(1994). In P. Bertelson, P. Eelen, G. d'Ydewalle (Eds.) International Perspectives on Psychological Science, Volume 1: Leading Themes. Erlbaum: Hillsdale, NJ.

4 The Interaction of Nature and Nurture in Development: A Parallel Distributed Processing Perspective

James L. McClelland
Department of Psychology, Carnegie Mellon University, Pittsburgh, USA

Parallel distributed processing (PDP) models provide a rich set of resources for exploring issues of nature, nurture and their interaction in cognition development. I present the essential aspects of the PDP (or connectionist) framework, and I draw parallels between the child as learner and the mechanisms of learning in connectionist systems. The remaining sections discuss some of the implications of this framework for our understanding of the acquisition of knowledge. I point out that many lines of argument that have typically been given in support of nativist approaches need to be reconsidered in the light of the characteristics of PDP models of learning and development. The first of these sections points out that connectionist models offer a dramatic advance over classical associationist approaches to learning. The second illustrates how stage-like progressions can be understood in terms of the typical learning trajectories seen in connectionist models. The third section considers the meaning and possible sources of early competence from a PDP perspective, and the fourth considers how connectionist models may shed light on the fact that some of the structure of human behaviour appears to be imposed by the learner. In all, the chapter amounts to an argument that connectionist models allow us to see ways in which experience might lead to the rich and interesting cognitive structures and developmental progressions that have often been taken as supportive of nativist approaches.

INTRODUCTION

Where do cognitive abilities come from? Are they born in us, innate endowments of nature? Are they products of experience, plain and simple? Or do they arise through the interaction of the characteristics of the organism and the

environment? These questions have stood at the centre of the study of mind for centuries. Some of the most prominent researchers of our century—Chomsky, Skinner and Piaget—have taken each of these views, and each has spawned large followings within psychology and the larger scientific community.

Which position is right? The pure empiricist tradition no longer holds much sway; and there is certainly a very active movement in the field today that favours the nativist position. Yet I think there is a feeling in many quarters that experience must play a larger role than the mere setting of parameters in an otherwise innately predetermined cognitive system.

In this chapter, I will address this question. I will argue for an interactionist position, similar in some ways to Piaget's. My main aim, though, will not be to champion interactionism *per se*. Rather, it will be to suggest that the parallel distributed processing framework (aka the connectionist framework) has broad implications for our understanding of cognitive development. This framework, I will argue, provides mechanisms and ideas that allow us to explore the interplay of nature and nurture, and that suggest how experience may be the engine that drives development, through channels shaped by both innate constraints and the structure of the environment.

I begin with a presentation of the essential aspects of parallel distributed processing that are relevant to the points I hope to make about cognitive development, and I draw parallels between the child as learner/experiencer and the mechanisms of learning in connectionist systems. The remaining sections discuss some of the implications of this framework for our understanding of the representation and acquisition of knowledge.

First, I will point out that connectionist models offer a dramatic advance over other forms of learning, particularly classical associationist approaches. This point is crucial, since it relates quite strongly to the question of what is learnable. Second, I will explore the time-course of development. In particular, I will consider stage-like progressions. Here I will demonstrate how connectionist models implement mechanisms of cognitive change very close to some of the mechanisms proposed by Piaget, and how they can address some of the puzzles that have plagued other experiential accounts of the sources of progress from stage to stage. Third, I will consider the meaning and possible sources of early competence, from a connectionist point of view. Finally, I will consider how connectionist models may shed light on the fact that some of the structure of human behaviour appears to be imposed by the learner.

A few of the points made in the chapter have been made previously in McClelland (1989). Bates and Elman (1993), Karmiloff-Smith (1992a, 1992b) and Plunkett and Sinha (1991) have written on related topics and many of their arguments have contributed to the evolution of the viewpoints expressed here.

THE PDP FRAMEWORK

Parallel distributed processing (PDP) provides us with a framework for thinking about the mechanisms that represent, acquire and use knowledge. The framework is described in detail in the first four chapters of Rumelhart, McClelland and the PDP Research Group (1986c). Here I touch only on points relevant to the present discussion.

First, PDP assumes that cognitive processes arise from the interactions of large numbers of simple processing units, organised into modules. A very generic example of such a network is shown in Fig. 4.1. Within each module, each unit computes a simple function of the inputs it receives from other units. Crucial to us will be the fact that this is a continuous but non-linear function, like the one shown in Fig. 4.1. Second, the PDP approach assumes that the knowledge that governs processing is stored in the strengths of the connections among the units. Such knowledge allows the pattern present on one set of units to give rise to other patterns on other sets of units; or for the pattern of activation at one point in time to give rise to a successor at the next moment. Third, the PDP approach assumes that acquisition of knowledge occurs through the adjustment of connection strengths. These adjustments, in turn, are driven by signals arising ultimately from external inputs to the network. I discuss this aspect more fully below.

