Parallel Distributed Processing

Implications for Psychology and Neurobiology

Edited by
R. G. M. MORRIS
University of Edinburgh

CLARENDON PRESS · OXFORD
1989

2

Parallel distributed processing: implications for cognition and development[1]

JAMES L. McCLELLAND

Introduction

What kind of processing mechanism is the mind? Is it a sequential information processing machine, like the von Neumann computer? Or is it a massively parallel processor? The fact that human thought takes place in a device consisting of some tens of billions of neurons seems to support the parallel view. Yet until recently, there has been little attention to this fact among those who study the higher mental processes, and little convergence in the study of mind and brain.

Feldman (1981) has pointed out that the human brain places constraints on the methods that might be used to implement human thought. Neurons are relatively sluggish, noisy processing devices, compared to today's computers. Yet people can perceive a visual scene at a glance and recognize an object in about half a second. Feldman estimates that this leaves time for perhaps a hundred processing steps; but sequential algorithms for perception generally require hundreds of thousands. The facts imply that we exploit the brain's obvious capacity for parallel processing.

In view of this, researchers have begun to work toward theories of mental processes that rely on these parallel capabilities. These models are variously known as parallel distributed processing (PDP) models, neural models, or, perhaps most generally, connectionist models. Work is proceeding in several directions. Cognitive scientists seeking to provide a characterization of the nature of human thought have turned to building computational models in which a number of interconnected processors work in concert in performing some information processing task. Meanwhile, neuroscientists seeking to understand the functional properties of neural circuits are also building computational models, exploring the collective properties of ensembles of neuron-like processing units. These enterprises often have somewhat different goals; yet each informs and enriches the other, and each is pursued with the hope that someday these two directions of research will converge upon a shared understanding of brain and mind.

In this chapter, I take primarily a congitive perspective. First, I review the

connectionist framework, stressing how basic aspects of cognition – representation, processing, knowledge, and learning – are captured in the connectionist framework. Next, the framework is applied to fundamental questions about the development of human thought, and some of the implications that the framework has for basic questions about cognition and development are illustrated.

The connectionist framework

The term 'connectionist models' was introduced by Feldman (1981; Feldman and Ballard 1982). In these papers, the term was used to refer to a class of models that compute by way of connections among simple processing units. Another phrase often used to describe some connectionist models is *parallel distributed processing* or PDP models (Rumelhart, McClelland, and the PDP Research Group 1986; McClelland, Rumelhart, and the PDP Research Group 1986). PDP models are instances of connectionist models that stress the notion that processing activity results from the processing interactions occurring among rather large numbers of processing units.

In this article I intend the phrase 'the connectionist framework' to encompass all kinds of connectionist models. The framework may be thought of as providing a set of general assumptions about basic aspects of information processing, and a set of soft constraints on the range of specific assumptions that might be made. In what follows I consider each of several aspects of an information processing system. I describe the general assumptions connectionist models make about these aspects and I characterize some of the specific assumptions that might be made. The presentation draws heavily on Rumelhart, Hinton, and McClelland (1986), which can be consulted for further details.

Primitives and their organization

Like all cognitive models, connectionist models must propose some building blocks and some organization of these building blocks. In connectionist models, the primitives are *units* and *connections*. Units are simple processing devices which take on activation values based on a weighted sum of their inputs from the environment and from other units. Connections provide the medium whereby the units interact with each other; they are weighted, and the weights may be positive or negative, so that a particular input will tend to excite or inhibit the unit that receives it, depending on the sign of the weight (we shall return to these matters when we consider the dynamics of processing below).

Any particular connectionist model will make assumptions about the

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

number of units, their pattern of connectivity to other units, and their interactions with the environment. These assumptions define the architecture of a connectionist model. The set of units and their connections is typically called a *network*.

It should be noted that a very wide variety of architectures is possible. Two are shown in Figs 2.1 and 2.2. One of these, in Fig. 2.1 (from the distributed

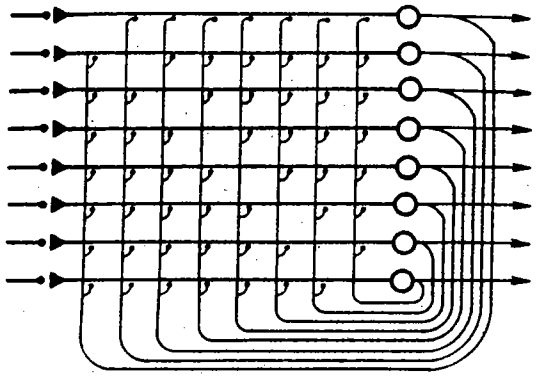


Fig. 2.1 A fully connected autoassociator network, with connections from each unit to every other unit. Each unit receives input from outside the network, and sends output outside the network. All connections may be either positive or negative in this very general formulation. (From J. L. McClelland and D. E. Rumelhart, 1985, 'Distributed memory and the representation of general and specific information', *Journal of Experimental Psychology: General*, **114**, 162. Copyright 1985 by the American Psychological Association. Reprinted by Permission.)

model of memory examined by McClelland and Rumelhart 1985) shows a set of completely interconnected units, each receiving input from the environment, and each projecting back to the environment. In some sense, the network in this diagram is the most general possible connectionist architecture, in that all others involve restrictions of this general case. For example, some units may receive no input from the environment; some may send no output outside the net; and some of the interconnections among units in the network may be deleted. There may, furthermore, be restrictions on the values of some of the connections. In the general case, each may be positive or negative, but the architecture may prescribe, for example, that a certain group of units have mutually inhibitory connections of fixed strength.

Figure 2.2 gives an example of a more restricted architecture, from the interactive activation model of visual word recognition (McClelland and Rumelhart 1981). In this model, units stand for hypotheses about displays of letter strings at each of three levels of description: a feature level, a letter level, and a word level. There are excitatory connections (in both directions)

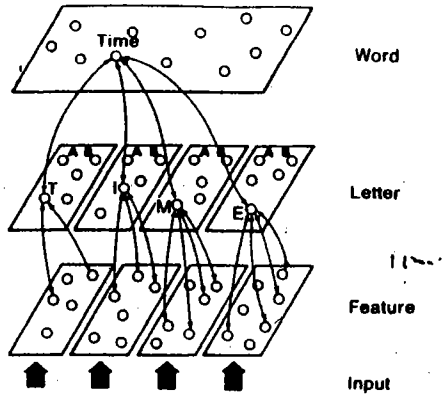


Fig. 2.2 A sketch of the network used in the interactive activation model of visual word recognition (McClelland and Rumelhart 1981). Units within the same rectangle stand for incompatible alternative hypotheses about an input pattern, and are mutually inhibitory. Bi-directional excitatory connections between levels are indicated for one word and its constituents. (From J. L. McClelland, 1985, 'Putting knowledge in its place: a scheme for programming parallel processing structures on the fly', *Cognitive Science*, **9**, 115. Copyright 1985 by Ablex Publishing. Reprinted by permission.)

between mutually consistent units on adjacent levels, and inhibitory connections between mutually inconsistent units within the same level. Thus the unit for T in the first letter position excites and is excited by the units for features of the letter T, as well as the units for words that begin with T. This unit also inhibits, and is inhibited by, units for other letters in the same letter position.

Active representation

Representations in connectionist models are patterns of activation over the units in the network. In some ways, these kinds of patterns are similar to representations in other frameworks; after all, representations in a computer are ultimately patterns of 0s and 1s. There are differences, however. For one thing it is quite natural for connectionist representations to be graded, in the sense that each unit's activation need not be one of two binary values. In some models, activations are restricted to binary or some other number of discrete values, but more typically each unit may take on a continuous activation value between some maximum and minimum. A more important difference is this: connectionist representations are truly active, in the sense that *they give rise to further processing activity directly*, without any need for a central processor or production-matching-and-application mechanism that examines them and takes action on the basis of the results of this examination.

Models differ in terms of the extent to which individual processing units can be identified with particular conceptual objects, such as letters, words, concepts, etc. The models illustrated in Figs 2.1 and 2.2 represent endpoints on a continuum. In the distributed model of memory, each conceptual object is thought of as a pattern of activation over a number of simple processing units. In the interactive activation model of word perception, on the other hand, each unit stands for a primitive conceptual object, such as a letter, a word, or a distinct visual feature. A large number of models lie between these two extremes (see Hinton, McClelland, and Rumelhart 1986, and Feldman 1988, for general discussions of the issue of distributed representation).

Processing

Processing in connectionist models occurs through the evolution of patterns of activation over time. This process is governed by assumptions about the exact way in which the activations of units are updated, as a function of their inputs. Updating can be synchronous (all units updated simultaneously) or asynchronous (units updated in random order). Updating generally occurs as follows. First, a *net input* is computed for each unit to be updated. The net input is the sum of the activations of all of the units that project to it, with each contributing activation weighted by the weight on the connection from the contributing unit to the receiving unit.[2] The net input may also include a *bias* term associated with the unit, as well as a term for input arising from outside the network. Thus for unit i, the net input is given by:

$$net_i = \sum_j w_{ij} a_j + bias_i + input_i.$$

Here j runs over all the units with connections projecting to unit i. The net input can then be used to set the new activation of the unit according to some monotonic but non-linear function such as the one shown in Fig. 2.3. Alternatively, the net input can be used to set the activation of the unit probabilistically to one of two discrete values (usually 1 or 0). Another possibility is that the net input may act as a force, tending to drive the activation of the unit up or down a small amount in each time step.[3]

It is typical to use some form of non-linear activation function, so that the activation of a unit is not simply set equal to the net input or some weighted average of the net input and the previous activation of the unit. Non-linearities are typically necessary for two reasons. 1. Linear networks are subject to explosive growth of activation due to positive feedback loops unless the weights are severely constrained (see Shrager, Hogg, and Huberman 1987). 2. Many computations require a layer of non-linear units between input and output. Without non-linearities, multiple layers of units add no additional

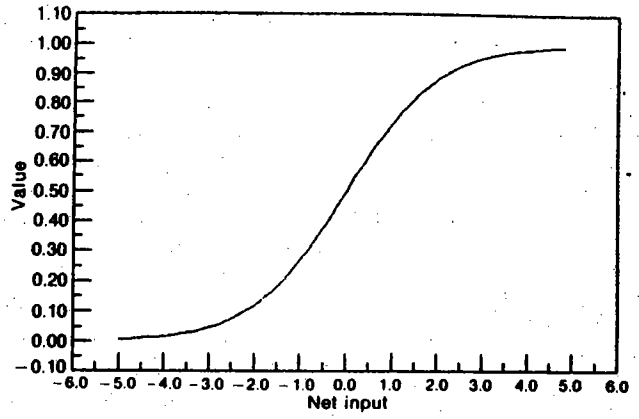


Fig. 2.3 The logistic function, a smooth non-linear function that is frequently used in relating activations of units to their net inputs. This function is often used to set the activation of a unit to a value between 0 and 1, or to set the activation of the unit to 1 or 0 probabilistically, with the probability determined by the value of the function.

computational power over that offered by a single layer (see Rumelhart, Hinton, and McClelland 1986, for further explanation).

