgan Kaufmann/Pitman. [LS]
(1991) Relevance of connectionism to AI: A representation and reasoning perspective. In: *Advances in connectionist and neural computation theory, vol. 1,* ed. J. Barnden & J. Pollack. Ablex. [LS]
(1997) A model for rapid memory formation in the hippocampal system. In: *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society,* 680–85. Erlbaum. [LS]
(1999a) Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence* 11:79–108. [LS]
(1999b) Recruitment of binding and binding-error detector circuits via long-term potentiation. *Neurocomputing* 26–27:865–74. [LS]
(1999c) A biological grounding of recruitment learning and vicinal algorithms. Technical Report TR-99–009, International Computer Science Institute, Berkeley, CA, 94704, April 1999. [LS]
Shastri, L. & Ajjanagadde, V. (1993) From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal asynchrony. *Behavioral and Brain Sciences* 16:417–94. [CL-S, aMP, LS]
Shastri, L., Grannes, D. J., Narayanan, S. & Feldman, J. A. (in press) A connectionist encoding of parameterized schemas and reactive plans. In: *Hybrid information processing in adaptive autonomous vehicles,* ed. G. K. Kraetzschmar & G. Palm. Lecture notes in computer science. Springer-Verlag. [LS]
Shepard, R. N. (1958) Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review* 65(4):242–56. [aMP]
(1987) Towards a universal law of generalization for psychological science. *Science* 237:1317–23. [aMP]
Shultz, T. R. (1998) Generative network models of cognitive development: Progress and new challenges. *NIPS*98 workshop on development and maturation in natural and artificial structures.* Morgan Kaufmann. [TRS]
Shultz, T. R. & Lepper, M. R. (1996) Cognitive dissonance reduction as constraint satisfaction. *Psychological Review* 103:219–40. [TRS]
Sloman, S. A. & Ripa, L. J., eds. (1998) *Similarity and symbols in human thinking.* MIT Press. [SO]
Sloman, S. A. & Rumelhart, D. E. (1992) Reducing interference in distributed memories through episodic gating. In: *From learning theory to cognitive processes: Essays in honor of William K. Estes,* ed. A. S. Healy, S. Kosslyn & R. Shiffrin. Erlbaum. [aMP]
Smith, E. R. & DeCoster, J. (1998) Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology* 74:21–35. [TRS]
Smith, J. D. & Minda, J. P. (1998) Prototypes in the mist: The early epochs of

category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:1411–36. [rMP]
Smith, P. L. (1995) Psychophysically principled models of visual simple reaction time. *Psychological Review* 102:567–93. [AH]
Smolensky, P. (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159–216. [SF]
(1991) Connectionism, constituency, and the language of thought. In: *Meaning in mind: Fodor and his critics,* ed. Loewer, B. & Rey, G. Blackwell. [ISNB]
Snowden, J. S., Griffiths, H. L. & Neary, D. (1996) Semantic-episodic memory interactions in semantic dementia: Implications for retrograde memory function. *Cognitive Neuropsychology* 13(8):1101–37. [aMP]
Spellman, B. A., Ullman, J. B. & Holyoak, K. J. (1993) A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues* 49:147–65. [TRS]
Squire, L. R., Cohen, N. J. & Nadel, L. (1984) The medial temporal region and memory consolidation: A new hypothesis In: *Memory consolidation,* ed. H. Weingarten & E. Parker. Erlbaum. [aMP]
St. John, M. F. & Shanks, D. R. (1997) Implicit learning from an information processing standpoint. In: *How implicit is implicit learning?,* ed. D. C. Berry. Oxford University Press. [AK]
Taylor, I. & Greenhough, M. (1994) Modelling pitch perception with adaptive resonance theory artificial neural networks. *Connection Science* 6:135–54. [rMP]
Tetewsky, S., Shultz, T. R. & Buckingham, D. (1994) Assessing interference and savings in connectionist models of recognition memory. Paper presented at the 35th Annual Meeting of the Psychonomic Society, St. Louis, MO. [TRS]
Thagard, P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* 12:435–502. [TRS]
