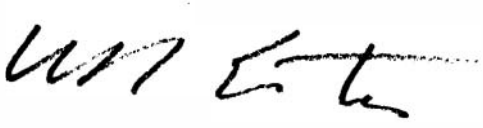

Annu. Rev. Psychol. 1991. 42:1–28

COGNITIVE ARCHITECTURES FROM THE STANDPOINT OF AN EXPERIMENTAL PSYCHOLOGIST

W. K. Estes

Department of Psychology, Harvard University, Cambridge, Massachusetts 02138

KEY WORDS: connectionist models, distributed memory, category learning, memory format, memory traces and associations

---

CONTENTS

INTRODUCTION

At times the concept of an architecture seems to me entirely too grand for psychology at the present stage of theory development. But at other times, it seems that I have been seeking an appropriate architecture for cognition

0066-4308/91/0201-0001$02.00

during the whole half century of my professional life. When I entered research in psychology in the 1940s, the term architecture was not in use, but there was much concern with the problem of establishing a suitable framework for theories in the broad area now termed cognition. The nature of the proper framework seemed obvious.

> Since propositions concerning psychological events are verifiable only to the extent that they are reducible to predictions of behavior under specified environmental conditions, it appears likely that greatest economy and consistency in theoretical structure will result from statement of all fundamental laws in the form R = f(S), where R and S represent behavioral and environmental variables, respectively. (Estes 1950:94)

My, such confidence. Could any future turn of events overturn the insight that the stimulus-response architecture is basic to, and indeed sufficient for, psychological theory? Such a turn of events was, in fact, not far in the future. During the next decade and a half, the focus of my research moved from animal learning and conditioning to human visual processing and short-term memory, and it is remarkable what a shift of research interest can do for one's theoretical outlook.

> [T]he type of theory to which we are evidently being led by a wide range of current experimental developments differs in a number of major respects from classical association and stimulus-response theories. There is now a large amount of detailed and extensive evidence which indicates that theoretical interpretations will gain more than they will lose in the way of parsimony by accentuating and sharpening the distinction between the processes of learning and response selection. A striking simplification in the interpretation of many learning phenomena is achieved at a stroke if we conceive the result of an organism's experiencing a sequence of events to be, not simply the strengthening or weakening of the constituent stimulus-response connections, but rather the establishment in memory storage of a *representation* of the entire sequence . . . (Estes 1969:185–86, Ital. added)

That passage appeared at a time when I had one hand still in conditioning research while the other was working in information processing. But it wasn't long before both hands were at work together in the new framework.

> Just as the physical sciences can be conceived as the study of energy in its many aspects, the behavioral and social sciences can be characterized in terms of their concern with the processing and transformation of information. The adaptation of living organisms to an ever-changing environment depends upon the ability to acquire information about environmental regularities and to use this information as the basis of adaptive response. (Estes 1975b:1)

From that time on, although learning continued to be one of my central research interests, I viewed it, not as a simple mechanism for strengthening and weakening response tendencies, but as a collection of processes responsible for the building and elaborating of memory structures. I and others who bridged the gap between the old and newer paradigms faced the task of

discovering how to replace the comfortable but now clearly inadequate stimulus-response framework with a framework, or architecture, capable of accommodating the variety of models flourishing in the broad domain of information processing.

TOWARD A COGNITIVE ARCHITECTURE

The concept of cognitive architecture has at present no generally accepted definition and can only be understood by observing it in use. The concept was imported into the cognitive literature from computer science, and not only its meaning but also its applicability in the new context is unsettled.

In computer science, architecture refers to the general characteristics of a computer that make programming possible. By far the most familiar version is the von Neumann architecture, the basis of virtually all digital computers. In this architecture, informational units, the physical embodiments of strings of binary digits, are stored in locations that can be accessed either by their addresses or their contents. The machine carries out its computations by sequencing through the stored items and, in effect, applying operators such as comparison or logical combination. The informational units take on meaning by virtue of their correspondence to familiar symbols like numbers, letters, or mathematical operators. Thus, in artificial intelligence, the programming languages that enable computers to manifest human-like cognitive functions can be characterized as symbol processors.

In view of the many similarities between the computer and the human being, viewed as information processing systems, which have done much to spark the development first of artificial intelligence and then of cognitive science, it has seemed to many investigators that the symbol-processing architecture should carry over from one realm to the other. For cognitive science, the architecture would characterize the set of informational (as distinguished from neural) structures and symbol-manipulating processing that underlies all of the specific cognitive models and theories. This expectation has in fact been realized for some subdomains of cognition—for example, problem solving (Newell & Simon 1972). In other subdomains, however, popular lines of research of the 1970s demanded concepts like spreading activation (Collins & Loftus 1975) and automatization (Schneider & Shiffrin 1977) that seemed quite out of the spirit of the symbol-processing architecture.

How, then, are we to arrive at a satisfactory characterization of a general cognitive architecture? I can see two possible routes, one direct and the other indirect. The direct route is for some individual investigator, or possibly group of investigators, to develop and present a proposed architecture, just as is routinely done for more limited theories. Recently there have been several

major proposals of this sort. One of these is the Adaptive Control of Thought (ACT*) architecture of Anderson (1983), another the State, Operator, and Result (SOAR) architecture of Newell and his associates (Laird et al 1987; Newell 1990). Perhaps prematurely, some would include in this category a new contender, the parallel, distributed processing (PDP) or "connectionist" architecture (McClelland & Rumelhart 1986; Rumelhart & McClelland 1986a).

A difficulty with the direct approach is that formulating a whole architecture is an enterprise of such complexity that the large number of decisions about details must reflect the preferences and biases of the formulator, and the product is very difficult to evaluate. Unlike the situation with experimental investigation and limited theory construction, we have no stock of tested methods for the construction of whole architectures, nor generally accepted standards for accessing their merits. As pertinent evidence, consider that Anderson's deservedly influential architecture has run through some half dozen versions in a short period of years (Anderson 1976, 1983), the shifts often unaccompanied by any empirical developments that appear compelling to an outside observer.

The alternative I see to the direct approach is an indirect one based on the idea that if cognitive theory has any general architecture, it must have evolved over the last century of research on cognition. If so, then the architecture should be discoverable by adapting the standard methods of scientific investigation, that is, by tracing the development of the lines of theory that have been influential over the century, examining their similarities and differences, and discovering whether there are commonalities of structure general enough to qualify as the basis of an architecture. Whether this inductive approach will work is an open question, of course, and I expect this essay to accomplish no more than to provide an illustration that may set the stage for discussion.

In this chapter, I sketch the evolution of the notion of a cognitive architecture over the last half century from my own standpoint as an observer and participant; I conclude with a discussion of current issues and prospects. My interest in what may seem to be an esoteric concept derives from its relevance to my longtime concern with the problem of comparing and testing mathematical and computer models in psychology (Estes 1975a, 1986a). Verbally formulated theories are notoriously difficult to test because of the inadequacy of verbal arguments for deriving their implications or even for ascertaining when the implications of two such theories differ. But the problem does not automatically vanish when theories are cast as mathematical or computer simulation models, for superficial differences can mask basic commonalities. The task of determining when apparent differences between theories are testable requires their examination within a common framework, and for

theories of cognition this common framework would be the cognitive architecture.

A salient aspect of my personal history is the corresponding of shifts in ideas about architecture with shifts in loci of research activity. If we regard a research domain—for example, cognition—as a collection of different kinds of experimental subdomains, theory construction is relatively straightforward. We choose for any given subdomain the concepts that prove most serviceable for its interpretation, thus generating what may be termed local architectures (see Table 1).

However, suppose we take our task to be, not interpreting various clusters of experiments, but interpreting the cognizing organism? Then the experimental clusters are just the results of looking at the organism from different perspectives, and we need some way to fit the interpretations of them together in a more comprehensive structure. The question then arises whether as ambitious a goal as Newell's "general cognitive architecture" (Newell 1990) is a feasible target. We cannot foresee whether achieving such a goal is possible, but continuing pursuit of that goal may nonetheless be the best way to ensure that cognitive science will achieve some generality of theory in spite of the enormous complexities of its subject matter and the pressures to settle for heuristic principles and local models.