Knowledge

Crucial to the very idea of cognition is the notion that information processing is guided by knowledge. We recognize the word THE as a definite article because of knowledge we have about the relation between letter strings and linguistic forms. We infer that a spoon may have been used if we hear 'the man stirred the coffee' because of knowledge we have about the kinds of instruments that are used for stirring. In many models, these kinds of knowledge would be stored in tables. For example, information about THE would be stored in a table called a lexicon, listing correspondences of letter strings and the linguistic objects they represent.

In connectionist models, knowledge is stored in the connections among the processing units. This assumption works together with the assumptions that connectionist models make about representations. An active representation on a set of units, together with the knowledge stored in connections, will give rise to new patterns of activation on the same or on other units.

Typically in connectionist models, connection strengths are real-valued. In models whose connections are set by assumption, it is typical to assume homogeneity of connection strengths as much as possible, to avoid excessive

degrees of freedom. In models that learn, however, connection strengths are typically allowed to take on whatever values the learning process gives them; parsimony arises from the use of a homogeneous principle of learning.

Learning

If knowledge is in the connection weights, learning must occur through the adjustment of these weights. This weight adjustment process is assumed to occur as a by-product of processing activity. Some knowledge can in fact be built into connectionist models, in the form of initial connection strengths, before there has been any learning, but it is common to explore the limits of what can be acquired through connection strength adjustment with minimal pre-wiring. The initial architecture of the network serves to impose constraints on the learning process; these can in many cases greatly facilitate learning and generalization, if these constraints are appropriate to the problem the network is given to learn.

A wide variety of 'learning rules' for tuning connections has been proposed. A recent review is provided by Hinton (in press). Generally, these rules state that the adjustment that is made to each connection should be based on the product of a 'pre-synaptic' term, associated with the unit that is sending input through the connection, and a 'post-synaptic' term, associated with the unit that is receiving input through the connection. For example, the *Hebb rule*, as used by Anderson (1977), makes the change in the strength of a connection proportional to the product of the activation of the sending unit and the receiving unit.[4] Learning through connection strength adjustment is very different from learning processes in most other types of models. It is governed by simple mathematical expressions, and results in knowledge that is completely implicit, in that it is embedded inextricably in the machinery of processing, and is completely inaccessible to introspection or report. However, it should be noted, that while the connection changes themselves are not accessible, the patterns of activation whose construction they make possible can be accessible to other parts of the processing system.

The environment

Though it has been implicit in what I have said already, there is another aspect of connectionist models that deserves comment, namely their *environment*. The environment consists of an ensemble of possible patterns that might be presented to the network. In most cases, these patterns are thought of as separate events; each one presented when the network is in a resting state, then left on until processing is complete. However, input patterns can have a richer

temporal structure, or course; each event may consist of a sequence of events, or of a graded progression of input activations.

For networks with fixed connections, the environment simply defines the domain of inputs on which the network might be tested. For networks in which the connections are adjusted as a result of processing experience, however, the environment plays a crucial role in determining exactly what is learned. Thus models that aim to capture aspects of cognitive development through connectionist learning include among their assumptions a specification of the details of the experience that gives rise to the resulting developmental sequence. In many cases, these assumptions play a major role in determining the success or failure of the modelling effort.

The spirit of the thing

The connectionist framework is cast, not as a list of specific detailed assumptions, but as a set of *general principles* and some guidelines that provide weak constraints on the range of variants that fall within the scope of these principles. Indeed, as Rumelhart, Hinton, and McClelland (1986) noted, it is possible to build a von Neumann computer out of connectionist primitives, if they are organized in accordance with the von Neumann architecture. It thus becomes important to focus on the spirit of the connectionist framework. Generally, connectionist models of cognitive processes have been constructed expressly to exploit the capability for parallelism inherent in the approach, to make use of the graded capabilities of patterns of activation, and to capture the incremental nature of human learning in many tasks through the adjustment of connection strengths based on signals arising in the course of processing.

The microstructure of cognition

Finally, it is worth pointing out that the connectionist framework is not incompatible with other levels of description in cognitive science. Thus, there is nothing inconsistent with connectionist models in the claim that a cognitive system may traverse a sequence of states in a temporally extended cognitive task such as solving an arithmetic problem. According to the connectionist approach one would tend to view each such step in the process of solving the problem as a new state of the processing network. Indeed, Rumelhart *et al.* (1986) describe a network that performs a mental tic-tac-toe simulation, settling into a sequence of states representing the results of the successive mentally simulated moves made by each player.

There are important differences between conventional and connectionist models of sequential behaviour. In connectionist models, the states need not

be so discrete as they generally are in other models (Rumelhart and Norman 1982; Jordan 1986; Smolensky 1986). Furthermore, the powerful constraint–satisfaction characteristics inherent in the connectionist framework are not typically exploited by conventional models of sequential processing. The idea that each step is a sequential process involves a massively parallel constraint–satisfaction process seems like a promising starting place for a new way of thinking about the macrostructure of cognition.

The point that connectionist models characterize the microstructure of cognition applies not only in respect of time, but also in respect of the structure of the processing system and in respect of the description of the computational operations that the system is performing. Structurally, a processing system may consist of many parts, and for some purposes it may be adequate to describe its structure in terms of these parts and the flow of information between them. Computationally, too, it may often be useful and illuminating to describe what a part of such a system computes without referring specifically to the role in this computation that is played by the specific units and connections. The claim is, though, that it will be necessary to delve more deeply than this to provide a full description of the mechanisms of cognition.

Are connectionist models mere implementations?

In allowing that there may be a macrostructure to thought, connectionists may seem to suggest that their models merely describe the implementation details of a processing system that would be best characterized more abstractly. However, we simply do not know exactly what level of description is the appropriate one for characterizing many behavioural phenomena. Many who have turned to connectionist models have done so because these models have seemed to provide exactly the right level of description for characterizing certain kinds of cognitive processes. Just where the bounds of usefulness of the connectionist framework may lie seems at this point to be one of the very open questions. Since there is little in cognitive psychology that we understand perfectly at present, we are not in a position to say which aspects of cognition might be explainable without recourse to a model of the microstructure.

Connectionist models and cognitive development

In the preceding part of this chapter, I have tried to give an overview of the connectionist framework for cognitive modelling. Here, I consider the question: Does the connectionist framework have any implications for the answers that we give to basic questions about human cognition? I will argue that it does. The questions are ones that arise within the field of cognitive development; they are motivated by dramatic behavioural phenomena. Several different kinds of answers have been given to these questions. We will see how the connectionist framework opens them anew and suggests what may be different answers in many cases.

The phenomena

The field of cognitive development is replete with examples of dramatic changes in children's thinking as they grow older. Here I give three examples: (1) failures of conservation and compensation; (2) progressive differentiation of knowledge about different kinds of things; and (3) U-shaped learning curves in language acquisition.

Failures of conservation and compensation

Perhaps the best-known phenomena in cognitive development are the dramatic failures of conservation that Piaget has reported in a wide range of different domains. One domain is the domain of liquid quantity. A child of 3 years is shown two glasses of water. The glasses are the same, and each contains the same amount of water, and the child sees that the amount is the same. But when the contents of one of the glasses is poured into a wider container, the child will say that there is less liquid in the wider container.

It is typical to say that this answer given by the young child reflects a failure to recognize two things: (1) that quantity is conserved under the transformation of pouring from one container to another; and (2) that greater width can compensate for less height. Many tasks are specifically designed to tap into the child's ability to cope with these kinds of compensation relations between variables.

One such task developed by Inhelder and Piaget (1958), the so-called *balance-beam task*, is illustrated in Fig. 2.4. In this task, the child is shown a

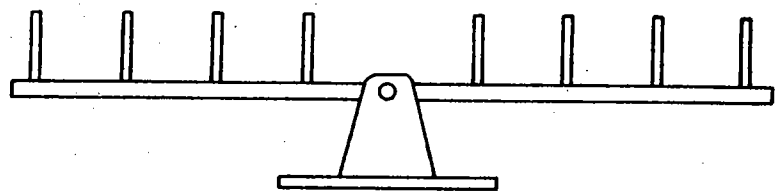


Fig. 2.4 Balance beam of the kind first used by Inhelder and Piaget (1958), and later used extensively by Siegler (1976, 1981; Siegler and Klahr 1982). (Reprinted from Siegler 1976, Fig. 1, with permission.)

balance beam with pegs placed at evenly spaced intervals to left and right of a fulcrum. On one peg on the left are several weights; on one peg on the right are several weights. The beam is immobilized, and the child is asked to judge which side will go down, or whether they will balance. We will have occasion to examine performance in this task at length below; for now it suffices to note that young children (up to about 6 or 7 years in this case) typically respond as if the distance from the fulcrum was completely irrelevant. They will say the beam should balance if the weight is the same on both sides, regardless of distance. Otherwise they say the side with the greater weight will go down. These children, then, appear to miss the fact that lesser weight can be compensated for by greater distance. Typically by the age of 11 years or so children have some appreciation for this trade off; the details of the developmental progression are quite interesting, as we shall see below.