The Role of the Innate Endowment

Let us now consider how this framework allows us to explore the interaction of innate endowment and experience as determinants of the course and outcome of knowledge acquisition. The innate endowment has several aspects. Among them is the gross architecture of the system—the modules, the number of processing units in each, and their initial connectivity (Rumelhart, Hinton, & McClelland,

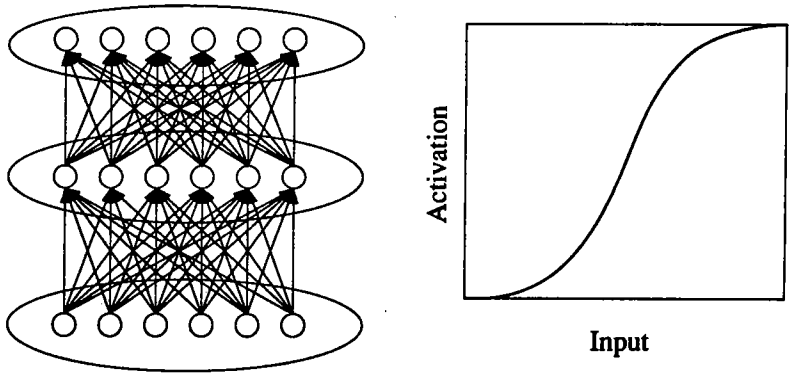

FIG. 4.1 A generic connectionist network, together with its activation function.

1986a). We know of course that the brain is not at all fully connected, and the organisation of this connectivity determines which units will be in line to receive relatively direct auditory input, which to receive visual input, which to combine these two sources of information, and so on.

A second aspect of the innate endowment concerns the basic rules that govern the propagation of activation and the adjustment of connection strengths. These rules, to the extent that they differ in different parts of the system, provide a means whereby modality- or domain-specific constraints can be imposed on the cognitive system. To the extent that they are general throughout the system, they provide a set of pervasive common mechanisms for processing and learning.

A third aspect of the innate endowment concerns the detailed parameters of different parts of the system. Even when the basic rules of processing and connection strength are held constant, differences in parameters can lead to mechanisms with very different characteristics. As one example, O'Reilly (1992) took a network (illustrated in Fig. 4.2) consisting of two sets of internal units, each receiving input from the same set of inputs, and each sending output to all of the output units. The internal units in the two pools differed only in the rate at which their activations changed over time. Those on the left changed activation slowly when their inputs changed, while those on the right changed their activation very rapidly in response to changing inputs. The network was asked

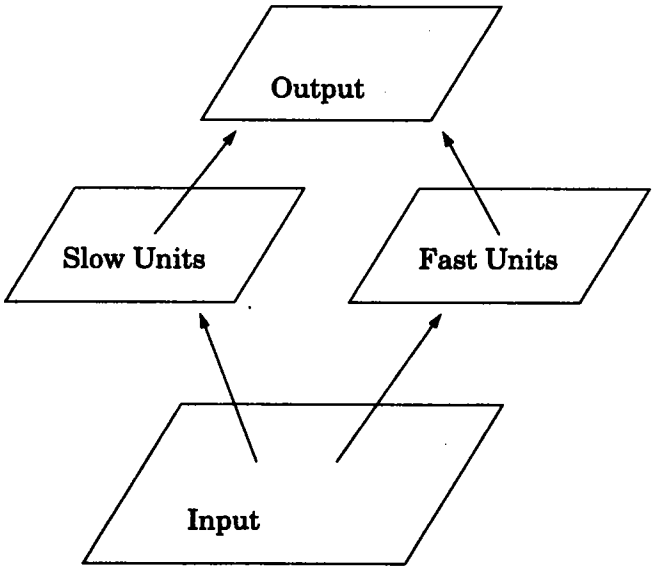

FIG. 4.2 O'Reilly's network. The slow units in the module on the left become "what" units and the fast units in the module on the right become "where" units.

to perform the task of learning to indicate the location and identity of input patterns. Some of the output units were used to specify identity, and the others were used to specify shape, with both sets of hidden units connected to all of the output units. Each input pattern arose at a discrete point in time and then persisted for several time steps, moving about from place to place within the input array, as if the network were receiving input at different retinal locations over a series of eye fixations. Thus, the identity of a pattern tended to persist for several time steps, while the location of the pattern moved around. The result was that the pool of units with the slow rate of change became specialised for detecting the identity of the object, while the units with the faster rate of change became specialised for detecting its location. This is but one example of a burgeoning body of work showing how modules with slightly different initial parameters or biases can become differentially specialised through experience. This particular case also emphasises the joint roles of environment and innate structure. It is the combination of the fact that object identity remains the same as the retinal location changes with the fact that the units in different modules differ in their temporal parameters that determines the outcome in this case.

The Role of Experience

Now let us focus on the role of experience. Here I will begin by proposing a general way of thinking about how experience might propel cognitive development, and then I will relate this idea back to parallel distributed processing (McClelland, 1989). We begin by thinking of the child experiencing an environment involving an ongoing sequence of events. We assume that the child, while alert and attending, is always making implicit predictions for what might happen next. Let me emphasise the implicit nature of the process. I don't mean that the child is consciously aware of asking, and posing answers to the question, "What's coming next? What's coming next?" I do mean, though, that his or her cognitive system is in fact anticipating the future, and that a reaction can occur if these expectations are violated. Behaviourally, such reactions are accompanied by orienting responses—eye movements, pupil dilation, increases in gaze duration, etc. They also, at least in language processing, generate large and robust evoked potentials, such as the N400 of Kutas and Hillyard (1980). Note that matching or mismatching is always a matter of degree in PDP systems—it is a matter of the consistency between one pattern of activation and another—and in general actual events will always differ to some degree from what is anticipated.