MEMORY FORMAT: ASSOCIATION OR TRACE

One essential constituent of a cognitive architecture is a specification of the form of information storage in memory. Two more or less parallel approaches to this problem have run through the history of memory theory, one centered in the tradition of association theory and the other in the concept of a memory trace. Though its roots are traceable to the British associationists James Mill and David Hume, the concept of association was given its first formulation as a theoretical principle with experimental interpretations by Ebbinghaus (1964

Table 1 Local architectures for subdomains of cognitive research

Subdomain	Local architecture
conditioning	stimulus-response
language	rules, semantic nets
classification	array structures
short-term memory	list structures
perception	multidimensional space
knowledge acquisition	propositional networks
problem solving	problem spaces, production systems

[1885]); its continuing elaboration over the next several decades was thoroughly reviewed by Robinson (1932). The central idea was that any form of memorization results in the laying down of associations between units, with the property that if an association is formed between units A and B, then later activation of A tends to lead to reactivation of B. For the earlier associationists, the units were vaguely defined ideas; for Ebbinghaus, they were mental representations of the elements, usually words or nonsense syllables, of the lists he so laboriously studied. Learning consists, not in modifying the units, but only of establishing associations between units.

Almost coextensive with association theory has been the development of models of memory based on the concept of a memory trace, or engram, the modern version dating from the work of Hollingworth (1913, 1928) and Semon (1921). In this tradition, it is assumed that any learning experience results in the deposit of a trace in the memory system. Whatever is perceived may enter into the trace, which typically takes the form of a sensory image. Perceived or learned relationships among objects or events are embodied in the trace itself, rather than in associations among units. Memory traces give rise to reconstruction or recall of an experience by virtue of the process of *redintegration,* whereby later perception of some portion of the stimulus pattern comprising a trace leads to reactivation of the entire pattern, as when a glimpse of some portion of a familiar face or scene gives rise to an image of the whole.

As a preliminary to examining architectural properties of current models, I give a brief sketch of each of these lines of theory, organized in terms of some salient theoretical attributes.

Association Theory

On the whole, there has been remarkably little change in the structural assumptions of association theory over the last century. Ebbinghaus's formulation was based on a single layer of interconnected associative units. Following study of a list of items, A, B, C, D, and E, in order (the capital letters denoting any type of item), direct pairwise associations would be established between a starting signal and A, between A and B, between B and C, and so on, providing the basis for subsequent recall of the whole list on presentation of the starting signal. This structure would be fragile, however, for if the connection between A and B were impaired, the remainder of the list would be unrecallable. For this reason and others, Ebbinghaus admitted also indirect associations between items such as A and C or A and D that were not contiguous during study. The indirect associations are typically weaker than the direct ones, but they produce a structure that is less fragile in the face of possible interfering factors.

As research on simple learning, both human and animal, progressed over

several decades, it became apparent that the simple association model could not explain why learning sometimes occurs on an all-or-none basis but sometimes requires many repetitions of an experience or why temporal spacing of learning experiences is a critical factor in retention. The remedy I proposed in my first contribution to learning theory was to introduce a layer of abstract units (originally termed "stimulus elements," later "memory elements") that were interposed between stimuli and responses; on a learning trial, a sample of these units, corresponding to the stimulus aspects attended to by the learner, could become associated with the correct ("reinforced") response category (Estes 1950). Also, the units were assumed to fluctuate in level of availability over time, providing a mechanism to account for temporal aspects of learning and retention (Estes 1955).

The minimal structural assumptions of classical association and stimulus sampling models seemed satisfactory in an age of high concern for parsimony and operationism but offered few resources for addressing problems of organization in memory. This limitation came to be felt acutely when list and paired-associate memorization were replaced by free recall as the experimental paradigm of choice for studies of verbal learning in the 1950s. In a standard free recall study, a subject hears or reads a list of words, presented singly, then attempts to recall the words in any order. Typically there proves to be little correlation between presentation order and recall order; rather, words tend to be clustered in recall, with semantically related words tending to occur adjacently in recall regardless of their input positions (Bousfield 1953). This observation gave rise to the idea that associative links may fan out from a studied word to a number of others semantically related to it, allowing growth of a hierarchical structure of the kind illustrated in Figure 1 (Mandler 1967); then recall can be effected by proceeding from the topmost node of the hierarchy downward, reading out the cluster of words associated with each lower-order node as it is encountered. The topmost node may be viewed as corresponding to a representation of the list as a unit (hence the common designation *list marker*), the next level of nodes to category labels, especially if category labels are supplied by the experimenter prior to recall, and the nodes at lower levels to members of categories. However, these identifications are not essential to the concept of a hierarchical structure, and the upper level nodes may be regarded as abstract constituents of the structure, *control elements* in the general hierarchical model of Estes (1972a), that serve an organizational function but have no specific empirical referents.

This notion of a hierarchical memory structure has been widely extended simply by redefining the types of units that correspond to the nodes at various levels. Thus, for the purpose of representing the mental lexicon—that is, an individual's long-term memory for vocabulary—the nodes are taken to correspond to words and semantic categories (Collins & Quillian 1972); for the

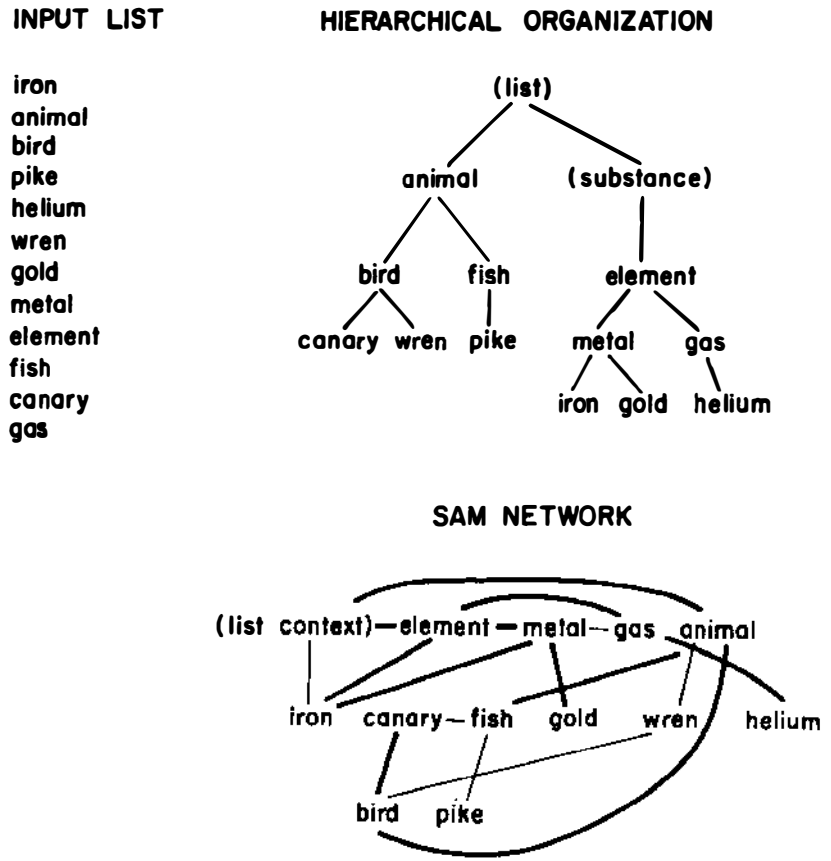

Figure 1 Memory for a list of words presented in a hypothetical free-recall experiment is represented in hierarchical and nonhierarchical network structures. Widths of lines in the SAM diagram signify strengths of associative connections (presumably deriving from different degrees of familiarity in the learner's experience).

interpretation of factual memory, the nodes are taken to correspond to concepts (Anderson & Bower 1973) or to propositions (Anderson 1976, 1983).

The Search of Associative Memory (SAM) model of Raaijmakers & Shiffrin (1981) is perhaps the most general current representative of association theory and has certainly yielded the most detailed interpretations of a wide variety of phenomena of recall and recognition (Clark & Shiffrin 1987; Gillund & Shiffrin 1984; Gronlund & Shiffrin 1986); Shiffrin et al 1989). On first encounter, the SAM model seems much more elaborate than any of its predecessors. The elaboration has, however, occurred largely through the

augmentation of the classical scheme by a variety of control processes, with little actual change in the basic structure. In the free recall example, each word studied is represented by a node in a memory network, with connections between those representing words that were adjacent in the input list or that were rehearsed together during study. As shown in the lower part of Figure 1, the network resembles the hierarchical structure except that superordinate-subordinate relationships are not immediately apparent. However, in the SAM network, associations differ in strength (shown by light and heavy lines in the figures), and the stronger associations would be expected to be generally the same as those represented in the hierarchy. Thus, it appears that the differences between the hierarchical model and SAM are at the level of process other than at the level of structure.