Progressive differentiation of ontological categories

Other researchers, studying different domains, have noticed other kinds of developmental progressions. Keil (1979) studied children's judgements about whether you could say things like 'A rabbit is an hour long'. He supposed such judgements tapped children's knowledge about different kinds of things. In these judgements, Keil was interested not in whether the child saw a sentence as true or false, but in whether the child felt that one could make certain kinds of predictions (e.g. that something is an hour long) when the something is a member of a certain 'ontological category' (e.g. living thing). Keil found that children were much more permissive in their acceptance of statements than adults were, but their permissiveness was not simply random. Rather, children would accept statements that overextended predicates to categories near the ones that they typically apply to, but would not extend them further. Thus some children will accept predications like 'The rock is asleep', but not 'The rock is an hour long'. It was as though children's knowledge of what predicates apply to particular categories becomes progressively more and more differentiated, as illustrated in Fig. 2.5.

U-Shaped learning curves in language development

Early on, children often get certain kinds of linguistic constructions correct which they later get wrong; only much later do they recover their former correct performance. One example is the passive construction, applied to semantically biased materials, such as 'The man was bitten by the dog'. (See Bever, 1970, for a discussion of the development of the use of the passive construction.) Early in development, children correctly interpret such sentences; they appear to be using information about what roles the different

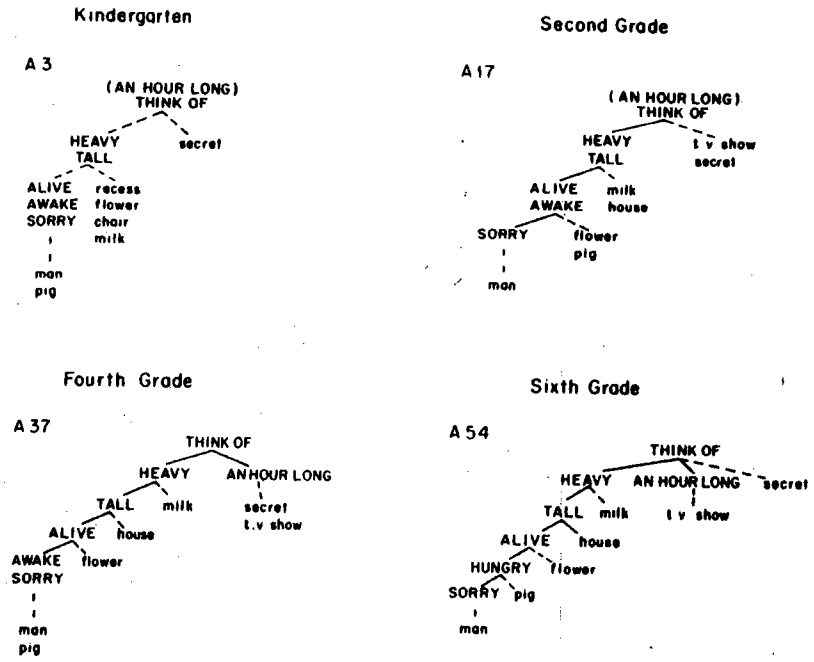


Fig. 2.5 Four different 'predictability trees' illustrating the progressive differentiation of concepts as a function of age. Terms in capitals at internal nodes in the trees represent predicates, and terms in lower case at terminal nodes in the trees represent concepts that are spanned by all the predicates written on nodes that dominate the terminal. A predicate spans a concept if the child reports that it is not silly to apply either the predicate or its negation or both to the concept. Thus the first tree indicates that the child will accept 'The man is (not) alive', and 'The chair is (not) tall', but will not accept 'The chair is (not) alive'. Parentheses indicate uncertainty about the application of a predicate. (Redrawn from Keil, 1979, Appendix C, with permission.)

nouns typically play in the action described by the verb, since they tend to be correct only when the correct interpretation assigns the nouns to their typical roles. At an older age, children respond differently to such sentences, treating the first noun-phrase as the subject; semantic constraints are over-ridden, and there is a tendency to interpret 'The man was bitten by the dog' as meaning 'The man bit the dog'. Finally, children interpret the sentence correctly again, but for a different reason. It would appear that they now understand passives in general, since at this stage they can also interpret semantically neutral and even reverse-biased sentences (such as 'The dog was bitten by the man') correctly.

The questions

The phenomena reviewed above raise basic questions about cognitive development. Three of these questions are:

1. Are these different phenomena simply unrelated facts about development in different domains?
2. Are there principles that all of these phenomena exemplify?
3. If there are principles, are they domain specific, or are they general principles about development?

Different kinds of developmental theorists have answered such questions in very different ways. To Piaget, each failure of compensation or conservation reflected a single common developmental stage; the phenomena were intrinsically related by the characteristics of the stage, and these characteristics provided the basis for explanation.

Others have taken a very different approach. Keil (1979), following Chomsky's analogous argument for language, argues for *domain specific principles* of development. His view is that each cognitive domain has its own laws that provide constraints on what can be learned. These constraints limit the hypotheses that the child can entertain, thereby making it dramatically easier for the child to acquire adult abilities in the face of the impoverished information that is provided by experience with the world.

The main thrust of the remainder of this chapter is to argue that recent developments in connectionist learning procedures suggest a dramatic alternative to these kinds of views. The alternative is simply the hypothesis that these diverse developmental phenomena all reflect the operation of a single basic learning principle, operating in different tasks and different parts of the cognitive system.

The learning principle

The principle can be stated in fairly abstract terms as follows:

Adjust the parameters of the mind in proportion to the extent to which their adjustment can produce a reduction in the discrepancy between expected and observed events.

This principle is not new. It might well be seen as capturing the residue of Piaget's accommodation process, in that accommodation involves an adjustment of mental structures in response to discrepancies. (See Flavell, 1963, for a discussion of Piaget's theory.) It is also very similar to the principle that governs learning in the Rescorla–Wagner model of classical conditioning (Rescorla and Wagner 1972). What is new is that there exists a learning procedure for multilayer connectionist networks that implements this principle. Here, the parameters of the mind are the connections among the units in the network, and the procedure is the back-propagation procedure of Rumelhart, Hinton, and Williams (1986; see Hinton, this volume).

The learning principle lies at the heart of a number of connectionist models that learn how to do various different kinds of information processing tasks, and that have applications to phenomena in cognitive and/or language development. Perhaps the simplest such model is the past-tense model of Rumelhart and McClelland (1986). The development of that model pre-dated the discovery of the back-propagation learning procedure, thereby forcing certain simplifications for the sake of developing an illustration of the basic point that lawful behaviour might emerge from the application of a simple principle of learning to a connectionist network. Subsequent models have used back-propagation to overcome some of these limitations. Included in this class are NETtalk (Sejnowski and Rosenberg 1987) and a more recent model of word reading (Seidenberg, Patterson, and McClelland, this volume). The present effort grew out of two observations of similarities between the developmental courses seen in models embodying this principle, and the courses of development seen in children: First, the course of learning in a recent model of concept learning by Rumelhart (in preparation) is similar to aspects of the progressive differentiation of concepts reflected in Keil's (1979) studies of predictability. Second, the course of learning in a recent model of sentence comprehension by St John and McClelland (1988) mirrors aspects of the progression from reliance on semantic constraints, to reliance on word order, to, finally, reliance on complex syntactic patterning such as the passive voice. I do not mean to claim that the models in question are fully adequate models of the developmental progression in either case; I only claim that they seemed suggestive: they raised the possibility that part of the explanation of these and other developmental phenomena might be found in the operation of the learning principle as it adjusts connection strengths in a network subjected to patterns arising in its environment.

The remainder of this chapter presents an experiment assessing the applicability of this conjecture to another developmental phenomenon, namely the acquisition of the ability to take both weight and distance into account in the balance beam task described above. The task has been studied extensively by Siegler and his colleagues (Siegler 1976, 1981; Siegler and Klahr 1982), and quite a bit is known about it. I will first review the developmental findings. Then I will describe a connectionist model that captures these phenomena by applying the learning principle stated above.

Development of judgements of balance

In an important monograph, Siegler (1981) studied children's performance in the balance beam task and three other tasks in which two cues had to be taken

into account for correct performance. In all cases, as in the balance beam task, the correct procedure requires multiplication. For example, in the balance beam task, to determine which side will go down, one must multiply the amount of weight on a given side of the beam by the distance of that weight from the fulcrum. The side with the greater product will go down; when the products are the same, the beam will balance.

Siegler studied children in several age groups, as well as young adults. Each child was asked to judge 24 balance problems. In each case, the beam was immobilized so that there was no feedback. The 24 problems could be divided into four of each of six types:

1. *Balance*. In this class of problem, the weight is the same on both sides of the beam and the weight is the same distance from the fulcrum on both sides.
2. *Weight*. In these problems, the weights differ but distance from the fulcrum is the same on both sides.
3. *Distance*. Here the weight is the same on both sides, but the distance from the fulcrum differs.
4. *Conflict*. Here both weight and distance differ and are in conflict, in that the weight is greater on one side but the distance from the fulcrum is greater on the other. There are three types of conflict problems:
   (a) conflict-weight. In these cases, the side with the greater weight has the greater torque (i.e. the greater value of the product of weight times distance).
   (b) conflict-distance. In these cases, the side with the greater distance has the greater torque.
   (c) conflict-balance. Here the torques are the same on both sides.

Siegler's analysis of children's performance assumed that children use rule-governed procedures. Four such procedures or *rules* as Siegler called them are shown in Fig. 2.6. Each of these rules corresponds to a distinct pattern of performance over the six problem types. For example, children using Rule I should say the side with the greater weight will go down in weight problems and in all three types of conflict problems. They should think the beam will balance on balance problems and distance problems. In general, the mapping from the rules to expected performance is extemely straightforward. The only point that needs explication is the instruction *muddle through* when weight and distance conflict in Rule 3. In practice it is assumed to mean 'guess randomly among the alternatives', so that $\frac{1}{3}$ of the responses would be left-side-down; $\frac{1}{3}$ right-side-down, and $\frac{1}{3}$ balance.

Siegler compared the performance of each child tested with each rule, and counted discrepancies from predicted performance based on the rule. Children who scored less than four discrepancies from a given rule were scored as using that rule.