Thagard, P., Holyoak, K. J., Nelson, G. & Gochfield, D. (1990) Analog retrieval by constraint satisfaction. *Artificial Intelligence* 46:259–310. [TRS]
Thagard, P. & Millgram, E. (1995) Inference to the best plan: A coherence theory of decision. In: *Goal-driven learning,* ed. A. Ram & D. B. Leake. MIT Press. [TRS]
Thelen, E. & Smith, L. B. (1994) *A dynamic systems approach to the development of cognition and action.* MIT Press. [PCMM]
Thomas, E., Van Hulle, M. & Vogels, R. (1999) The study of the neuronal encoding of categorization with the use of a Kohonen network. In: *Proceedings of the Fifth Neural Computation and Psychology Workshop. Connectionist Models in Cognitive Neuroscience, Birmingham, England,* 218–27. Springer-Verlag. [RMF]
Thorndike, E. L. (1911) *Animal intelligence: Experimental studies.* Macmillan. [CM]
Thorpe, S. (1995) Localized versus distributed representations. In: *The handbook of brain theory and neural networks,* ed. M. A. Arbib. MIT Press. [aMP]
Thurstone, L. L. (1927) Psychophysical analysis. *American Journal of Psychology* 38:368–89. [CB, aMP]
Tippett, L. J., Glosser, G. & Farah, M. J. (1996) A category-specific naming impairment after temporal lobectomy. *Neuropsychology* 34:139–46. [rMP]
Usher, M. & McClelland, J. L. (1995) On the time course of perceptual choice: A model based on principles of neural computation. Technical Report PDP.CNS.95.5, Dept. of Psychology, Carnegie Mellon University. [AH, aMP]
Valiant, L. (1994) *Circuits of the mind.* Oxford University Press. [aMP, LS]
Van der Maas, H. L. J. & Molenaar, P. C. M. (1992) Stage-wise cognitive development: An application of catastrophe theory. *Psychological Review* 99:395–417. [PCMM]
Van Santen, J. P. H. & Bamber, D. (1981) Finite and infinite state confusion models. *Journal of Mathematical Psychology* 24:101–11. [aMP]
Van Zandt, T. & Ratcliff, R. (1995) Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin and Review* 2(1):20–54. [aMP]
Visser, I., Raijmakers, M. E. J. & Molenaar, P. C. M. (1999) Prediction and reaction times in implicit sequence learning. Technical Report 1999–02, Developmental Psychology Institute of the University of Amsterdam. http://develop.psy.uva.nl/users/ingmar/ [IV]
Vogels, R. (1999) Categorization of complex visual images by rhesus monkeys. Part 2: Single cell study. *European Journal of Neuroscience* 11:1239–55. [RMF, rMP]
Von Helmholtz, H. (1885/1954) *On the sensations of tone as a physiological basis for the theory of music.* Dover. [CM]
Wang, G., Tanifuji, M. & Tanaka, K. (1998) Functional architecture in monkey inferotemporal cortex revealed by *in vivo* optical imaging. *Neuroscience Research* 32(1):33–46. [SRS]
Warrington, E. K. & Shallice, T. (1984) Category-specific semantic impairments. *Brain* 107:829–59. [RMF]
Werbos, P. J. (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences. Unpublished Ph. D. dissertation, Harvard University. [GAC]
Wickelgreen, W. A. (1979) Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review* 86(1):44–60. [LS]
Yellott, J. I., Jr. (1977) The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology* 15:109–44. [CB, aMP]
Young, A. W. & Burton, A. M. (1999) Simulating face recognition: Implications for modelling cognition. *Cognitive Neuropsychology* 16:1–48. [AMB]
Young, M. P. & Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex. *Science* 256:1327–31. [aMP, MPY]
(1993) An analysis at the population level of the processing of faces in the inferotemporal cortex. In: *Brain mechanisms of perception and memory: From neuron to behavior,* ed. T. Ono, L. R. Squire, M. E. Raichle, D. I. Perrett & M. Fukuda. Oxford University Press. [CM, arMP, MPY]
Zohary, E. (1992) Population coding of visual stimuli by cortical neurons tuned to more than one dimension. *Biological Cybernetics* 66:265–72. [SRS]
Zohary, E., Shadlen, M. N. & Newsome, W. T. (1994) Correlated neuronal discharge rate and its implication for psychophysical performance. *Nature* 370:140–43. [SRS]
Zorzi, M., Houghton, G. & Butterworth, B. (1998) Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance* 24:1131–61. [aMP]