This analysis suggests a distinction between the features common to all associative models, which may reasonably be taken to comprise the architecture, and the features that distinguish among them. All associative models are based on networks of nodes and associative links, the products of learning being the establishment of links or the strengthening of already established ones. The nodes are units whose function is to enter into associations; they may be classified according to the kinds of objects they correspond to in empirical interpretations, but they are opaque in that there is no direct access to their internal structures. In SAM, the units are referred to as images and described as containing clusters of information. In actual implementations of the model, however, the information contained in an image is manifest only in the probabilities of evocation of the responses, usually words, associated with it.

Trace Theory

TRACE MODELS FOR MEMORIZATION In the second main type of memory theory, the product of a learning experience is the laying down of a memory trace, the information accrued being represented in the trace itself rather than in connections among traces. The beginnings of formalization of the concept appeared when Hollingworth (1913) added the notions of a threshold and of strength, or degree of familiarity, of a trace, which varied with repetition and recency of activation. In Hollingworth's version, a percept that activated a stored trace with a strength falling above the threshold would be recognized as "old"—that is, as representing a previously experienced pattern of objects or events—whereas a strength falling below the threshold would yield nonrecognition. Except for some refinements contributed by the importation of signal detectability theory (Murdock 1974), the threshold model has served as the canonical interpretation of recognition down to the present.

The concept of a memory trace was important in gestalt psychology (Koffka 1935), the main new assumption being that a trace is not a static

representation but rather one that changes autonomously over time in the direction of better conformity to gestalt laws of symmetry, "good figure," and the like. Attempts to adduce experimental support for autonomous processes yielded mixed results, however, and the concept has remained a bypath in the evolution of present-day memory theory.

The next major development in trace theory was the introduction of the idea that information is stored in a trace in the form of values on a set of attributes or dimensions. This concept was formalized by Bower (1967) in his model based on a multicomponent memory trace; with a strong boost from an influential article by Underwood (1969), it quickly became the standard interpretation. Now a memory trace was conceived as a vector, or list, of features, each representing a value on an attribute or dimension (usually, but not necessarily, binary). Important assumptions in Bower's version were that features could fluctuate in level of availability over time in the manner of stimulus sampling theory (Estes 1955) and could be independently lost as a consequence of factors responsible for forgetting. Bower showed that trace theory, so elaborated, could yield quantitative accounts of numerous phenomena of recognition and simple verbal learning and forgetting.

RELATIONAL INFORMATION IN TRACE MODELS An important limitation of the multicomponent trace model, as of stimulus sampling models, was the lack of any principled way of handling relational information. An adequate theory must be able to account for the fact that both human learners and higher animals readily discriminate between patterns and their components. It is trivial, for example, to remember that the number of the hotel room one has checked into is 49 even though there are rooms 4 and 9 on the same floor assigned to other people, but an unelaborated multitrace model would have to predict interference. This problem was a focus of interest for members of my Mathematical Psychology Laboratory at Rockefeller University in the 1970s. It was one of our group, Douglas Medin, who came up with the insight that sensitivity to relations among elements of a pattern need not be based on the storage of some kind of special relational information (as relational features) in the memory trace but may, rather, emerge in the computations performed when perceived patterns are compared to stored representations. In the discrimination model he formulated, features, or attributes, of a stimulus are stored in a multicomponent trace, and the process for accessing stored representations is a computation of a multiplicative measure of the similarity of a perceived stimulus pattern to each member of the relevant memory array (Medin 1975). Any decision called for on the part of the learner is based on these similarity measures, and because of the multiplicative character of the similarity function, discrimination of patterns from their components can occur automatically. The simple but powerful idea that access to stored

memories is achieved by a computation based on similarity between perceived and stored patterns is embodied in many contemporary theories, among them the "resonance" model of Ratcliff (1978) and the "Minerva II" simulation model of Hintzman (1986). The outcome of many decades of evolution of trace theory seems, then, to be a schema based on the storage of multi-component memory vectors in partitioned arrays accessible via similarity computations—evidently as deserving of the appellation "architecture" as the general schema of association theory.

A Common Architecture for Trace and Association

Have we at this point arrived at two architectures, one for association and one for trace theory? I surmise that actually they are homologous, the familiar networks and hierarchical structures of current association theory being just graphical depictions of relationships among percepts and memories that are common to both kinds of theory. For a simple illustration of the correspondence, consider the memorization of a list of words W1, W2, W3, . . . Wn. In an associative model the memory structure formed could be depicted as

S' – W1' – W2' – W3' – . . . –Wn',

the units corresponding to the starting context, S, and the words of the list. In a trace model the corresponding structure would be a list of vectors whose elements are the units of the associative structure.

(S' W1')
(W1' W2')
(W2' W3')
.
.
(Wn–1' Wn').

Now the starting context, S, being most similar to the first trace, would have the highest probability of reactivating it when presented, and therefore of producing recall of W1; the result of recalling W1 would similarly be most likely to activate the trace including W1', and so on. A similar analysis has been presented by Greeno et al (1978). If the associative scheme were interpreted as a particular application of the SAM model and the trace structure as a particular application of the exemplar-memory model, even the formulas for response probability would be identical in form, one entering sums and ratios of associative strengths and the other sums and ratios of similarities, both in the manner of Luce's (1963) choice model. The exemplar trace model may be viewed as a special case of the SAM model in which associative strengths are constrained to be values on a similarity dimension. The set of structural properties common to association and trace models I henceforth refer to, for brevity, as the *array* architecture.

Some psychological models that appear very different in form from those reviewed above may nonetheless be shown to share the same architecture. An important case is the class of geometric models of memory (Cunningham & Shepard 1974; Hutchinson & Lockhead 1977). These first appeared in the semantic memory literature just when hierarchical network models were approaching their peak of popularity. The geometric models did not receive wide attention, perhaps in part because their applications were largely limited to studies of similarity judgments in the tradition of psychophysical scaling. Following the scaling approach, Hutchinson & Lockhead (1977) showed that words could be assigned to positions in a multidimensional space, consistently across different methods of collecting similarity data, and that reaction times for discriminations between words were directly related to distances between them. Their results suggested that priming effects, a cornerstone of the network models, were readily interpretable in the metric framework.

The geometric model is useful in bringing out relationships between phenomena of semantic memory and psychophysics that would not otherwise be apparent, whether or not it exemplifies a unique architecture. Hutchinson & Lockhead offered, without a proof, the conjecture that it is actually isomorphic to feature and network models. It is not obvious that these classes of models are well enough defined to make a general proof possible, but Nosofsky (1984) has shown that the isomorphism does hold for models of categorization and identification formulated in the array framework. Building on earlier work of Shepard (1958), Nosofsky showed that with similarity between memory vectors in an array related to distance between corresponding points in a cognitive space by an exponential function, an exemplar-memory model is equivalent to a geometric model. The correspondence is not unique, for a particular array model can be mapped onto different geometric models based on different metrics, but it does seem clear that geometric models need not be considered to assume a different architecture from that of array models.

Summary of the Array Architecture

What is common to all of the models that I have subsumed under the array architecture? A central assumption is that memory of any experience with objects or events can be stored in the form of a multicomponent trace, the components being features or values on attributes or dimensions. (I use the latter two terms interchangeably.) These traces can be viewed as vectors or lists, as nodes in a network, or as points in a multidimensional space. When any one trace is activated, most commonly by occurrence of an appropriate percept (i.e. one with the same feature values or a subset of them), it may in turn activate others. The tendency for this activation to occur is interpretable as a similarity relation between traces, as strength of an associative link, or as

distance in a cognitive space. A measure of strength of activation in any one of these interpretations can be converted into any of the others by a simple transformation, so they are not empirically distinguishable. Again, a memory array can be partitioned into subarrays corresponding to categories; a network or cognitive space can be similarly divided into regions.