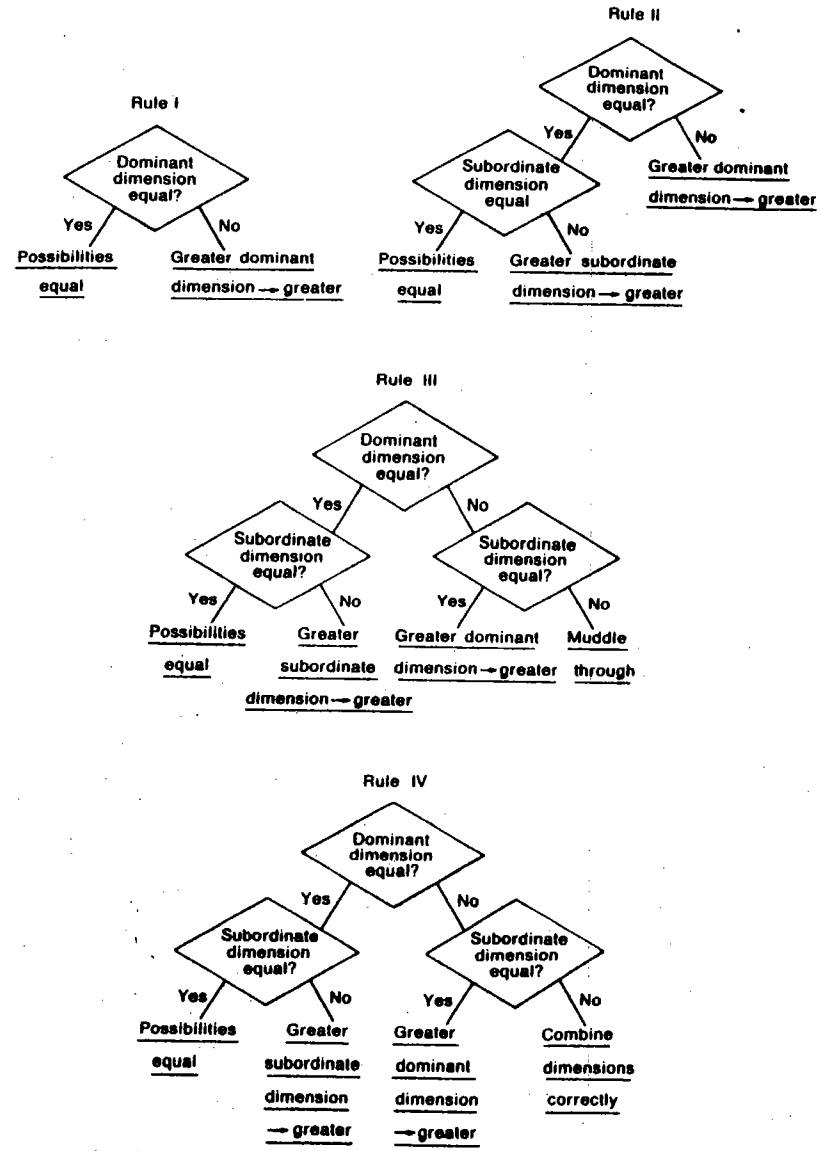


Fig. 2.6 Siegler's (1976, 1981) four 'rules' for answering balance beam problems. Each rule is in fact a full procedure, rather than a single rule. (Reprinted with permission from Siegler (1981), Fig. 1.)

For our purposes, there are four basic findings that emerge from Siegler's analysis:

1. *Lawful behaviour.* In general, performance of children over the age of 5 years is extremely regular in the balance beam task. Overall, about 90 per cent of children tested conform to one of the four rules.
2. *Developmental progression.* As children get older, they appear to progress through the use of the different rules. The progression from Rule 1 to Rule 3 can be thought of as a progression in which at first the weight cue is relied on exclusively, while at the end distance and weight are both taken into account. In between (Rule 2), distance is taken into account only if it does not conflict with the weight cue. Children aged 5–7 years typically use Rule 1, and college students typically use Rules 3 or 4. Many college students do not have explicit knowledge of the torque principle. Children younger than age 5 years tend not to be scorable strictly in terms of one of the rules; however, they appear to show an increasing tendency to behave in accordance with Rule 1.
3. *Generality.* The same four rules appear to be adequate to characterize performance in all three of the domains that Siegler studied. Though the developmental progression was not identical across cases, there was in all cases a trend from simpler to more complex rules with development.
4. *Lack of correlation between domains.* Even though children seem to progress through the same rules in different domains, they do not do so in lock-step; the correlation across domains is low, particularly in terms of the higher-numbered rules, so that children who are showing Rule 3 behaviour in one task may be showing Rule 1 behaviour in another and Rule 4 in a third.

The simulation model

The model I have developed to capture Siegler's findings is sketched in Fig. 2.7.[5] Of course, the model is a drastic oversimplification of the human mind and of the task; but as we shall see it allows us to capture the essence of Siegler's findings, and to see them emerge from the operation of the learning principle described above.

The model consists of a set of input units, to which balance problems can be presented as patterns of inputs; a set of output units, over which the answer to each problem can be represented; and a set of hidden units, between the input and the output. Connections run from input units to hidden units and from hidden units to output units.

The input units can be divided into two groups of 10. One group is used to

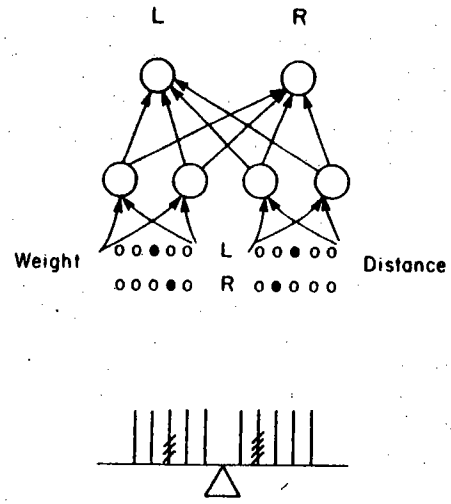


Fig. 2.7 The network used in the simulation of the development of performance in the balance beam task.

represent information about weight and the other is used to represent information about distance. In each case I have chosen to use an input representation that imposes as little structure as possible on the input patterns. Each possible value of weight or distance from the fulcrum is assigned a separate unit. The ordering of values from low to high is not given in this representation; the network will have to learn this ordering. For the convenience of the reader, the units are arranged in rows according to which side of the beam they are from, and within each row they are arranged from left to right in order of increasing weight or distance from the fulcrum; but this ordering is unknown to the model before it is trained, as we shall see.

Though the two dimensions are not intrinsically structured for the model, the design of the network does impose a separate analysis of each dimension. This separation turns out to be critical; I will consider the implications of this architectural simplification below. The separation is implemented as follows: there are separate pairs of hidden units for each dimension. Two hidden units receive input from the weight input units and two receive input from the distance input units.

Each of the four hidden units projects to each of the two output units. The left output unit can be thought of as a 'left-side-down' unit, and the right one as a 'right-side-down unit'. Thus a correct network for the task would turn on the output unit corresponding to the side with the greater torque, and would turn off the unit for the other side. For balance problems, I assume that the network should turn both units on half-way. Note that this coding of the output

patterns does tell the network that balance is between left-side-down and right-side-down.

Processing

Balance problems of the kind studied by Siegler can be processed by the network by simply turning on (i.e. setting to 1) the input units corresponding to a particular problem and turning off (i.e. setting to 0) all other input units. The input from the problem illustrated in Fig. 2.7 is shown by using black to indicate those input units whose activations are 1.0, and white for the units whose activations are 0.

The inputs are propagated forward to the hidden units. Each hidden unit simply computes a net input:

$$net_i = \sum_j w_{ij} a_j + bias_i.$$

Here j ranges over the input units. Each hidden unit then sets its activation according to the logistic function:

$$a_i = \frac{1}{1 + e^{-net_i}}$$

In these equations, w_{ij} is the strength of the connection to hidden unit i from input unit j, a_j is the activation of input unit j, and $bias_i$ is the modifiable bias of hidden unit i. This bias is equivalent to a weight to unit i from a special unit that is always on.

Once activations of the hidden units are determined, the activations of the output units are determined by the same procedure. That is, the net input to each output unit is determined based on the activations of the hidden units, the weights from the hidden units to the output units, and the biases of the output units. Then the activations of the output units are determined using the logistic function.

Responses

The activations of the output units are real numbers between 0 and 1; to relate the model's performance to the balance beam task, these real-valued outputs must be translated into discrete responses. I used the following simple translation: if the activation of one output unit exceeded the activation of the other by 0.333, I took the answer to be 'more active side down'. Otherwise, the answer was assumed to be 'both sides equal'.

Learning

Before training begins, the strengths of these connections from input to hidden units and from hidden to output units are initialized to random values uniformly distributed between +0.5 and −0.5. In this state, inputs lead to

random patterns of activity over both the hidden and output units. The activations of the output units fluctuate approximately randomly between about 0.4 and 0.6 for different input patterns. The network comes to respond correctly only as a result of training. Conceptually, training is thought of as occurring as a result of a series of experiences in which the network is shown a balance problem as input; computes activations of output patterns based on its existing connection weights; and is then shown the correct answer. The signal that drives learning is the difference between the obtained activation of each output unit and the correct or target activation for that unit. The back-propagation procedure of Rumelhart, Hinton, and Williams (1986) is then used to determine how each connection strength in the network should be adjusted to reduce these differences. The procedure is described in Hinton's chapter in this volume, and it would be redundant to describe it here. Suffice it to say that it exactly implements the learning principle stated above, and restated here in network terminology:

Adjust each weight in the network in proportion to the extent to which its adjustment can produce a reduction in the discrepancy between the expected event and the observed event, in the present context.

Here the 'expected event' is the pattern of activation over the output units that is computed by the network; the observed event is the pattern of activation that the environment indicates these units have; and the present context is the pattern of activation over the input units. Note that the direction of change to a connection (positive or negative) is simply the direction that tends to reduce the discrepancy between computed output and the correct or target output.