In any case, as events occur, they permit the child to compare implicit predictions with what actually happens next, and thereby to learn to make better predictions for the future. The essential assumption is that the child uses the following procedure to adjust the predictions:

> Adjust each parameter of the mind in proportion to the extent that its adjustment will reduce the discrepancy between predicted and observed events.

The most straightforward and typical way of relating this idea to connectionist networks is as follows (see, e.g. Elman, 1990): First, we imagine that the input to the system represents the current situation, from which a prediction will be made. This input may be signals arising from outside the system, or the system's own internal state. Second, we imagine that the output represents the prediction that the system makes for what may happen next. Finally, we imagine that the actual next event arises as a pattern of activity on these units. The rule for learning becomes:

> Adjust each connection weight in the network in proportion to the extent that its adjustment will reduce the difference between the output of the network and the actual next event.

The algorithm standardly used for adjusting connection strengths, the back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986b), is simply an algorithm for calculating the relevant quantities—in particular, the extent to which an adjustment to each weight will reduce a measure of the differences across the output units.

Given such a learning rule, experience drives changes in connection weights, thereby causing the network to learn how to represent and use inputs to make better predictions for the future. Just this approach has been used to train networks to do a number of very important and interesting things, which we will turn to in a moment. For now, though, I want to highlight briefly three key features of the connectionist approach to knowledge and knowledge acquisition.

Properties of Connectionist Knowledge

First, the knowledge is implicit, in the very sense that this term has been used in the developmental literature by Karmiloff-Smith (1986, 1991, 1992a), and in the memory literature by a number of investigators (see Schacter, 1987). It is knowledge that is embedded in specific processing mechanisms. It is knowledge that can be used to make predictions, perceptual anticipations and pattern completions, and to subserve the gradual acquisition of domain-specific skills. Violations of predictions made on the basis of such knowledge can serve as the basis of making various kinds of judgements, as when a person is asked to make a judgement about the grammaticality of a potential example of a natural or artificial language (Knowlton, Ramus, & Squire, 1992; Reber, 1976, 1989). But the knowledge itself is not directly available as such, either for report, for explicit reasoning, or any other purpose.

Second, the knowledge is inherently graded in nature. This is a most important point, and one that will pervade much of the rest of this chapter. When a network is first placed in an environment, its connection weights might be initialised to small, random values, within broad connectivity constraints. Its predictions, then, might be initially weak and random. As experiences occur, predictions gradually become attuned to situations, and after more and more time has passed, the network becomes structured in accordance with the contents of the domain. At first we would say the network knows nothing, and at the end we would say the network knows the structure of the domain, but in the middle our ordinary terminology and ways of thinking about what it means to "know" something break down. The network has a kind of partial knowledge—it generates predictions that incline towards being correct, or are correct only in certain cases, or that capture the gross regularities without capturing the subtler details. The process is completely gradual and there is no special discrete point at which we would say the network now knows, and before this, it did not know.

Third, acquisition is gradual and incremental. Gradualness is important, as we will see in our consideration of an example below, because it allows the overall direction of change to be determined by the overall structure in the domain that it is learning rather than specific individual stimuli. It is incremental, in that each change builds on earlier changes. As in Piaget's developmental theory, where cognitive structure emerges from accommodation and assimilation, each new adaptation is but a slight variation and extension of the system that resulted from previous acts of adaptation (see Flavell, 1963, for discussion).

Given the implicit nature of the knowledge embodied in PDP systems, it may seem to many researchers at first glance that PDP systems have relatively little to offer any effort to understand cognitive development. This will be especially true for researchers who think of "real" knowledge as explicit knowledge, or who think of discontinuous change, such as insight, as the hallmark of development. There is certainly no denying that explicit knowledge exists, or that the development of explicit knowledge is a major aspect of cognitive development. But I believe that implicit knowledge, of the kind that is built into connectionist systems through experience, is more fundamental, in that it structures the very representations of experience itself that provide the input to explicit thought processes. Certainly, we all appreciate from linguistics the subtlety that implicit knowledge can have, and the role that it can play in shaping our representations of linguistic stimuli. I share the view of Karmiloff-Smith (1992a) that such implicit knowledge is as important in other domains, and serves, as it does for language, as the substrate on which explicit cognition is ultimately built. On this view, our understanding of explicit cognition will be enriched by a deeper understanding of this substrate. Therefore, the rest of this chapter will focus on the implications of PDP for thinking about the nature and acquisition of knowledge of this implicit kind.

IMPLICATIONS OF THE PDP APPROACH

My main point in this chapter is to show that the PDP approach has profound implications for our understanding of development. In general, my argument is that with the tools provided by the PDP framework we can call into question some of the tenets of nativist approaches on the one hand, and breathe new life into the interactionist position on the other.

The rest of this chapter will pursue this point through four specific arguments. First, I will show that connectionist learning rules allow us to extend the scope of what is learnable far beyond what was thought possible under earlier approaches, thereby reducing the need many authors have seen to build specific knowledge in from the start. Second, I will argue that PDP allows us to understand stage-like progressions in development more clearly than was possible under Piaget's interactionist account. Third, I will point out how PDP provides us with a language in which to rethink the meaning of many of the interesting phenomena of early competence in infants. Fourth, I will suggest how PDP allows us to understand how a learning process may structure the domain that is being learned.