Nearly all cognitive theories now make a distinction between long-term memory and short-term, or working, memory. The former is conceived to be of indefinitely large capacity but slow access, the latter of sharply limited capacity and rapid access, and also subject to cognitive operations like comparison and counting. It does not seem that this distinction need be built into the architecture, however. In the array framework, for example, we need only assume that the temporal-spatial context of a learning experience is represented in the memory trace but decays in availability as a function of time following the experience.

COMPOSITE DISTRIBUTED MEMORY

In array models, an encoded memory trace is conceived to be stored in a distinct, content-addressable location and to preserve its identity, so that it is meaningful to speak of carrying out cognitive operations on individual traces. In contrast, distributed memory models, which began to appear in the early 1970s, are based on an architecture in which the individual identity of a trace is lost. Typically, a percept is encoded as a vector of feature values, as in an array model; however, the vector is not stored in a memory location, but is, rather, added into a composite trace that represents cumulative memory. In the model of Anderson (1973), successively encoded vectors are literally added (by matrix addition). Thus, if two successive learning experiences were encoded as $(1,-1,-1,1)$ and $(1,1,-1,-1)$, the composite memory for these experiences would be $(2,0,-2,0)$. If the vectors are orthogonal (roughly speaking, if their components are uncorrelated), the evocation of a percept corresponding to any one of the traces on a later test will revive that trace, providing a basis for recognition. Technically, a test vector, say $(1,1,-1,-1)$, is applied to the composite memory vector by a form of matrix multiplication, and the product, multiplied by a scalar if appropriate, is the same as the test vector. If the vectors are large, then this result may be obtained to a good approximation even if there are departures from orthogonality. Thus, in general, the result of a recognition test is not revival of a particular stored trace, but, rather, generation of a vector similar to one or more stored traces. A model with a very similar mathematical structure has been formulated and applied to recall as well as recognition by Pike (1984).

A distributed memory model that has been progressively elaborated and applied to many different types of situations is the theory of distributed

associative memory (TODAM) of Murdock and his associates (Murdock 1979, 1982; Murdock & Hockley 1989) and a closely related composite holographic associative recall model, CHARM (Eich 1982; Metcalfe & Fisher 1986). These models differ from those of Anderson (1973) and Pike (1984) in using some different mathematical operations. To enter new vectors into the composite memory, they employ *convolution*, a combination of addition and multiplication, and to obtain information from the composite memory, *correlation*, the inverse of convolution.

Two questions immediately come to mind regarding a proposed architecture requiring a mathematical formalism and methods so different from those familiar to psychologists in the tradition of trace and association theory: What suggested this novel architecture, and how is it faring in current memory theory? The prime answer to the first question appears to lie in the intuition of some investigators that a distributed, composite memory has properties much more congenial to what is known about the brain than those of traditional theories (Anderson 1973; Anderson et al 1977). Brain function is conspicuous for redundancy, tolerance of noise, and resistance to disturbance from localized damage—all properties that seem out of keeping with localized storage of discrete memory traces. Also, trace and association theories that had appeared prior to about 1970 had been so austerely simple in structure that they seemed to offer little prospect of helping to interrelate and integrate research in different paradigms, such as recognition, recall, and classification. In its continuing development, the family of composite memory models has shown enough promise in this respect to present an interesting alternative to the quite different lines of elaboration of classical association theory by Anderson & Bower (1973), Estes (1972a), and Raaijmakers & Shiffrin (1981).

The typically impatient psychologist is likely to ask at this point why, in the course of a decade or so of research, no one has carried out an experimental test to determine which type of model is superior. Such tests have been attempted, but none has yielded a decision, and I suspect that none is likely to. Still, it is possible to say something about evaluation. In the hands of their proponents, the composite models appear to be serving about as well as array models for instigating instructive experiments and bringing out theoretically interesting relationships among different empirical domains (Eich 1982; Metcalfe & Fisher 1986; Murdock & Lamon 1988; Murdock 1989). The latter function is illustrated by the fact that Murdock's convolution model has been shown to imply a formal relationship between recall and recognition (Murdock 1989; Murdock & Hockley 1989) comparable in parsimony and elegance to those that have been derived by Gillund & Shiffrin (1984), for example, for the SAM model and by Nosofsky (1986, 1988) for the exemplar-memory model.

THE CONNECTIONIST ARCHITECTURE

Perhaps the only reason why the composite models developed by Metcalfe-Eich and Murdock are not more widely influential than they currently appear to be is that they have been overshadowed somewhat by the sudden flowering of what are termed *connectionist,* or *parallel-distributed-processing* (PDP) models, a family sharing some architectural features with the convolution models but stemming from quite different origins. The basic elements of connectionist models are nodes and links, as in association models. The nodes are, however, simple, homogeneous units that do not, individually, correspond to external referents; their only properties are levels of activity and the capability of transmitting activation over the links between nodes. A connectionist network comprises two or more layers of nodes, typically a lowest level, in which nodes are activated by inputs from perceptual channels, a highest level, in which nodes receive activation from lower levels and in turn activate response mechanisms, and one or more intermediate levels of "hidden units," which receive input from lower levels and transmit it upward. Each link between nodes has an associated weight, which determines the strength of transmission over it, and memory resides wholly in the pattern of weights, which is modified ("updated") by the inputs to the network on each learning experience. The richness of the architecture is greatly amplified by the hidden units, which have no direct correspondence to input or output nodes but in the course of learning may come to act as feature detectors or classifiers.

Immediate precursors of the current wave of connectionist activity in cognitive science include PDP network models of visual processing (Ballard et al 1983; Marr 1982), mathematical investigations of feature detection and classification in adaptive networks (Grossberg 1976), and development of an "interactive activation" network model of processes in reading (McClelland & Rumelhart 1981). Some unity was brought to a heterogeneous collection of theoretical developments by the imagination and initiative of two leaders of the movement within psychology in assembling a two-volume handbook, including both summary and tutorial presentations of the formal concepts and methods of connectionism and a sampling of applications of PDP models to several cognitive domains (McClelland & Rumelhart 1986; Rumelhart & McClelland 1986a).

The special relevance of connectionist models to my long-term research interests and to this essay is that they have offered the promise of rescuing learning theory from its near eclipse during several decades in which cognitive research has been dominated by concern with problems of information-processing, representation in memory, and cognitive operations. Whether we

look at research and theory on short-term memory, semantic memory, propositional networks, or even the earlier distributed memory models, we see almost no efforts to interpret in any detail the processes whereby information comes to be stored in memory. The connectionist movement, in contrast, brings a new emphasis on learning. In traditional information-processing models, mechanisms like feature detectors are laid down as part of the architecture; items of information from input channels are deposited in memory buffers or stores to await retrieval; motives and goals may be important in the minds of the investigators, but they receive no explicit representation in the models. In connectionist models, feature detectors are generated from initially "blind and dumb" nodes and links; items of input information gain memorial representation by a process of weight adjustment; the capacity to recognize and classify perceptual patterns develops by means of a learning process that is driven by an overall tendency toward error correction. An "error," in this context, is a discrepancy between the current output ("response") of a network and a target, or goal, specified by a *teaching signal,* which may be either supplied by the environment or (a critically important property) internally generated. Thus, in the connectionist approach, learning is not viewed as a subsidiary problem to be left for consideration after the big problems are taken care of, but, rather, is a major aspect of the cognitive system from the start. For an investigator who grew up scientifically in the golden age of learning theory, this renewed emphasis on learning is a welcome development.

We have recently seen an abundance of discussions of general properties of learning in connectionist networks and similar architectures (Grossberg 1987; Hinton & Sejnowski 1986; Rumelhart & Zipser 1986; Rumelhart et al 1986; Stone 1986). As a complement to these, I wish to compare detailed properties of representatives of the network and array model families as developed and applied in a particular line of research, namely human category learning.

A COMPARISON OF MODELS FOR CATEGORY LEARNING

Category Learning in Array Models

Research on the learning of categories has followed two main strands. Historically prior was what Smith & Medin (1981) called the "classical approach," dealing with learning of categories that are sharply delimited by necessary and sufficient properties and definable in terms of simple verbal rules, like those of formal school subjects like geometry and grammar. In the information-processing approach dating from Hovland (1952), the standard experimental task is classifying multidimensional stimulus patterns into categories defined in terms of logical combinations of attributes, the prime research question

being how task difficulty is related to complexity of the logical rules (Bourne 1970; Hunt 1962; Shepard et al 1961). The body of theory generated in this tradition took the form of hypothesis-testing models (Hunt 1962; Trabasso & Bower 1968). Thus the main focus was on performance, perhaps the reason why work in this tradition dwindled as the center of interest for cognitive psychologists shifted during the 1970s from problems of performance to problems of representation in memory.