Environment

As I pointed out at the beginning of this chapter, the environment in which a network learns plays a very strong role in determining what it learns, and particularly the developmental course of learning. The simulations reported here were based on the assumption that the environment for learning about balance problems consists of experiences that vary more frequently on the weight dimension than they do on the distance dimension. Of course, I do not mean to suggest that all the learning that children do and that is relevant to their understanding of balance takes the form of explicit balance problems of the kind my network sees. Rather, my assumption that the experience on balance problems is dominated by problems in which there is no variability in weight is meant as a proxy for the more general assumption that children generally have more experience with weight than with distance as a factor in determining the relative heaviness of something.[6] The specific assumptions about the sequence of learning experiences were as follows. The environment consisted of a list of training examples containing the full set of 625 possible

problems involving 25 combinations of possible weights (1–5 on the left crossed with 1–5 on the right) crossed with 25 combinations of possible distances (1–5 steps from the fulcrum on the left crossed with 1–5 steps from the fulcrum on the right). Two corpuses were set up. Problems in which the distance from the fulcrum was the same on both sides were listed five times each in one corpus, and 10 times each in the other corpus. Other problems were listed only once in each corpus.

Training and testing regime

Four simulation runs were carried out, two with each of the two corpora just described. In each run, training consisted of a series of epochs. In each epoch, 100 patterns were chosen randomly from the full list of patterns in the corpus. In each epoch, weight increments were accumulated over the 100 training trials and then added into the weights at the end of the epoch, according to the momentum method described in Rumelhart, Hinton, and Williams (1986, p. 330); parameters were $\eta = 0.075$, $\alpha = 0.9$).

After weight updating at the end of each epoch, the network was given a 24-item test, containing four problems of each of the six types described above, taken from an experiment of Siegler's. (A few of the examples had to be modified since Siegler's experiment had used up to six pegs.)

A comment on the simulation model

The model described above obviously simplifies the task that the learner faces and structures it for him to some degree. In particular, it embodies two principal assumptions which are crucial to the successful simulations we will consider below:

1. *Environment assumption.* The model assumes that the environment is biased, so that one dimension—in this case weight—is more frequently available as a basis for predicting outcome than the other.
2. *Architecture assumption.* The model assumes that the weight and distance dimensions are analysed separately, before information about the two dimensions is combined.

Both these assumptions are crucial to the success of the model. In an unbiased environment, both cues would be learned equally rapidly. Effects of combining the cues from the start, as prescribed by the architecture assumption, are more complex, but suffice it to say for now that the apparent stagelike character of performance is much less clear unless this assumption is adopted.

An important topic for further research will be to examine what variants of these assumptions might still allow the model to be successful. For example, regarding the environment, differences in salience (i.e. strength of input

activations) and structuredness of the dimensions might also produce similar results.

The issue of structuredness of the dimensions is a key point that needs to be considered as it relates to the present simulation. For both dimensions, the input representations encode different weights and distances from the fulcrum using distinct units. This means that different values are distinguishable by the model, but they are not structured for it; for example, the input itself provides no indication that a distance or weight of 3 is between 2 and 4. The network must learn to represent the weights and distances in structured ways in order to solve the balance problem. Below we will see that it does so.

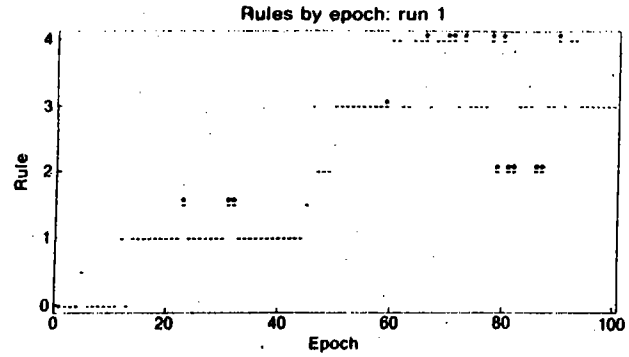
Results

In general, performance of the model conformed to one of the four rules described by Siegler. Over the four runs, the model fit the criteria of one of Siegler's four rules on 85 per cent of the occasions, not counting an initial, pre-Rule 1 period discussed below (in Siegler, 1981, the conformity figure is about 90 per cent). Of course, the model was not consulting these rules or following the step-by-step procedures indicated in them; rather its behaviour was simply scorable by Siegler's criteria as consistent with the succession of rules. Excluding the initial period, failures to fit the rules were of three types: (1) cases in which a rule fit except for a position bias that gave difficulty on balance problems; (2) cases in which performance was borderline between Rules 1 and 2; and (3) combinations of these two problems. (Siegler (personal communication) does find some borderline cases between Rule 1 and Rule 2, but the position bias cases are not typical of children's performance.)

Overall development trends

Epoch by epoch performance in each of the four runs is shown in Figs 2.8 and 2.9. One generally observes the expected developmental progression. Each simulation run is slightly different, due to differences in the random starting weights and the sequence of actual training experiences, but there are clear common trends. Over the first 10 epochs or so, the output of the model was close to 0.5 on all test patterns; by our scoring criteria, all these outputs count as 'balance' responses, but of course they really represent a stage in which neither weight nor distance governs performance. The next few epochs represent a transition to Rule 1, in that in this phase the model is showing some tendency to activate the output unit on the side with the greater weight, but this tendency is variable across patterns and the discrepancy between the activations of the output units is not reliably greater than 0.33 when the weights differ.

After this brief transition, performance of the model had generally reached the point where it was responding consistently to the weight cue while



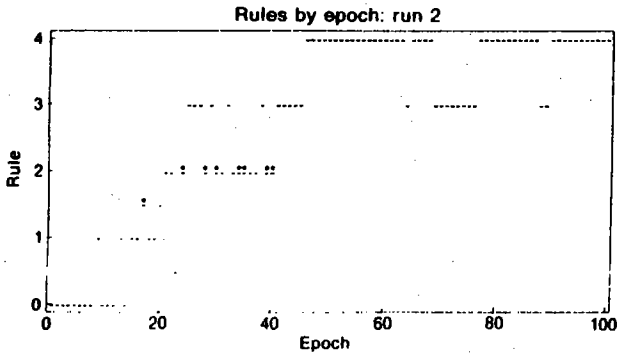
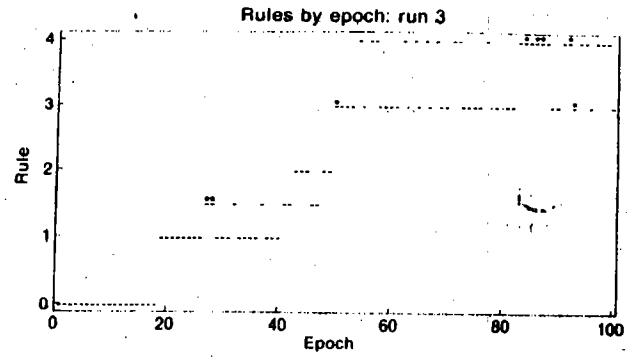


Fig. 2.8 Epoch-by-epoch performance of the simulation model in the two runs with a 5 to 1 bias favouring problems in which distance did not vary. Performance is scored by rule. Cases marked by * missed a rule due to position bias. Rule 0 corresponds to always saying 'balance', and occurs at the beginning of training. Rule 1.5 corresponds to performance on the borderline between Rules 1 and 2.

systematically ignoring the distance dimension. This pattern continued for several more epochs. There was a brief transitional period, in which the model behaved inconsistently on the *distance* problems crucial to distinguishing between Rule 1 and Rule 2 behaviour. After several epochs in this phase, use of the distance cue reached the point where performance on all types of conflict problems become variable. The model generally continued in this phase indefinitely, sometimes reaching the point where its performance was generally scorable as fitting Rule 4 and sometimes not.

The variability in the model's performance from epoch to epoch is actually



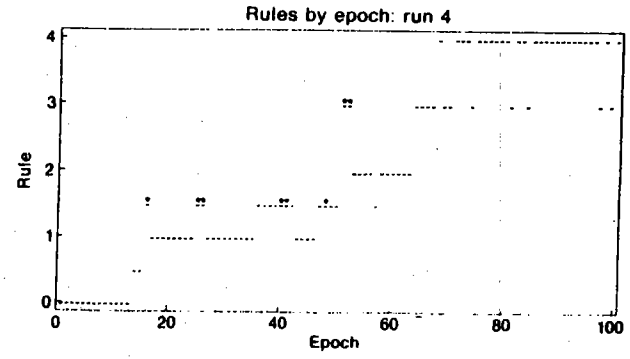


Fig. 2.9 Epoch-by-epoch performance of the simulation model in the two runs with a 10 to 1 bias favouring problems in which distance did not vary. Performance is scored by rule, as in Fig. 2.8.

quite consistent with test re-test data reported in Siegler (1981). Rule 2 behaviour is highly unstable, and there is some instability of behaviour in other rules as well.

Performance in each phase

Seigler's criteria for conformity to his rules allow for some deviations from perfect conformity; in fact only 83 per cent of test problems must be scorable as consistent with the rule. Given this, it is interesting to see whether the discrepancies from the rules that are exhibited by the model are consistent with human subects' performances. In general, they seem to be quite consistent, as

Fig. 2.10 indicates. Each panel shows percentage correct performance by the model averaged over the tests on which the model scored in accordance with one of the four rules. Also shown are data from two groups of human subjects as well as the pattern of performance that would be expected from a perfect rule user.

For Rule 1, the model differs very little from human data. For Rule 2, again the correspondence to human data is very close. Both the model and the humans show some slight tendency to get *conflict-distance* problems correct, and to occasionally miss *distance* and *balance* problems. For both Rule 1 and Rule 2, the tendency to miss *balance* problems is slightly greater in the model than in the children's data. For Rule 3, the model exaggerates a tendency seen in the human data to be correct on *conflict-weight* problems more often than on *conflict-distance* problems. The major discrepancy in the data is that the model is too accurate on *conflict-balance* problems. For Rule 4, the model again exaggerates a tendency seen in the human data to have residual difficulties with conflict problems.

With the exception of the *conflict-balance* problems in Rule 3, the human data seem to fall about half-way between the model and perfect correspondence to the rules. It is tempting to speculate that some human subjects—particularly Rule 4 subjects—may in fact use explicit rules such as the torque rule some of the time. It is, indeed, easy for the adult subjects who contribute to the Rule 4 results to follow the torque rule if instructed specifically in this rule. However, it is evident that the subjects who fall under the Rule 4 scoring criteria do not in fact adhere exactly to the rule. Perhaps this group includes some individuals performing on the basis of implicit knowledge of the trade-off of weight and distance as well as some who explicitly use the torque rule, and perhaps some individuals use a mixture of the two strategies.