The chapter does not lay out its own detailed theory of development, nor does it even use the PDP approach to make specific predictions. Rather, it uses example simulations to demonstrate specific points that contrast with viewpoints often taken in more nativist approaches. Taken together, the examples support the more general claim that PDP can lead us to rethink much of the evidence that has been taken in support of the view that much essential content knowledge is innate and that experience plays only an elaborating or parameter-setting role in development.

Extending the Domain of the Learnable

In order to bring out the way in which connectionist learning allows us to extend the domain of the learnable, and to underscore its relevance to basic questions about the sources of our conceptual knowledge, I will focus on an example used by Keil (1987) in his research on conceptual development. Keil considers how children come to shift their categorisation behaviour from an early reliance on an ensemble of characteristic features to a later reliance on core relational properties. Among the examples he uses is the use of kinship terms such as "uncle". He reports that children shift from an initial acceptance of descriptions such as "a fellow who was not related to anyone in your family but was a big pal of your dad's and brought you presents on your birthday and Christmas" to a willingness to accept descriptions such as "a two year old who was your mom's brother". Keil stresses that the developmental pattern indicates that the child's conceptual development reflects the emergence of a "relational system" relevant to the specific domain of knowledge, rather than piecemeal acquisition of knowledge about, say, uncles but not aunts, or any kind of domain general change in overall approach.

Although he acknowledges that experience plays a role in these developments, Keil argues in several places (1981, 1987, 1991a, 1991b) for the idea that the ability to use experience in the service of the construction of such relational systems of meaning depends on the use of *a priori*, domain-specific knowledge. As part of his argument he points out a challenge for any learning-based account that does not rely on such *a priori* knowledge. He argues that classical associationist accounts—accounts he takes as paradigmatic of experience-based approaches to knowledge acquisition—appear to lack the crucial ability to shift from a reliance on merely correlated features to a reliance on deeper relational properties. The reasons for this, he argues, lie in the basis on which the learning mechanisms provided by such accounts learn and use what they have learned. Keil (1991a) notes two key properties of associationist accounts: learning occurs by contiguity (for example, contiguity of a situation with an outcome, or of a current event with the next event) and generalisation occurs by similarity, permitting responding to novel inputs because of their similarity to familiar cases. Thus, for example, I associate the sight of a rose with the smell of a rose by contiguity, learning to predict the smell from the visual appearance. I know that another rose will smell as sweet because it looks similar to the first rose.

In essence, Keil's argument is that children—and certainly adults—do not always generalise on the basis of similarity. Rather, they come to rely, sometimes almost exclusively, on deeper, relational information. Lacking an experience-based alternative to associationism that overcomes this limitation, he points to the work of Spelke and others (e.g. Spelke, 1991; Spelke, Breinlinger, Macomber, & Jacobson, 1992) for evidence of very early (and therefore putatively innate) knowledge in a number of domains, and he suggests that children may have a handful of innate "proto-theories" that allow them to determine just what the appropriate bases are or generalisation, in each of several domains.

Now let us take a look at this argument, in light of PDP models, and see how it holds up. At first glance, it seems that PDP models may not change things much, since these models do learn from contiguity of situations and outcomes, as I have already explained. Furthermore, it is also true that they tend to generalise by similarity: similar patterns of activation on one layer of units tend to produce similar patterns of activation at the next. This may be why so many cognitive scientists and philosophers with otherwise widely divergent views essentially dismiss connectionist models of learning. Putnam has stated, for example, that "things that cannot be expressed in terms of correlations between hard-coded variables cannot be found by the algorithm", and are not able to recover the structure that lies behind the surface. Fodor has stated, "The question is whether you can get some success that you could not get by just doing statistics . . . I think of them [connectionist models] as analog statistics packages."[1]

[1] The quotations from Fodor and Putnam have been used with permission of the authors (personal communication, 1992). The statements were originally made in interviews with P. Baumgartner, who conveyed these quotations to me for comment.

But there is something very wrong with all these arguments. Although connectionist models are similar in some ways to associationist models, and although they are extensions of existing statistical techniques, there is a crucial difference. In connectionist models, the very basis of determining what is similar to what is part of what is discovered in the course of learning. This is not just a matter of emphasising some dimensions and de-emphasising others (although this is what happens in some cases), but of *totally recoding each input so as to map it into a similarity space whose structure depends*, not on the surface properties of the inputs, but *on the demands that are made by the task of predicting outcomes from inputs*. This sensitivity of the internal representations to the task they are required to perform dramatically increases the range of what such models are able to achieve compared to associationist models or standard statistical techniques.

This point has been illustrated many times in connectionist learning research. The importance of constructing representations was the theme of Rumelhart et al. (1986b) and almost every interesting application of connectionist learning since their paper has made essentially this point. The example I have chosen to use here is one of the first to demonstrate this point, by Hinton (1986). It happens to use an example that relates specifically to Keil's kinship examples, as is therefore particularly appropriate. I spend some time reviewing this example here because its point appears to have been lost on many very important researchers and commentators.