In the other principal strand of research on concept learning, research is addressed to acquisition and representation of what may be termed "fuzzy sets," that is, categories that do not have sharp boundaries and are best definable in terms of family resemblance or probability distributions. This approach was a center of attention for me and my associates from the mid 1950s because it provided an ideal research context for the testing and development of models based on stimulus sampling theory (Estes 1950). The favorite experimental paradigm of that period is very similar to some that are popular today, though the connection has generally been missed by writers of introductions to research articles, perhaps largely because the vocabulary used to describe it has changed. Studies that would now be characterized in terms of classification or categorization were reported in the earlier period under the label "discrimination learning," a term now almost wholly confined to the animal learning literature.

In a typical study of that period (Estes et al 1957), subjects viewed a display panel containing a row of 12 light bulbs. They were instructed, in effect, that on each of a series of trials some subset of the lights would be illuminated and that they should try to assign it to one of two alternative categories. Correct assignment would be indicated by a feedback signal. Different probability distributions defined over the display positions determined the samples of lights drawn for the two categories. The collections of samples that occurred on the two types of trials would be termed fuzzy sets in modern parlance. The stimulus sampling model that the study was intended to test predicted quite accurately the asymptotic proportions of correct responses for groups that learned with different category base rates as well as transfer performance on tests given at the end of the learning series with subsets of lights not previously seen.

The conditions of the 1957 study differed from those characteristic of related current work in that the populations of stimulus patterns associated with the categories were very large, so that individual patterns would rarely have recurred even during a learning series of several hundred trials. It occurred to me that the success of the stimulus sampling model might have been peculiar to that constraint, so I carried out several followups with similar designs but much smaller population sizes (and, in tune with the changing tenor of the times, a change from meaningless signal lights to symbols for

medical symptoms, or the like, as the component features of category examplars). The result was that the model in its original form broke down and could be brought into accord with the new data only with the added assumption that repeated experience with individual subsets of features led to the patterns' being encoded as units (Atkinson & Estes 1963; Estes 1972b; Estes & Hopkins 1961). It was apparent that the stimulus sampling model was not rich enough in structure to provide a satisfactory linkage between the two distinct versions needed to handle learning with large and small categories. It was not so apparent how to mend matters.

The next step forward needed a fresh viewpoint, and one was finally supplied by the extension of the discrimination model of Medin (1975) to the interpretation of human category learning (Medin & Schaffer 1978).

Although preserving some of the basic ideas of stimulus sampling theory, Medin & Schaffer's model presented a distinctly new look. The subject in a category learning experiment was assumed, not to associate individual cues with responses, but rather to store in memory on each learning trial a featural representation of the perceived exemplar together with its category tag. When asked to assign a pattern to a category, the learner was conceived to compare it to each of the stored representations, compute the similarity by the multiplicative rule, and then generate a choice probability for each category proportional to its summed similarity to the test exemplar. This model immediately aroused much interest because it not only yielded predictions of transfer effects under some novel experimental routines but also accounted for phenomena that had been taken to support prototype models, as, for example, the fact that a stimulus pattern corresponding to a category prototype presented for the first time on a test at the end of learning is likely to be correctly categorized with higher probability than patterns that have actually occurred during learning (Medin & Schaffer 1978). Numerous applications of various special cases of this exemplar-memory model (for which some investigators prefer the less mnemonic designation *context model*) during the next several years yielded consistently good accounts of asymptotic learning data and transfer to new patterns following learning, and tended to support the exemplar-memory model over prototype models (Busemeyer et al 1984; Estes 1986b; Nosofsky 1984, 1986). My own related work went further in demonstrating similarly good accounts of the detailed course of category learning over hundreds of trials and brought out the close parallelism between categorization and recognition that is implied by the exemplar-memory model (Estes 1986b).

When treading any primrose path, one is likely to run eventually into brambles, and this one proved no exception. It would be expected on theoretical grounds that such closely related processes as identification and classification of the same set of stimulus patterns should be predictable by an

adequate model without changes in paramater values from one task to the other, but a direct test by Nosofsky (1984) yielded an apparent negative result for the exemplar-memory model. I say "apparently" because, although the model provided good fits to both identification and categorization, it was at the cost of a drastic change from one situation to the other in the attentional weights associated with stimulus dimensions in Medin & Schaffer's formulation of the model (to reflect a direct relationship between selective attention and feature validity). This finding was not unanticipated by Nosofsky, and he accomplished a partial rescue of the model by showing that during categorization learning, the values of the attentional weights (estimated from performance data) moved systematically in the direction of the values that would be optimal for efficient categorization. Thus, it seemed that the exemplar-memory model plus an auxiliary model for attentional learning might be able to account for both identification and categorization. Since no such auxiliary model has been formulated, the issue remains open. However, it will be seen in the next section that a more elegant solution to the problem than grafting mechanisms onto the exemplar-memory model may be forthcoming.

In my own studies related to the exemplar-memory model, I have employed only a special case that does not include parameters for dimensional weights, in order to simplify the problem of testing hypotheses about storage and retrieval processes. This version fared well enough at accounting for the details of learning in situations where selective attention would not be expected to play a significant role (Estes 1986a,b); but, even with this qualification met, the model has run into difficulties when called on to predict across a change in conditions. In a recent study, for example, I set out to test a particular aspect of the model having to do with hysteresis. Subjects were given the task of learning to classify artificial words into grammatical categories, the set of stimulus patterns having different probability distributions in each of three categories. For one category, conditions were constant throughout a 240-trial learning series, but the probability distributions for the other two categories were switched after trial 60 for an early-shift group and after trial 180 for a late-shift group. According to the exemplar-memory model, many more patterns, with their category tags, would be stored in the memory array by the point of the shift for the late-shift than for the early shift group; and therefore performance would be impaired after the shift until enough patterns could be correctly stored to outweigh the ones now incorrectly stored in the two shifted categories. The quantitative predictions of the model were exactly as advertised, but the predicted post-shift impairment for the late-shift group was much greater than that observed (Estes 1989).

Again, the setback for the model is not fatal, of course. In further analyses (not yet published), I have found that, as one might expect, the model can be brought into line with the shift data by adding the assumption of a decay-like

process whereby a stored pattern declines in availability as an exponential function of trials following storage, so that, in effect, the retrieval and similarity-comparisons of the model only operated on a limited set of recently stored patterns. This assumption seems a quite natural one, and I now regard it as part of the "standard equipment" of an updated exemplar-memory model. Nonetheless, as in the case of predicting from identification to categorization, would it not be fine if the model under test proved able to handle the new results without requiring elaboration? The continuing elusiveness of that goal for models of the array family prepared me to be immediately much interested in the potentialities of connectionist models when they were introduced into categorization research.

Category Learning in PDP Network Models

A stripped-down PDP model, based on a connectionist architecture but lacking hidden units, was developed and applied to categorization learning by Gluck & Bower (1988b). Their model, denoted an "adaptive network," includes an input node for each member of the feature set used to generate category exemplars in an experiment, an output node for each category, and a link, with an associated weight, from each input to each output node. The probability of a categorization response is based on the summed output of the system to each of the category nodes in response to an input pattern. It is assumed that each learning trial comprises presentation of a stimulus pattern, which activates a set of feature nodes, computation of the system's categorization response, and presentation of a feedback, or "teaching," signal indicating the correct category. Learning is accomplished by a set of functions that update the associative weights for each node active on a trial in such a way that the weights move toward the target value specified by the teaching signal. The functions are similar to the learning functions of stimulus sampling models except that they embody a competitive property in that the increment to a weight on any trial is reduced to the degree that other active nodes already predict the correct category. Formally, the increment, or decrement, is proportional to the difference between the current output of the network for the given input pattern and the target output, a property deriving from the conditioning model of Rescorla & Wagner (1972) and the "delta rule" familiar in adaptive network theory (Stone 1986; Widrow & Hoff 1960).