Further correspondences between the model and child development

So far we have seen that the balance beam model captures the pattern of development seen in the studies of Siegler (1976, 1981). There are two further aspects of the developmental data which are consistent with the gradual buildup of strength on the distance dimension that we see in the model:

1. Wilkening and Anderson (in press) present subjects with one side of a balance beam, and allow them to adjust the weight on the other side at a fixed distance from the fulcrum to make the beam balance. Over the age range of 9–20 years, in which children are generally progressing from late Rule 1 or Rule 2 to Rule 3 or Rule 4, according to Siegler's methods, they find an increasing sensitivity to the distance cue. Unfortunately, it is difficult to be sure whether this reflects different numbers of subjects relying on the distance cue, or (as we see in the model) differences in degree of reliance among those who show some sensitivity to the distance cue.

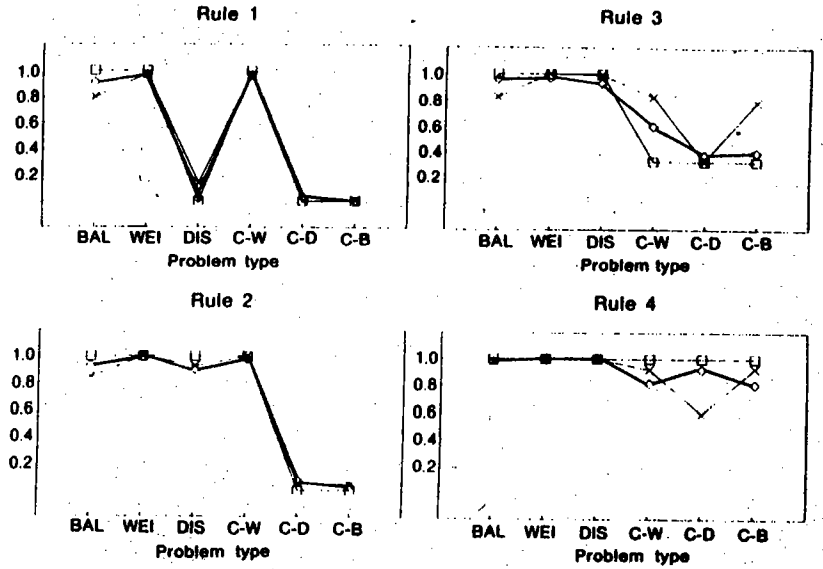


Fig. 2.10 Children's performance by problem type on the balance beam task, together with the performance of the simulation model and expected performance based on each rule. The heavy line with diamonds indicates children's performance. The model's performance is given by the light line with × s, while performance predicted from the rule is given by the light line with squares. For each child and each test of the simulation, performance was pre-categorized according to the best-fitting rule. Then, percentage correct responses by problem type were calculated averaging over children or simulation tests falling into each rule.

2. For children who exhibit Rule 3 on Siegler's 24-item test, careful assessment with a larger number of conflict problems indicates the use of cue compensation strategies, rather than random guessing (Ferretti *et al.* 1985). Thus children are not simply totally confused about conflict problems during this stage but have some sensitivity of relative magnitudes of cues, as does the model. The exact degree of correspondence of the model's performance and human performance on these larger tests remains to be explored.

The mechanism for developmental change

Given the generally close correspondence between model and data, it is important to understand just how the model performs, and how its performance changes. To do this, it is helpful to examine the connections in the network at several different points in the learning process: Figure 2.11 displays

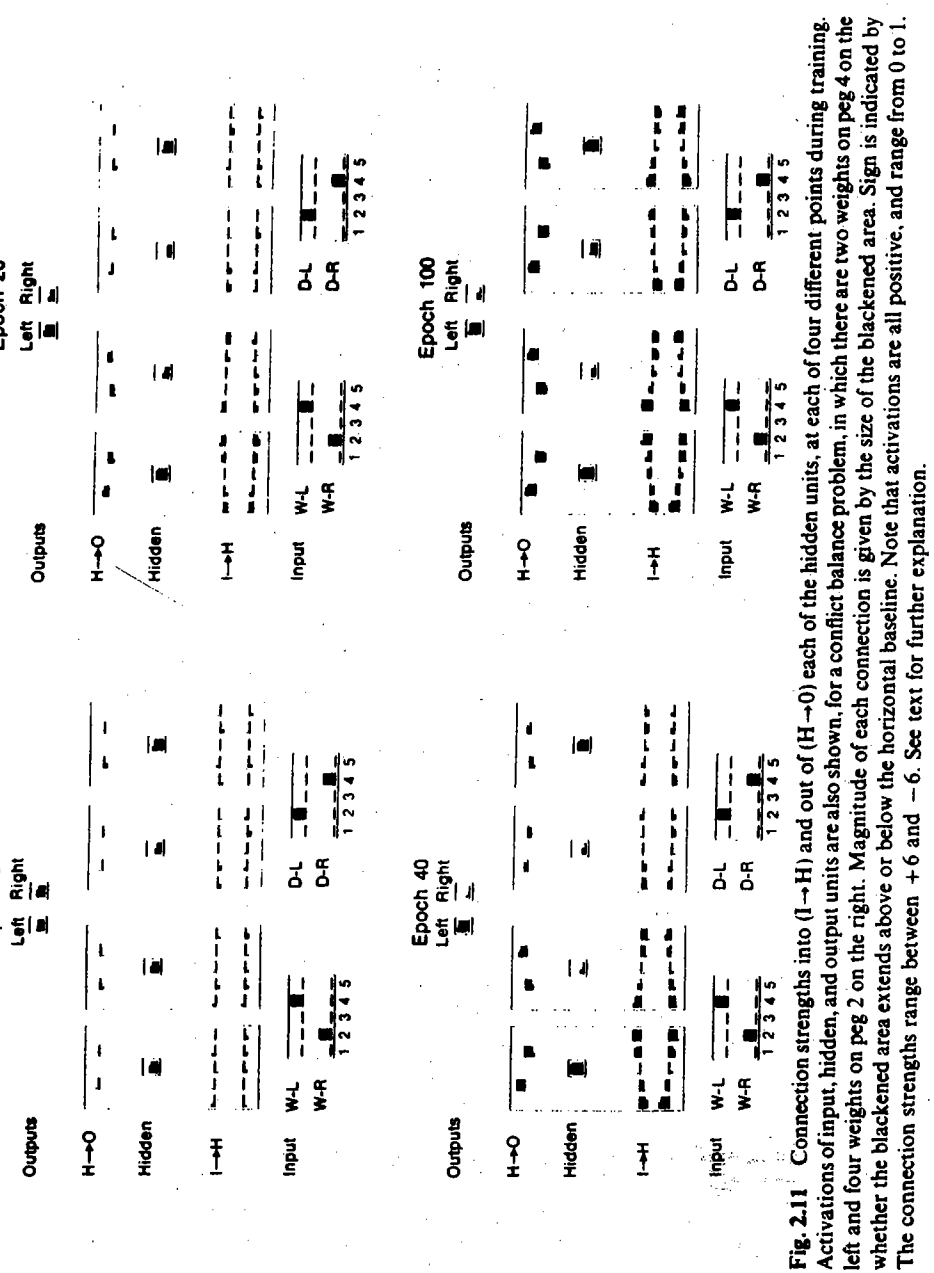


Fig. 2.11 Connection strengths into (I→H) and out of (H→O) each of the hidden units, at each of four different points during training. Activations of input, hidden, and output units are also shown, for a conflict balance problem, in which there are two weights on peg 4 on the left and four weights on peg 2 on the right. Magnitude of each connection is given by the size of the blackened area. Sign is indicated by whether the blackened area extends above or below the horizontal baseline. Note that activations are all positive, and range from 0 to 1. The connection strengths range between +6 and −6. See text for further explanation.

the connections from the run that produced the results shown in the top panel of Fig. 2.9, at four different points during learning. At epoch 0, before any learning; at epoch 20, early in the Rule 1 phase; at epoch 40, at the end of the Rule 1 phase; and at epoch 100, when the simulation was terminated. Each of the four subrectangles in each panel shows the weights coming into and out of one of the four hidden units. The two on the left receive input from the weight dimension, and the two on the right receive input from the distance dimension.

In the first panel, before learning begins, all the connection strengths have small random values. In this situation, the output of the hidden units is not systematically related to magnitudes of the weights or distances, and is therefore of no use in predicting the correct output. At this point, the hidden units are not encoding either relative weight or relative distance, and are therefore providing no useful information for predicting whether the left or right side should go down.

The first phase of learning consists of the gradual organization of the connections that process the amount of weight on each side of the balance beam. Recall that the network receives problems in which the distance cue varies much less frequently than problems in which the weight cue varies. Learning to rely on the weight cue proceeds more quickly than learning to rely on the distance cue as a simple result of this fact. The rate of learning with respect to each type of cue is relatively gradual at first, but then speeds up, for reasons that we will explore below. The relatively rapid transition from virtually unresponsive output to fairly strong reliance on the weight cue represents the brief transition to Rule 1 responding. The result of this phase, in the second panel of the diagram, is a set of connections that allow the hidden units on the left to reflect the relative amount of weight on the left v. the right side of the balance beam. The leftmost hidden unit is most strongly excited by large weights on the left and small weights on the right, and most strongly inhibited by large weights on the right and small weights on the left. The activation of this unit, then, ranges from near 0 to near 1 as the relative magnitude of weight ranges from much more on the right to much more on the left. Correspondingly, this unit has an excitatory connection to the left-side-down output unit, and an inhibitory connection to the right-side-down output unit. The second hidden unit mirrors these relationships in reverse. At this point, then, the hidden units can be said to have learned to represent something they were not representing before, namely the relative magnitude of the inputs. Note that this information is not explicitly contained in the inputs, which simply distinguish but do not order the different possible values of weight on the two sides of the balance beam.

At this point, the connection strengths in the distance part of the network remain virtually unchanged; thus, at the hidden unit level, the network has not yet learned to encode the distance dimension.