Hinton trained a connectionist network to answer questions about the kinship relations among two groups of individuals, one called English and the other called Italian (Fig. 4.3). Within each group, there were parallel sets of kinship relations, but there were no relations between the groups. Hinton trained a connectionist network on the relations among these individuals in the following way. The task of the network was to take queries that can be glossed "Person1's Relation is _____" (e.g. "Colin's Uncle is _____") and to then respond with the correct completion or completions of the proposition (Person2). As shown in Fig. 4.4, he set up two pools of input units, one consisting of a unit for each possible filter of the Person1 role, and the other consisting of a unit for each of the kinship relations. This allowed him to present queries by simply activating the appropriate Person1 unit and the appropriate Relation unit. He also set up a set of output units, for Person2, again consisting of one unit for each person, to allow the network to activate a unit corresponding to the individual or individuals at the other end of the relation. In between there were separate groups of hidden units, one for the network to use for an internal representation of P1, one to represent the relation, one to represent the combination of P1 and the relation, and one to represent P2.

The important point is that, at the input and output levels, the persons and the relations are each represented by distinct input units. Initially, the connections from the input units to the corresponding representation units are random, so

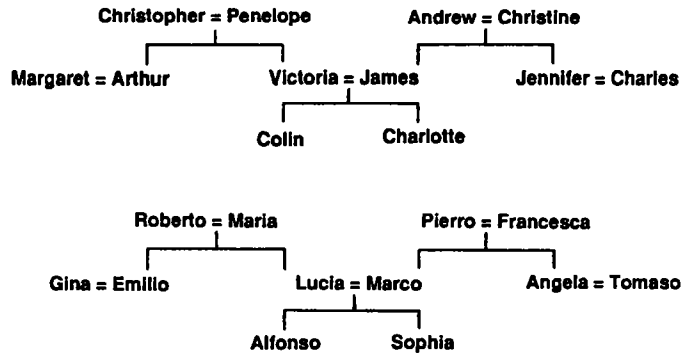

FIG. 4.3 The family trees underlying the training corpus used by Hinton (1989). The symbol "=" means "married to". From *Parallel Distributed Processing: Implications for psychology and neurobiology* (p. 49), by R.G.M. Morris. New York: Oxford University Press. Copyright 1989 R.G.M. Morris. Reproduced by permission.

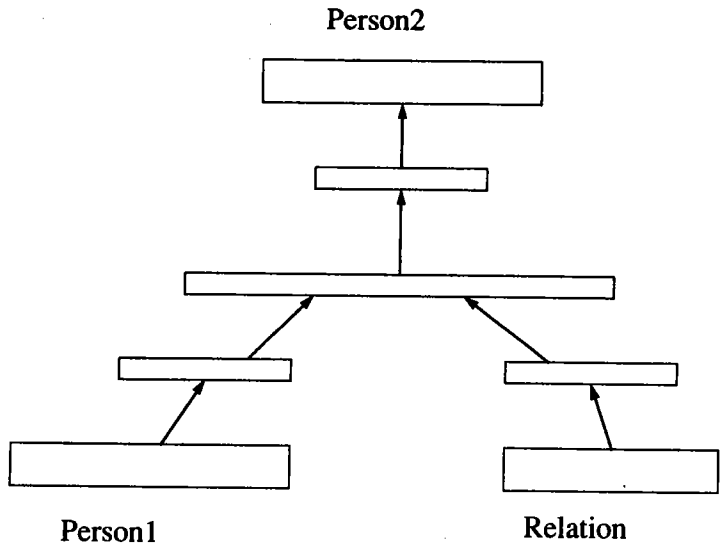

FIG. 4.4 The network used by Hinton (1986) to learn the family trees shown in Fig. 4.3. Reproduced by permission.

that the similarity relations among the initial internal representations are arbitrary and unrelated to the task. During the course of training, the connection weights throughout the network change. On each training trial, the network uses its

existing connection weights to predict the correct completion or completions; the output is compared to the correct response; and the connections are adjusted to reduce the difference using back-propagation. The weights from the Person1 input units to the Person1 representation units determine how each person is represented. The representations capture what might be called a theory of the underlying structure behind the set of specific propositions on which the network was trained. They represent as similar those individuals who play similar roles, within and across the two family trees shown in Fig. 4.3.

We can try to visualise the discovered similarity structure[2] by imagining the patterns of activation on these person-representation units as points in a six-dimensional space (where each dimension corresponds to the activation of one of the six units). Since such spaces are difficult to visualise, and since units in the network tend to be used redundantly anyway, we can use principal components analysis to pick out the important dimensions. When we do this, we find that there are three strong components. We can then plot the individuals in a three-dimensional space, representing where each individual falls with respect to each of these three components. The three-dimensional structure is still somewhat difficult to grasp in the projected view presented in Fig. 4.5, but it can be brought out by using the fact that the representations of the individuals

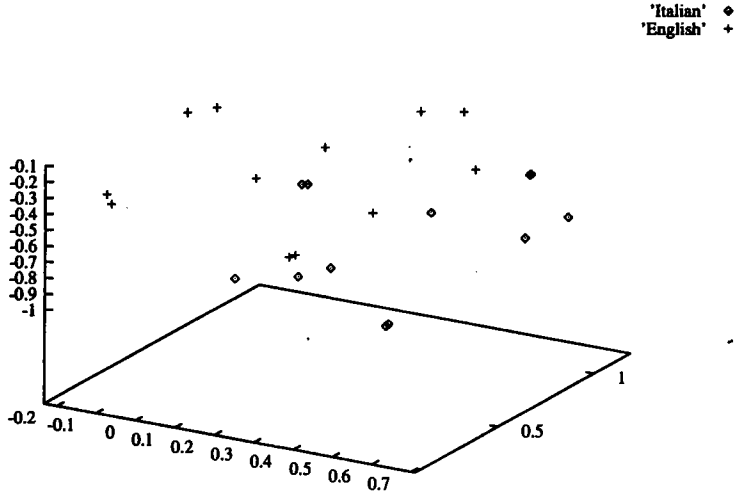

FIG. 4.5 The internal representations of the individuals in the two family trees, displayed as points in a multidimensional space.