Gluck & Bower noted that even a very simple categorization problem of suitable design could yield an interesting test of differing predictions by exemplar ("pattern-matching," in their terms) models and their network model. In their study (Gluck & Bower 1988b, Exp 1), subjects were given a task simulating that of a diagnostician. Stimulus patterns were generated from a set of four features, labelled as medical symptoms, and subjects were to assign each pattern presented (interpreted as the symptom chart of a hypothetical

patient) to one of two disease categories. The two categories, A and B occurred with probabilities .25 and .75, respectively; on each type of trial, the features occurred with the probabilities shown in Table 2. Prime interest is in the subjects' response to Feature 1 when it was tested alone at the end of the learning series and they were asked to estimate the probability of either category in its presence. It will be seen that, owing to the unequal category probabilities, Feature 1 will be expected to occur equally often in both categories over the learning series. Consequently, the prediction of the exemplar-memory model (or a stimulus sampling model) is that the subjects' probability estimates for categories A and B should each be equal to .5 on the test. In contrast, the network model predicts a large bias for Category A—the result observed in Gluck & Bower's study. Shortly after this demonstration, Estes et al (1989) replicated this finding and went on to show that the network model yielded an account of the detailed course of learning considerably superior to that of the exemplar-memory model.

A major limitation of the simple network model is that it can only learn categorizations for situations in which the features of category exemplars combine independently—that is, the features are uncorrelated within categories. Thus the network could not, for example, exhibit learning in Experiment 2 of Estes (1986b), where all members of the feature set were invalid (that is, occurred equally often in each category) but some pairs of cues were partially valid (that is, occurred with different frequencies in the two categories)—although both the subjects and the exemplar-memory model did exhibit significant learning.

The minimal elaboration of the network needed to get around this difficulty is to allow for three levels of nodes, the first level representing individual features and the second level patterns of features, with both levels being connected to the category nodes at the third level (Estes 1988). In a realization of this elaborated network that might be termed the feature/pattern (F/P) model, learning of a new categorization begins with the model having only the simple network structure, but as each presented exemplar activates its feature nodes, a pattern node is added to the network. Henceforth, the pattern

Table 2 Feature probabilities on trials assigned to each category (Gluck & Bower 1988b)

	Category	
Feature	A	B
1	.6	.2
2	.4	.3
3	.3	.4
4	.2	.6

node is activated whenever its set of feature nodes is active, thus acting as an AND gate in network parlance, and it is linked to nodes at the category level just as are the feature nodes. The system is still linear, since the outputs are assumed to be summed activations just as in the simple model, but the additional structure enables the network to learn problems like that of Estes (1986b, Exp. 2). In fact, on the data of that experiment, the F/P model (with one free parameter added to allow for different learning rates on features and patterns) closely matches the performance of the neo–exemplar memory model, and the same has been true for several sets of unpublished data. An important property of the F/P model is that, in a categorization task, the network can concurrently learn quite different things about exemplars and their features (for example, that presence of a particular pattern points strongly to one category whereas each of its features points to other categories). Consequently, the network model may prove significantly more powerful than exemplar-memory models for the prediction of transfer phenomena.

An experiment well designed to tax the capabilities of both the exemplar-memory and F/P models was conducted by J. B. Hurwitz as part of his Harvard dissertation study. The task for his subjects was learning to classify a set of strings of four binary features into two categories A or B. The category structure was a bit complex. Two of the string positions were filled with an exclusive-OR problem (denoted XOR1); the features in those positions were the same (11 or 22 in binary coding) on Category A and different (12 or 21) on Category B trials throughout learning. The other two positions were filled with a different problem, XOR2, for which the contingencies were reversed for some strings after the second 60-trial training block. By "reversed," I mean that if feature pairs 11 and 22 originally occurred in these positions only on A trials, after reversal they occurred only on B trials. Programming of exemplar and category occurrences was such that optimally efficient learners who attended only to the XOR1 letter positions would move quickly to 100% correct responding and remain at that level throughout the series; if they attended only to the XOR2 positions, they would approach 100% correct during the first two blocks but would revert to chance responding in the third block. This scheme sounds complex, but those of Hurwitz's subjects who learned at all moved steadily from chance to a level near 100% correct responding with very little disturbance from the reversal of the XOR2 problem. The picture was as though the subjects were able to do quite well at screening out the XOR2 letter positions and attending only to the XOR1 positions. Exemplar-memory models have no mechanism for producing this apparent selective attention, so it is not surprising that the neo–exemplar memory model gave only a poor account of the data, predicting much too large a drop in performance at the point of partial reversal. The network model, even in the F/P version, did little better [and the same is true of the "configural-cue" model of Gluck & Bower (1988a)].

The reason for special interest in Hurwitz's result is not just that it is the first massive failure recorded in this line of research for the models considered, but that Hurwitz went on to show that the data could be well handled by an elaboration of the F/P network model that included learning on the weights from input to pattern nodes by back-propagation of error signals from the output nodes (Rumelhart et al 1986). Thus, at this stage of the current wave of research on category learning, an adaptive network model with an architecture closer to that of typical connectionist models has proved able to cope with a novel and complex learning regime in a way well beyond the capabilities of the array models or simpler, linear networks that have been applied in this line of research to date.

PROBLEMS FOR CONNECTIONIST MODELS

Despite this and other successes, all is not smooth going for connectionist models. A disturbing note is sounded by reports that as soon as they are extended beyond the task of accounting for the concurrent learning of a set of materials, as a single categorization task or a single to-be-memorized list, they may prove unable to maintain several clusters of successively acquired memories simultaneously and therefore exhibit massive interference effects on recognition or recall tests (McCloskey & Cohen 1989; Ratcliff 1990). In the study of McCloskey & Cohen, for example, simulations of classical paired-associate list learning experiments like that of Barnes & Underwood (1959) yielded interference effects in recall much larger than those shown by human subjects, and this interference was not alleviated by any of the obvious remedial tactics, such as varying the number of hidden units, giving overtraining on the first of two successive lists, or including representation of list contexts. Further, the ambitious, and in many respects impressive, effort by Rumelhart & McClelland (1986b) to produce a connectionist model of the way children acquire certain linguistic competences has run into heavy criticism (Prince & Pinker 1988), including a claim that connectionist networks cannot, in principle, learn rules of the kind that are basic to language. Thus we have a curious situation: Connectionist models are built to learn, but there are reasons to question whether they can be made to learn like human beings.

Of course, these failures of connectionist networks have attracted wide attention, and there have already been reports of some results that limit somewhat the generality of the massive interference findings. Hetherington & Seidenberg (1989) have shown that the extremely large interference effects manifest when lists of items are learned successively are mitigated to some degree under a learning routine intended to be closer to that of vocabulary learning by children in a natural environment. Following a somewhat different approach, Brousse & Smolensky (1989) have found that once a hidden-

unit network has learned a subset of items from a large combinatorial domain, as, for example, strings of six letters, additional items can be learned rapidly and the increasingly large memory can be maintained without interference.

Returning to the more restricted successive list paradigm, which remains at the least an annoying pebble in the connectionist shoe, I can add that I have run my F/P network model on tasks very similar to those studied by McCloskey & Cohen (1989), including the Barnes & Underwood (1959) experiment, and have found interference only of the level seen in human subject data. There are too many differences between the specific models simulated to make interpretation of the different results feasible until the analyses are carried further. My model is a linear system, which should not be a critical difference in itself; and it has more a priori structure, in that the nodes at the second level are mapped onto particular input patterns, rather than having stimulus patterns correspond to patterns of activity across a layer of undifferentiated nodes. A stray thought that comes to mind is that the situation is a bit reminiscent of a segment of the history of learning theory. In the learning theories of the period 1930–1960, it was assumed that learning processes are basically the same at least for all of the higher animals and that learning in the individual organism starts from a *tabula rasa,* the counterpart of a network of homogeneous and mutually interconnected nodes. During the next decade, however, under the impact of ethology and the beginnings of modern neuroscience, the prevalent view shifted to one that recognized biological constraints on learning (Estes 1988; Hinde 1973; Shettleworth 1972); and it is now quite generally assumed that learning in any organism, human or subhuman, builds on a substrate of species-specific predispositions and products of previous learning. Implementing this more biologically founded orientation in connectionist learning models is a tall order; but the effort will surely have to be made sooner or later, and the results may cast some of the current problems in a new light.