Over the next 20 epochs, connections get much stronger on the weight

dimension, and we begin to see some organization of the distance dimension. While this is going on, the overt behaviour of the network remains Rule 1 behaviour. The network is getting ready for the relatively rapid transition to Rule 2 and then to Rule 3 which occurs over the next several epochs of training (as shown in the top panel of Fig. 2.9), but at epoch 40, the end of the Rule 1 phase, the distance connections are still not quite strong enough to push activations of the output units out of the balance range. With further learning, the distance cue becomes stronger and stronger; this first causes the distance cue to govern performance when the weights are in balance, giving rise to Rule 2 behaviour. Further strengthening causes the distance cue to win out in some conflict problems, giving rise to behaviour consistent with Rules 3 and 4. At epoch 100 of this particular run, the weight dimension maintains a slight ascendancy, so that with the *conflict-balance* problem illustrated, the model activates the left-side-down unit most, corresponding to the side with the greater weight.

A couple of aspects of the developmental progression deserve comment. Learning is slow at first and then accelerates, as shown in Fig. 2.12. As the diagram illustrates, the connection strengths are largely insensitive to differences early on, then go through a fairly rapid transition in sensitivity, and then level off again. The acceleration seen in learning is a result of an inherent characteristic of the gradient descent learning procedure coupled with the architecture of the network. The procedure adjusts each connection in proportion to the magnitude of the effect that adjusting it will have on the discrepancy between correct and actual output. But the effect of a given connection depends on the strengths of other connections. Consider the connection coming into a hidden unit from one of the input units. An adjustment of the strength of this input connection will have a small effect on the output if the connections from the hidden unit to the output units are weak. In this case, the input connection will only receive a small adjustment. If, however, the connections from the hidden units to the output units are strong, an adjustment of the strength of the input connection will have a much larger effect; consequently the learning procedure makes a much larger adjustment in this case. A slightly different story applies to the connections from the hidden units to the output units. When the connections from the input to the hidden units are weak and random, the activations of the hidden units are only weakly related to the correct output. Under these circumstances, the adjustments made to the output weights tend to cancel each other out, and learning progress is very slow. It is only after the input weights become organized that learning can proceed efficiently on the output side of the hidden units.

The story I am telling would be a very sad one, were it not for the fact that it is not all or none. It is not that there is *no* learning at all at first; if there were, there would be no gradual change to the point where learning becomes more rapid. Rather, it is simply that initially learning is *very* gradual; so gradual that

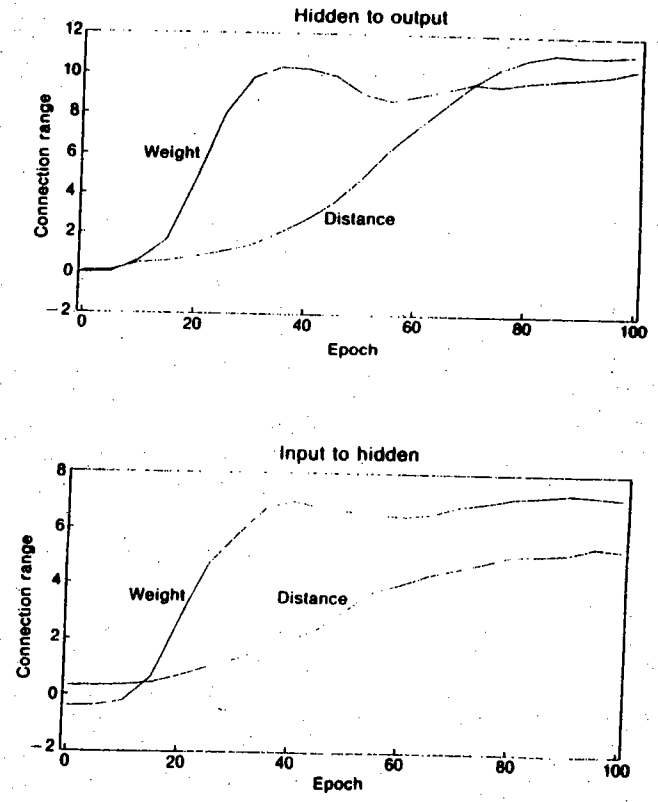


Fig. 2.12 Relative magnitude of connection strengths encoding weight and distance, as a function of training. Magnitude is given by the range of connection strengths (most positive minus most negative) coming into one weight or distance hidden unit (lower panel) and coming out of a weight or distance unit (upper panel).

it does not show up in overt behaviour. Gradually, though, this initially slowly learning accelerates, producing an increasing readiness to learn.

This differential readiness to learn allows the model to account for the results of an experiment described in Siegler and Klahr (1982), on the effects of training for young v. old Rule 1 children. They showed 5- and 8-year-old Rule 1 children a series of conflict problems. The children were allowed to try to predict which side would go down, and were then shown what actually happens. The results were striking. The older Rule 1 children were very likely to exhibit Rule 2 behaviour on a post-test. The younger children either continued to behave in accordance with Rule 1 or became inconsistent in their

responses. In further experiments on early Rule 1 children, Siegler and Klahr reported that these children do not represent the distance dimension correctly: when asked to reproduce a balance beam configuration, they could usually get the number of weights correct, but could rarely place them on the correct pegs. These findings are in complete conformity with the model: as we have seen, the model does not represent distance information early in Rule 1. Further simulations reported in McClelland and Jenkins (in preparation) show that the model can profit from conflict training of the sort used by Siegler and Klahr at the end of the Rule 1 phase but not at the beginning.

Shortcomings of the model

The model exhibits a striking correspondence with many aspects of the developmental facts, but does have a few shortcomings. Three failures to fit aspects of Siegler's data must be acknowledged. First, the model can never actually master Rule 4, though some subjects clearly do. Second, its behaviour during Rule 3 is slightly different from that of humans (though it should be noted that the 'human' Rule 3 pattern is actually a mixture of different strategies according to Klahr and Siegler 1978). Third, it can exhibit position biases which are uncharacteristic of humans, who seem (at least, from the age of 5 years on) to 'know' that there is no reason to prefer left over right.

There are other shortcomings as well. Perhaps the most serious is in the input representations, that use distinct units to represent different amounts of weight and distance. This representation was chosen because it does not inherently encode the structure of each dimension, thereby forcing the network to discover the ordering of each dimension. But it has the drawback that it prevents the network from extrapolating or even interpolating beyond the range of the discrete values that it has experienced.

Finally, Siegler has reported protocol data that indicate that subjects are often able to describe what they are doing verbally in ways that correspond fairly well to their actual performance. It is not true that all subject's verbalizations correctly characterize the rule they are using, but it is true, for example, that subjects who are sensitive to the distance cue mention that they are using this cue and those who are not tend not to mention it. The model is of course completely mute.

What are we to make of these shortcomings in the light of the overall success of the model? Obviously, we cannot take it as the final word on development of ability to perform the balance scale task. I would suggest that the model's shortcomings may lie in two places: first, in details of the encoding of inputs and of the network architecture; and second, in the fact that the model only deals with acquisition of implicit knowledge.

Regarding the first point, it would be reasonable to allow the input to encode similarity on each dimension by using input representations in which

each unit responded to a range of similar values so that neighbouring weights and distances produced overlapping input representations; furthermore, the inputs could well make use of a relative code of magnitude to keep values within a fixed range. This would probably overcome the interpolation and extrapolation problems (I have no stand on whether such codings are learned or pre-wired).

These kinds of fixes would not allow the model to truly master Rule 4. This is as it should be, since I believe Rule 4 (unlike the other rules) can only be adhered to strictly as an explicit (arithmetic) rule. Indeed, it must be acknowledged that there is a conscious, verbally accessible component to the problem-solving activity that children and adults engage in when they confront a problem like the balance beam problem. The model does not address this activity itself. However, it is tempting to imagine that the model captures the gradual acquisition mechanisms which establish the possible contents of these conscious processes. One can view the model as making available representations of differing salience as a function of experience; these representations might serve as the raw material used by the more explicit reasoning processes that appear to play a role. This is of course sheer speculation at this point. It will be an important part of the business of my ongoing exploration of cognitive development to make these speculations explicit and testable.

Implications of the balance simulation

The model captures several of the more intriguing aspects of cognitive development. It captures a stage-like character, while at the same time exhibiting an underlying continuity which accounts for gradual change in readiness to move on to the next stage. It captures that fact that behaviour can often seem very much to be under the control of very simple and narrow rules (e.g. Rule 1), yet to exhibit symptoms of gradedness and continuity when tested in different ways. It captures the fact that development, in a large number of different domains, progresses from an initial over-focusing on the most salient dimension of a task or problem—to the point where other dimensions are not even encoded—followed by a sequence of further steps in which the reliance on the initially unattended dimension gradually increases.

As mentioned previously, the model can be seen as implementing the accommodation process that lies at the heart of Piaget's theory of developmental change. Accommodation essentially amounts to adjusting mental structures to reduce the discrepancy between observed events and expectations derived from the existing mental structures. According to Flavell (1963), Piaget stressed the continuity of the accommodation process, in spite of the overtly stage-like character of development, though he never gave a

particularly clear account of how stages arise from continuous learning (see Flavell 1963, pp. 244–9 for a description of one attempt). The model provides such a description: it shows clearly how a continuous accommodation-like process can lead to a stage-like progression in development.

Changes in representation and attention through the course of development

When a balance beam problem is presented to the model, it sees it in different ways, depending on its developmental state. At all times, information is in some sense present in the input for determining what is the correct response. However, at first this information produces no real impression; weak, random activations occur at the hidden level and these make weak, random impressions at the output level. At the beginning of the Rule 1 behavioural phase, the model has learned to represent the relative amount of weight. The pattern of activation over the hidden units captures relative weight, since one unit will be more activated if there is more weight to the right, and the other will be more activated if there is more weight to the left; both units take on intermediate activations when the weights balance. At this point, we can see the model as encoding weight, but not distance, information. Indeed, as we have seen at this point the network could be said to be ignoring the distance cue; it makes little impact on activation, and learning about distance is very slow at this point. At the end of the Rule 1 phase, in spite of its lack of impact on overt behaviour, the network has learned to represent relative distances; at this point it is extremely sensitive to feedback about distance; it is ready to slip over the fairly sharp boundary in performance between Rule 1 and Rule 2. Thus, we can see the Rule 1 stage as one in which overt behaviour fails to mirror a gradual developmental progression that carries the model from extreme unreadiness to learn about distance at the beginning of this phase to a high degree of readiness at the end.