[2] The patterns used for the analysis discussed in this and the following paragraphs were kindly supplied by Geoff Hinton, who re-ran the network described in his 1986 paper to regenerate the data.

can be separated perfectly by a plane through the space shown in Fig. 4.5 that leaves all of the Italians on one side of the plane, and all of the English on the other. The dimension perpendicular to this plane can be called the "nationality dimension". The network discovers this dimension because it plays a very strong role in restricting the possible completions of a query, since Person1 and Person2 always have the same nationality.

We can then look at the projection of the English and the Italian groups onto the separating plane, and what we see is shown in Fig. 4.6. For both the English and the Italians, the older generation (squares) is at the top of the plane, the younger generation (X's) at the bottom, and the middle generation (triangles) is in between. Also, for both the English and Italians, the individuals in the left branches of the trees shown in Fig. 4.3 (open symbols) are located to the left of the plane, and the individuals in the right branches in Fig. 4.3 (filled symbols) are located to the right. In general, within generations, corresponding individuals

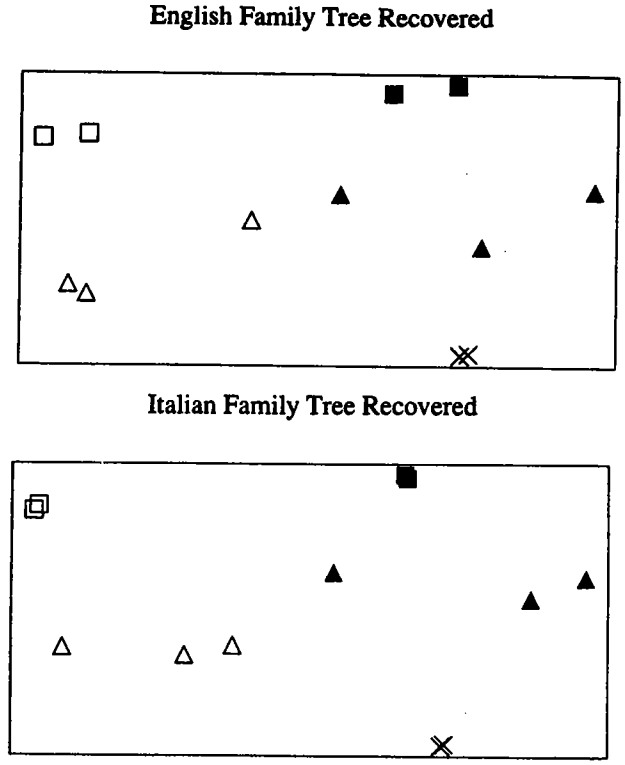

FIG. 4.6 Projections of the representations shown in Fig. 4.5 onto the plane that separates the English from the Italians. See text for explanation.

in the two trees occupy corresponding locations, with only slight differences between the position of each English and the corresponding Italian.

The point of reviewing this demonstration is to make clear that the network has learned to treat as similar, both within and between the two family trees, those individuals who occupy the most similar positions in each tree. The reason individuals at similiar locations in the tree are given similar representations is that these representations are useful for allowing the network to solve the problem posed by the query-completion task, since the completions of propositions involving individuals located at similar places in the family trees occupy similar tree positions.

Given the use of this basis for assigning similarity to the Person1 inputs, the network can generalise properly to cases that were not presented in training. In this case, eight of the relations were held back from the training corpus and only appeared during testing. The network used the representations it had derived for each individual during the course of training to generalise properly at the time of test. It was able to do this in a very simple way. The nationality dimension of the representation of Person1 determined the nationality of the response term. Otherwise, the representations of corresponding individuals in the different family trees were identical. Therefore, what it learned about an individual in one tree transferred to individuals in the other.

In short, the network learns by contiguity and generalises by similarity, but it differs in a crucial respect from associationist models and standard statistical techniques. Unlike these mechanisms, it assigns the similarity relations among the patterns as part of what it learns through the use of a simple gradual and incremental procedure for adjusting the strengths of the connections among units.

One of the key features of what happens in the process of assigning representations is the creation of categories—groups of patterns containing similar members—and distinctive features that differentiate members of different categories. As in the family trees example, such categories support generalisation. Elman (1990) considered this issue as it relates to syntax and St. John and McClelland (1990; McClelland, St. John, & Taraban, 1989) considered this matter in a connectionist model of sentence comprehension. Human language users can parse unfamiliar but grammatical sentences because they know the categories and subcategories of nouns and verbs. So can connectionist networks that discover these categories through gradual learning. In the St. John and McClelland case, as in Hinton's network, it is the discovered similarity structure, based on the roles each word plays in constraining the interpretation of the meaning of various sentences, that governs generalisation, rather than any surface similarity among particular input patterns.