REFLECTIONS

Is memory distributed? I find it hard to doubt that at the neural level the answer is yes, at least within the components of modular structures. The implication for constructors of cognitive models is not obvious, however, for models that differ only in their assumptions about distributed versus localized storage may not be differentiable at the behavioral level. The latter comment is relevant also to questions about composite memories. Convolution models and associative network models appear very different when diagrammed or described in words, but it is possible that one type can be mapped onto the

other. For example, following a given learning experience, the set of informational patterns recoverable from a composite memory by members of a set of recall cues might correspond one-to-one to the set of images stored at the nodes of an associative network. Much theoretical work is needed to assess the possibilities of such equivalences.

Are there built-in memory structures? The notion of a cognitive architecture may seem to connote a set of permanent memory structures, waiting to receive information and constraining the form of storage, as in the models of Anderson & Bower (1973), Atkinson & Shiffrin (1968), or Norman & Rumelhart (1970). But, alternatively, it may be assumed that the cognitive system has only very general built-in capabilities and creates memory structures on-line in response to task demands (Newell 1973). There may well be a peripheral-central gradient, with the more peripheral, or modality-specific, subsystems, like primary visual and auditory memory, tending to have more fixed structures. Examples of reasonably direct evidence for on-line formation of memory structures tailored to particular task demands can be found in research on ordered recall (Lee & Estes 1981) and memory for dates of events (Huttenlocher et al 1988).

Architecture or architectures? I think pluralism gets the nod, certainly for the present and quite possibly for a long way into the future. The complexity of the human cognitive system demands approaches from differing perspectives, and these must be expected to give rise to successions of limited models and restricted architectures. The idea that the connectionist and the symbol-processing architecture might fit neatly into the different levels of an overall theory (Smolensky 1988) is attractive, but in my view not ready for evaluation.

Parallel cultivation of symbol-processing and connectionist architectures need not imply anything like equality of effort or rate of progress, however. The former seems to be lagging in new theoretical development, but remains influential because concepts of symbol processing and array representation fit the intuitions of the majority of investigators in cognitive science. The connectionist approach has the enormous advantage of resonating more strongly with the current groundswell of interest in cognitive neuroscience, and brings a variety of new concepts, metaphors, and formal tools into cognitive theory. Intuitions can change, and they may have to do so if the connectionist movement, broadly conceived, continues to gain momentum at the rate that now seems likely from my perspective.

ACKNOWLEDGMENT

Preparation of this chapter was supported in part by Grant BNS 86-09232 from the National Science Foundation.

Literature Cited

Anderson, J. A. 1973. A theory for the recognition of items from short memorized lists. *Psychol. Rev.* 80:417–38

Anderson, J. R. 1976. *Language, Memory, and Thought.* Hillsdale, NJ: Erlbaum

Anderson, J. R. 1983. *The Architecture of Cognition.* Cambridge, MA: Harvard Univ. Press

Anderson, J. R., Bower, G. H. 1973. *Human Associative Memory.* Washington, DC: Winston

Anderson, J. A., Silverstein, J. W., Ritz, S. A., Jones, R. S. 1977. Distinctive features, categorical perceptions, and probability learning: some applications of a neural model. *Psychol. Rev.* 84:413–51

Atkinson, R. C., Estes, W. K. 1963. Stimulus sampling theory. In *Handbook of Mathematical Psychology,* ed. R. D. Luce, R. R. Bush, E. Galanter 2:121–268. New York: Wiley

Atkinson, R. C., Shiffrin, R. M. 1968. Human memory: a proposed system and its control processes. In *The Psychology of Learning and Motivation: Advances in Research and Theory,* ed. K. W. Spence, J. T. Spence, pp. 89–105. New York: Academic

Ballard, D. H., Hinton, G. E., Sejnowski, T. J. 1983. Parallel visual computation. *Nature* 306:21–26

Barnes, J. M., Underwood, B. J. 1959. "Fate" of first-list associations in transfer theory. *J. Exp. Psychol.* 58:97–105

Bourne, L. E. Jr. 1970. Knowing and using concepts. *Psychol. Rev.* 77:546–56

Bousfield, W. A. 1953. The occurrence of clustering in the recall of randomly arranged associates. *J. Gen. Psychol.* 49:229–40

Bower, G. H. 1967. A multicomponent theory of the memory trace. In *The Psychology of Learning and Motivation: Advances in Research and Theory,* ed. K. W. Spence, J. T. Spence, pp. 230–327. New York: Academic

Brousse, O., Smolensky, P. 1989. *Virtual memories and massive generalization in connectionist combinatorial learning.* Presented at Annu. Meet. Cognit. Sci. Soc., 11th, Ann Arbor

Busemeyer, J. R., Dewey, G. I., Medin, D. L. 1984. Evaluation of exemplar-based generalization and the abstraction of categorical information. *J. Exp. Psychol.: Learn. Mem. Cogn.* 10:638–48

Clark, S. E., Shiffrin, R. M. 1987. Recognition of multiple item probes. *Mem. Cog.* 15:367–78

Collins, A. M., Loftus, E. F. 1975. A spreading activation theory of semantic processing. *Psychol. Rev.* 82:407–28

Collins, A. M., Quillian, M. R. 1972. How to make a language user. In *Organization of Memory,* ed. E. Tulving, W. Donaldson, pp. 310–51. New York: Academic

Cunningham, J. P., Shepard, R. N. 1974. Monotone mapping of similarities into a general metric space. *J. Math. Psychol.* 11:335–65

Ebbinghaus, H. 1964. [1885]. *Memory.* Ed. transl. H. A. Ruger, C. E. Bussenius. New York: Dover

Eich, J. M. 1982. A composite holographic associative recall model. *Psychol. Rev.* 89:627–61

Estes, W. K. 1950. Toward a statistical theory of learning. *Psychol. Rev.* 57:94–107

Estes, W. K. 1955. Statistical theory of spontaneous recovery and regression. *Psychol. Rev.* 62:145–54

Estes, W. K. 1969. New perspectives on some old issues in association theory. In *Fundamental Issues in Associative Learning.* ed. N. J. Macintosh, W. K. Honig, pp. 162–89. Halifax: Dalhousie Univ. Press

Estes, W. K. 1972a. An associative basis for coding and organization in memory. In *Coding Processes in Human Memory,* ed. A. W. Melton, E. Martin, pp. 161–90. Washington, DC: Winston

Estes, W. K. 1972b. Elements and patterns in diagnostic discrimination learning. *Trans. NY Acad. Sci.* 34:84–95

Estes, W. K. 1975a. Some targets for mathematical psychology. *J. Math. Psychol.* 12:263–82

Estes, W. K. 1975b. The state of the field: general problems and issues of theory and metatheory. In *Handbook of Learning and Cognitive Processes.* Volume 1. *Introduction to Concepts and Issues,* ed. W. K. Estes, pp. 1–24. Hillsdale, NJ: Erlbaum

Estes, W. K. 1986a. Array models for category learning. *Cogn. Psychol.* 18:500–49

Estes, W. K. 1986b. Storage and retrieval processes in category learning. *J. Exp. Psychol.: Gen.* 115:155–74

Estes, W. K. 1988. Toward a framework for combining connectionist and symbol-processing models. *J. Mem. Lang.* 27:196–212

Estes, W. K. 1989. Early and late memory processing in models for category learning. In *Current Issues in Cognitive Processes: The Tulane Symposium on Cognition,* ed. C. Izawa, pp. 11–24. Hillsdale, NJ: Erlbaum

Estes, W. K., Burke, C. J., Atkinson, R. C., Frankmann, J. P. 1957. Probabilistic discrimination learning. *J. Exp. Psychol.* 54:233–39

Estes, W. K., Campbell, J. A., Hatsopoulos, N., Hurwitz, J. B. 1989. Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models. *J. Exp. Psychol.: Learn. Mem. Cognit.* 15:556–71

Estes, W. K., Hopkins, B. L. 1961. Acquisition and transfer in pattern-vs.-component discrimination learning. *J. Exp. Psychol.* 61:322–28

Gillund, G., Shiffrin, R. M. 1984. A retrieval model for both recognition and recall. *Psychol. Rev.* 91:1–67

Gluck, M. A., Bower, G. H. 1988a. Evaluating an adaptive network model for human learning. *J. Mem. Lang.* 27:166–95

Gluck, M. A., Bower, G. H. 1988b. From conditioning to category learning: an adaptive network model. *J. Exp. Psychol.: Gen.* 117:225–44