This developmental progression seems to resolve the apparent paradoxical relation between observed stage-like behavioural development and assumed continuity of learning. To me this is the most impressive achievement of the model: it provides a simple, explicit alternative to maturational accounts of stage-like progression in development.

It must be noted, however, that the success of the model depends crucially on its structure. In fact the results are less compelling if either of the following changes are made: (1) if balance is treated as a separate category, rather than being treated as the intermediate case between left-side-down and right-side-down; and (2) if the connections from input to hidden units are not restricted as they are here so that weight is processed separately from distance before the two are combined.

More generally, it is becoming clear that architectural restrictions on connectionist networks are crucial if they are to discover the regularities we humans discover from a limited range of experiences (Denker *et al.* 1987; Rumelhart, in preparation). This observation underscores that fact that the learning principle, in itself, is not the only principle that needs to be taken into account. There probably are additional principles that are exploited by the brain to facilitate learning and generalization. Just what these additional principles are and the extent to which they are domain-specific remains to be understood in more detail.

Extending this observation a step further, we can see the connectionist framework as a new paradigm in which to explore basic questions about the relations of nature and nurture. We may find that successful simulation of developmental processes depends on building in domain-specific constraints in considerable detail; if so this would support a more nativist view of the basis of domain-specific skills. On the other hand, it may turn out that a few other general principles in addition to the learning principle are sufficient to allow us to capture a wide range of developmental phenomena. In this case we would be led toward a much more experience-based description of development. In either case, it seems very likely that connectionist models will help us take a new look at these important basic questions.

Conclusions

The exploration of connectionist models of human cognition and development is still at an early stage. Yet, already, these models have begun to capture a new way of thinking about processing, about learning and, I hope the present paper shows, about development. Several further challenges lie ahead. One of these is to build stronger bridges between what might be called cognitive-level models and our evolving understanding of the details of neuronal computation. Another will be to develop more fully the application of congitive models to higher-level aspects of cognition. The hope is that the attempt to meet these and other challenges will continue to lead to new discoveries about the mechanisms of human thought and the principles that govern their operation and adaptation to experience.

Notes

1. The author would like to thank Eric Jenkins for showing the way toward a connectionist model of learning to perform the balance beam task. Thanks are due as well to Robert Siegler and Dave Klahr for useful discussions. This research was supported by ONR contracts N00014-86-K-0167 and N00014-86-K-0678, as well as NIMH Career Development Award MH00385.

2. In a slightly more general formulation, the net input may be the sum of *products* of the activations of groups of contributing units. In this formulation there is a weight

associated with each product, rather than each individual contributing activation. These product terms have no special computational significance, since the effects of multiplicative interactions among inputs can be accomplished by extra layers of units; see Williams (1986).

3. Some variants of connectionist models (e.g. Grossberg 1978) treat the excitatory and inhibitory inputs as separate forces, rather than aggregating them together in a single term.

4. The Hebb rule is about the simplest connectionist learning rule, and it is limited in what it can do, so it has recently been somewhat less popular than other learning rules (but see Linsker 1986*a*, *b*, and *c*). Three learning rules frequently used in current connectionist models are the *competitive learning rule*, the *delta rule* or *least-mean-squared procedure*, and the *generalized delta rule* or *back-propagation procedure* (see Hinton 1987, for details).

5. This model builds on an earlier model of stage transitions in the balance beam task by Jenkins (1986). I am indebted to Eric for indicating the applicability of connectionist models to cognitive development.

6. An alternative assumption which might account for the developmental data just as well is the assumption that the weight dimension is pre-structured before the child comes to consider balance problems, while the distance dimension is not. The assumption that distance varies less frequently than weight but that neither dimension is initially structured allows us to observe the structuring process for both dimensions.

References

Anderson, J. A. (1977). Neural models with cognitive implications. In *Basic processes in reading perception and comprehension* (eds. D. LaBerge and S. J. Samuels), pp. 27–90. Erlbaum, Hillsdale, NJ.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In *Cognition and the development of language* (ed. J. R. Hayes), pp. 279–362. Wiley & Sons, New York.

Denker, J., Schwartz, D., Wittner, B., Solla, S., Hopfield, J., Howard, R., and Jackel, L. (1987). *Automatic learning, rule extraction, and generalization* (AT&T Bell Labs Technical Report). AT&T Bell Labs., Holmdel, NJ.

Feldman, J. A. (1981). A connectionist model of visual memory. In *Parallel models of associative memory* (eds. G. E. Hinton and J. A. Anderson), pp. 49–81. Erlbaum, Hillsdale, NJ.

Feldman, J. A. (1988). Connectionist representation of concepts. In *Connectionist models and their implications: readings from cognitive science* (eds. D. Waltz and J. A. Feldman), pp. 341–63. Ablex Publishing Corporation, Norwood, NJ.

Feldman, J. A. and Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, **6**, 205–54.

Ferretti, R. P., Butterfield, E. C., Cahn, A., and Kerkman, D. (1985). The classification of children's knowledge: Development on the balance scale and included plane tasks. *Journal of Experimental Child Psychology*, **39**, 131–60.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget*. D. Van Nostrand Company, Inc., Princeton, NJ.

Grossberg, S. (1978). A theory of visual coding, memory, and development. In *Formal theories of visual perception* (eds. E. L. J. Leeuwenberg and H. F. J. M. Buffart). Wiley, NY.

Hinton, G. E. this volume.

Hinton, G. E. (in press). Connectionist learning procedures. *Artificial Intelligence*.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In *Parallel distributed processing: explorations in the microstructure of cognition. Volume I* (eds. D. E. Rumelhart, J. L. McClelland, and the PDP research group). Bradford Books, Cambridge, Mass.

Inhelder, B. and Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. Basic Books. NY.

Jenkins, E. A., Jr (1986). Readiness and learning: a parallel distributed processing model of child performance. Pittsburgh, PA: Carnegie-Mellon University, Psychology Department, 15213.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, Mass.

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

Klahr, D. and Siegler, R. S. (1978). The representation of children's knowledge. In *Advances in child development and behavior* (eds. H. W. Reese and L. P. Lipsitt), pp. 61–116. Academic Press, NY.

Linsker, R. (1986*a*). From basic network principles to neural architecture: emergence of spatial opponent cells. *Proceedings of the National Academy of Sciences USA*, **83**, 7508–12.

Linsker, R. (1986*b*). From basic network principles to neural architecture: emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences USA*, **83**, 8390–4.

Linsker, R. (1986*c*). From basic network principles to neural architecture: emergence of orientation columns. *Proceedings of the National Academy of Sciences USA*, **83**, 8779–83.

McClelland, J. L. (1985). Putting knowledge in its place: a scheme for programming parallel processing structures on the fly. *Cognitive Science*, **9**, 113–46.

McClelland, J. L., Jenkins, E. A., Jr (in preparation). Emergence of stages from incremental learning mechanisms: a connectionist approach to cognitive development. In *Architectures for Intelligence* (ed. K. van Lehn). Erlbaum, Hillsdale, NJ.

McClelland, J. L. and 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.

McClelland, J. L. and Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**, 159–88.

McClelland, J. L., Rumelhart, D. E., and the PDP research group. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*, Volume II. Bradford Books, Cambridge, Mass.

Rescorla, R. A. and 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* (eds. A. H. Black and W. F. Prokasy). Appleton-Century-Crofts, NY.

Rumelhart, D. E. (in preparation). *Generalization and the learning of minimal networks by back propagation.*

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume II (eds. J. L. McClelland, D. E. Rumelhart, and the PDP research group). Bradford Books, Cambridge, Mass.

Rumelhart, D. E. and Norman, D. A. (1982). Simulating a skilled typist: a study of skilled cognitive-motor performance. *Cognitive Science*, **6**, 1–36.

Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986). A framework for parallel distributed processing. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume I (eds. D. E. Rumelhart, J. L. McClelland, and the PDP research group). Bradford Books, Cambridge, Mass.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume I (eds. D. E. Rumelhart, J. L. McClelland, and the PDP research group). Bradford Books, Cambridge, Mass.

Rumelhart, D. E., McClelland, J. L., and the PDP research group. (1986). *Parallel distributed processing: explorations in the microstructure of cognition.* Volume I. Bradford Books, Cambridge, Mass.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume II (eds. J. L. McClelland, D. E. Rumelhart, and the PDP research group). Bradford Books, Cambridge, Mass.

St. John, M. F. and McClelland, J. L. (1988). *Learning and applying contextual constraints in sentence comprehension.* (AIP Technical Report). Carnegie Mellon University, Departments of Computer Science and Psychology, and University of Pittsburgh, Learning Research and Development Center, Pittsburgh, PA.

Seidenberg, M. S., Patterson, K. E., and McClelland, J. L. this volume.

Sejnowski, T. J. and Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145–68.

Shrager, J., Hogg, T., and Huberman, B. A. (1987, May). Observation of phase transitions in spreading activation networks (behavioral change as topology of spreading activation network). *Science*, **236**, 1092.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, **8**, 481–520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, **46** (189), 1–74.

Siegler, R. S. and Klahr, D. (1982). When do children learn? The relationship between existing knowledge and the acquisition of new knowledge. In *Advances in instructional psychology*, Volume 2 (ed. R. Glaser), pp. 121–211. Erlbaum, Hillsdale, NJ.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of

harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume I (eds. D. E. Rumelhart, J. L. McClelland, and the PDP research group). Bradford Books, Cambridge, Mass.

Wilkening, F. and Anderson N. H. (in press). Representation and diagnosis of knowledge structures. In *Contributions to information integration theory* (ed. N. H. Anderson).

Williams, R. J. (1986). The logic of activation functions. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume I (eds. D. E. Rumelhart, J. L. McClelland, and the PDP research group). Bradford Books, Cambridge, Mass.