The fact the connectionist models discover their own internal representations, and the fact that these representations can serve as a principled basis for non-trivial generalisations, appears to have been ignored in discussions, such as the one by Keil (1991a), where the point is to motivate a nativist approach by appealing to the insufficiency of experience-based approaches. The fact that associationist models, and not connectionist models, are discussed by Keil, simply indicates that the thinking underlying nativist approaches is based in part on an insufficient appreciation of the power of what can be accomplished by the newer and more powerful mechanisms of learning that are available in the connectionist framework.

I have chosen to focus on Keil's presentation of these issues because of his clarity in laying his position out and because of the specific relevance of Hinton's demonstration to his analysis of the acquisition of kinship knowledge. But the point is a very general one, and I believe that there has been a failure in the work of other key authors (among them Fodor and Putnam) to attend to the differences between connectionist learning mechanisms and pre-existing ideas such as associationist models or standard statistical techniques. This is unfortunate on two grounds. One is the point already made, that connectionist models are far more powerful than associationist models or standard statistical techniques. The other is that the use of connectionist models does not entail the claim often attributed to associationists that the organism starts as a *tabula rasa* with nothing more than domain-independent principles to guide it. On the contrary, as I have already indicated, it is quite easy to incorporate domain-specific variation. This allows an exploration of the full range of possibilities, ranging from pure *tabula rasa* formulations, to strongly pre-structured accounts. The appeal of connectionist models is not that they commit the researcher to one formulation or the other, but that they provide the opportunity to explore a wide range of possible accounts of the roles of nature and nurture in development. While it seems to me we can take the Hinton work as raising questions about the need for innate domain-specific knowledge of relevance to such things as kinship relations, I would not want to suggest that it proves the sufficiency of a blanket domain-general approach. After all, there is considerable pre-structuring of Hinton's architecture for the discovery of the relevant relational information. The extent to which this pre-structuring is crucial is not yet known.

The Time-course of Developmental Change

Another domain now available for exploration is the study of the time-course of change. In this domain, one question we may consider is, why is development so slow? As Flavell (1963, p. 49) put in his book about Piaget, "What prevents the organism from mastering in one fell swoop all that is cognizable in a given terrain?" The same question can be asked of connectionist networks. Here, in cases like the Hinton example, we can see one reason why learning is gradual. The reason is that the network must integrate the changes it makes to its connections over a sufficient sample of the environment, so that the changes can be guided by the structure that is present in it. It is only when the changes occur slowly enough that their overall direction is governed by the structure of the environment that the ability to represent that structure can emerge.

But now a puzzle seems to arise. If development is a slow and gradual process, why is it that it often appears to be marked by long periods of statis, punctuated by brief periods of relatively rapid change? What accounts, in Flavell's words, for its "velocity and acceleration" (Flavell, 1963, p. 49)? How is it, more specifically, that incremental mechanisms give rise to apparently qualitative change?

It is crucial to distinguish between the sense of stages intended in this chapter and the sense of stages that comes from the developmental theory of Piaget. In this chapter, stages are only meant to refer to periods of relative statis within domains, and stage transitions are meant as references to the transitions between these domain-specific features of the developmental profile. This conception of stages contrasts sharply with the Piagetian conception of broad stages cutting across all cognitive domains. It may be that Piaget's insistence on the centrality of these broad stages contributed to the difficulty in understanding how his views of the sources of cognitive change—accommodation and assimilation—could give rise to cognitive development.

In any case, the question of how and why domain-specific stage transitions occur remains puzzling for experience-based theories of all types. If we take an incremental view, it is not at all obvious why we should see apparent stability for long periods at all, especially if, as often happens, these long periods of stability are followed by relatively abrupt transitions to a new plateau.

One response to stage transitions has been to argue that they do not really exist—that they represent, perhaps, the crudeness of our methods for assessing childen's capacities at particular points in development. Another approach has been to think of them as arising from some maturational process. On this view, the child progresses from stage to stage based on some exogenous process, much as a butterfly progresses from stage to stage. Both of these kinds of things may be part of the story. In addition, however, insights from the study of learning in connectionist models may also shed some light on the matter.

I examined this issue in a study of developmental change in the balance scale task introduced by Inhelder and Piaget (1958), and subsequently studied extensively by Siegler (e.g. 1976, 1981) and many others. The task, quite simply, is to look at balance scale problems of the kind shown in Fig. 4.7, and to indicate which side will go down. The specific problem consists of providing some number of unit weights on each side, with the weights all on one peg, some number of steps from the fulcrum. Siegler identified several stages in the development of the ability to solve this kind of problem. In what I will call stage 0 (typical of children below the age of 5), he found that children tended to perform randomly in the task. In stage 1, children responded to the number of weights on each side, but ignored distance from the fulcrum. In stage 2, children took distance into account only when the weights on both sides were the same. In stage 3, they took both distance and weight into account, but behaved inconsistently when these two cues were placed in conflict. Stage 4, which is not always achieved even by

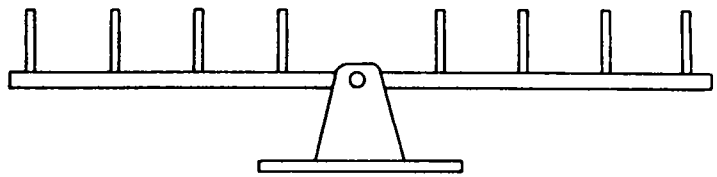

FIG. 4.7 A balance scale of the kind studied by Siegler (1976, 1981). Reproduced by permission, Academic Press, Inc.

adults, requires multiplying weight times distance when weight is greater on one side and the distance is greater on the other. The main interest for now is in the transitions through the early stages, particularly into and out of stage 1, and in the fact that children typically stay in stage 1 for several years. The question is, how is it that this stage-like character of development can arise from an incremental learning mechanism?