Greeno, J. G., James, C. T., DaPolito, F., Polson, P. G. 1978. *Associative Learning: A Cognitive Analysis*. Englewood Cliffs, NJ: Prentice-Hall

Gronlund, S. D., Shiffrin, R. M. 1986. Retrieval strategies in recall of natural categories and categorized lists. *J. Exp. Psychol.: Learn. Mem. Cognit.* 12:550–56

Grossberg, S. 1976. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 23:121–34

Grossberg, S. 1987. Competitive learning: from interactive activation resonance. *Cognit. Sci.* 11:23–63

Hetherington, P. A., Seidenberg, M. S. 1989. *Is there "catastrophic interference" in connectionist networks?* Presented at Annu. Meet. Cognit. Sci. Soc., 11th, Ann Arbor

Hinde, R. A. 1973. Constraints on learning: an introduction to the problems. In *Constraints on Learning*, ed. R. A. Hinde, J. Hinde, pp. 1–19. New York: Academic

Hinton, G. E., Sejnowski, T. J. 1986. Learning and relearning in Boltzmann machines. See Reumelhart & McClelland 1986, pp. 282–317

Hintzman, D. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychol. Rev.* 93:411–28

Hollingworth, H. L. 1913. Characteristic differences between recall and recognition. *Am. J. Psychol.* 24:532–44

Hollingworth, H. L. 1928. General laws of redintegration. *J. Gen. Psychol.* 1:79–90

Hovland, C. I. 1952. A "communication analysis" of concept learning. *Psychol. Rev.* 59:461–72

Hunt, E. B. 1962. *Concept Learning: An Information Processing Problem*. New York: Wiley

Hutchinson, J. W., Lockhead, G. R. 1977. Similarity as distance: a structural principle for semantic memory. *J. Exp. Psychol.: Hum. Learn. Mem.* 3:660–78

Huttenlocher, J., Hedges, L., Prohaska, V. 1988. Hierarchical organization in ordered domains: estimating dates of events. *Psychol. Rev.* 95:471–84

Koffka, K. 1935. *Principles of Gestalt Psychology*. New York: Harcourt Brace

Laird, J. E., Newell, A., Rosenbloom, P. S. 1987. SOAR: An architecture for general intelligence. *Artif. Intell.* 33:1–64

Lee, C. L., Estes, W. K. 1981. Item and order information in short-term memory: evidence for multilevel perturbation processes. *J. Exp. Psychol.* 7:149–69

Luce, R. D. 1963. Detection and recognition. In *Handbook of Mathematical Psychology*, ed. R. D. Luce, R. R. Bush, E. Galanter, 1:103–90. New York: Wiley

Mandler, G. 1967. Organization and memory. In *The Psychology of Learning and Motivation: Advances in Research and Theory*, ed. K. W. Spence, J. T. Spence, pp. 328–72. New York: Academic

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman

McClelland, J. L. Rumelhart, D. E. 1981. An interactive interaction model of context effects in letter perception. *Psychol. Rev.* 88:375–407

McClelland, J. L., Rumelhart, D. E. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2. Cambridge, MA: MIT Press

McCloskey, M., Cohen, N. J. 1989. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.: Adv. Res. Theory* 24:109–65

Medin, D. L. 1975. A theory of context in discrimination learning. In *The Psychology of Learning and Motivation*, ed. G. H. Bower, 9:263–314. New York: Academic

Medin, D. L., Schaffer, M. M. 1978. Context theory of classification learning. *Psychol. Rev.* 85:207–38

Metcalfe, J., Fisher, R. P. 1986. The relation between recognition memory and classification learning. *J. Exp. Psychol.: Learn. Mem. Cogn.* 14:164–73

Murdock, B. B. Jr. 1974. *Human Memory: Theory and Data*. Potomac, MD: Erlbaum

Murdock, B. B. Jr. 1979. Convolution and correlation in perception and memory. In *Perspectives on Memory Research*, ed. L.-G. Nilsson, pp. 105–19. Hillsdale, NJ: Erlbaum

Murdock, B. B. Jr. 1982. A theory for the storage and retrieval of item and associative information. *Psychol. Rev.* 89:609–26

Murdock, B. B. Jr. 1989. Learning in a distributed memory model. See Izawa, pp. 70–106

Murdock, B. B. Jr., Hockley, W. E. 1989. Short-term memory for associations. *Psychol. Learn. Motiv. Res. Theory* 24:71–108

Murdock, B. B. Jr., Lamon, M. 1988. The replacement effect: repeating some items while replacing others. *Mem. Cogn.* 16:91–101

Newell, A. 1973. Production systems: models of control structures. In *Visual Information Processing*, ed. W. G. Chase, pp., 463–526. New York: Academic

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard Univ. Press. In press

Newell, A., Simon, H. A. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall

Norman, D. A., Rumelhart, D. E. 1970. A system for perception and memory. In *Models of Human Memory*, ed. D. A. Norman, pp. 21–64. New York: Academic

Nosofsky, R. M. 1984. Choice, similarity, and the context theory of classification. *J. Exp. Psychol.: Learn, Mem. Cognit.* 10: 104–14

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol.: Gen.* 115:39–57

Nosofsky, R. M. 1988. Similarity, frequency, and category representations. *J. Exp. Psychol.: Learn. Mem. Cognit.* 14:54–65

Pike, R. 1984. A comparison of convolution and matrix distributed memory systems. *Psychol. Rev.* 91:281–94

Prince, A., Pinker, S. 1988. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–194

Raaijmakers, J. G. W., Shiffrin, R. M. 1981. Search of associative memory. *Psychol. Rev.* 88:93–134

Ratcliff, R. 1978. A theory of memory retrieval. *Psychol. Rev.* 85:59–108

Ratcliff, R. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol. Rev.* 98:285:308

Rescorla, R. A., Wagner, A. R. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In *Classical Conditioning II: Current Research and Theory*, ed. A. H. Black, W. F. Prokasy, pp. New York: Appleton-Century-Crofts

Robinson, E. S. 1932. *Association Theory Today*. New York: Century

Rumelhart, D. E., Hinton, G. E., Williams, R. J. 1986. Learning internal representations by error propagation. See McClelland & Rumelhart 1986, pp. 318–62

Rumelhart, D. E., McClelland, J. L., eds. 1986a. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press/Bradford Books

Rumelhart, D. E., McClelland, J. L. 1986b. On learning the past tense of English verbs. See McClelland & Rumelhart 1986, pp. 216–71

Rumelhart, D. E., Zipser, D. 1986. Feature discovery by competitive learning. See Rumelhart & McClelland 1986b, pp. 151–93

Schneider, W., Shiffrin, R. M. 1977. Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychol. Rev.* 84:1–66

Semon, R. 1921. *The Mneme*. London: George Allen & Unwin

Shepard, R. N. 1958. Stimulus and response generalization: deduction of the generalization gradient from a trace model. *Psychol. Rev.* 65:242–56

Shepard, R. N., Hovland, C. I., Jenkins, H. M. 1961. Learning and memorization of classifications. *Psychol. Monogr.* 75:1–41

Shettleworth, S. J. 1972. Constraints on learning. *Adv. Study Behav.*, 4:1–68

Shiffrin, R. M., Murnane, K., Gronlund, S., Roth, M. 1989. On units of storage and retrieval. In *Current Issues in Cognitive Processes: The Tulane Floweree Symposium on Cognition*, ed. C. Izawa, pp. 25–68. Hillsdale, NJ: Erlbaum

Smith, E. E., Medin, D. L. 1981. *Categories and concepts*. Cambridge, MA: Harvard Univ. Press

Smolensky, P. 1988. On the proper treatment of connectionism. *Behav. Brain Sci.* 11:1–59

Stone, G. O. 1986. An analysis of the delta rule and the learning of statistical associations. See Rumelhart & McClelland 1986b, pp. 423–43

Trabasso, T., Bower, G. H. 1968. *Attention in Learning: Theory and Research*. New York: Wiley

Underwood, B. J. 1969. Attributes of memory. *Psychol. Rev.* 76:559–73

Widrow, B., Hoff, M. E. 1960. Adaptive switching circuits. *Inst. Radio Eng., West. Electron. Show & Convent. Rec.* Pt. IV, pp. 96–104