In fact, this kind of stage-like progression is exactly what we typically see in connectionist learning models. The phenomenon requires a multi-layer network—one consisting of at least one layer of hidden units between input and output. In the simplest problems that require a two-layer architecture—like the exclusive–or problem, shown in Fig. 4.8—there are two stages. In this problem, it is standard to use a network consisting of two input units, two hidden units and one output, and the task the network faces is to learn that the output unit should be on if either input unit is on, but should be off if neither or both are on. In training such networks with multiple sweeps through the set of training examples, we see an early stage in which learning progresses very slowly, and the network is performing very poorly, producing the same response in every case—an activation of 0.5. There then follows a rapid transition, in which performance rapidly improves, until it levels off at a high level of accuracy, in which the output of the network approaches the right answer for all four training cases. In more complex cases, where some of the training examples are easier to master than others, there can be multiple plateaus, again punctuated by relatively rapid transitions.

The behaviour in the simplest cases can be understood in the following way. Early on, both input and output connections in the network are random. A change to a connection weight on the input side of the network has no useful effect, since the output connections are random; and a change to a connection weight on the output side of the network has no useful effect, since the input connections are random. Gradually, though, through the course of experience, the connection weights begin to become coordinated. The input weights begin to produce useful representations, and the output weights begin to produce useful predictions based on these representations. As this happens, it turns out that small changes to the weights begin to produce bigger and bigger changes in performance.

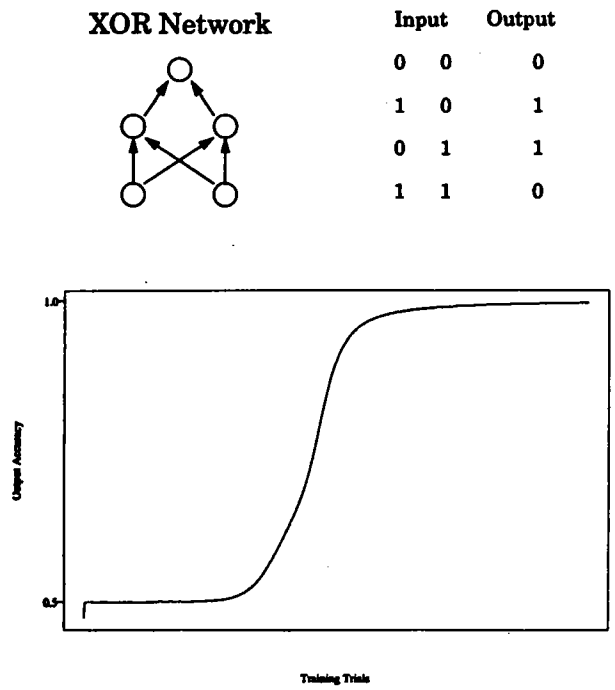

FIG. 4.8 XOR: The network, the input–output patterns and the time-course of acquisition.

While it seems like no progress is being made at all at first, in fact there is some progress all along. The early phases produce only very small changes in overt performance, but they are necessary to bring the network to the point where it is capable of more rapid and noticeable progress. Later rapid change depends upon the earlier, sometimes almost infinitesimally small adaptations.

In applying these ideas in an effort to understand development in the balance scale problem, we can envision a progression in which the kind of transition we see in the XOR network happens, first for one cue—the weight cue—and then for the distance cue. Indeed, it is possible to simulate the progression of stages in just this way (McClelland, 1989; McClelland & Jenkins, 1991; Schmidt & Schultz, 1992). In the simulations reported in McClelland (1989), I set up a network in which the inputs were a representation of a balance scale problem, indicating some number of weights some number of steps from the fulcrum on each side (Fig. 4.9). The task was simply to turn on one of two outputs, indicating which side should go down. The network was trained on a random sequence of balance scale problems, and was tested with the same problems that Siegler (1981)

used to assess children's performance at each point along the way. I ran the network several times, using different starting weights and different random sequences of training trials, and found that it captured several aspects of children's performance very nicely. In general, I found that the simulation conformed to one of Siegler's stages about 86% of the time—in children, the figure is around 93%. More gradual learning produces a higher level of conformity (Schmidt & Schultz, 1992). The network always progressed from stage 0 to stage 1, and then after a period of stability it exhibited a noisy transition through stage 2 to a stage representing something intermediate between Siegler's stages 3 and 4. In so doing, it captured a number of aspects of the data, including some of the features that indicate that children may use a continuous function, rather than discrete rules, in making their judgements in the balance scale task (Wilkening & Anderson, 1991; see Schmidt & Schultz, 1992, and McClelland & Jenkins, 1991, for further discussion).

I did find that in order to simulate the results, it was necessary to make two assumptions. First, I had to assume that weight and distance were encoded separately, as shown in the simple network, and combined only after separate encoding. Second, I had to provide some basis for the use of the distance cue to develop more slowly than the use of the weight cue. I got the results just shown using the assumption that weight varies more often in the relevant experience

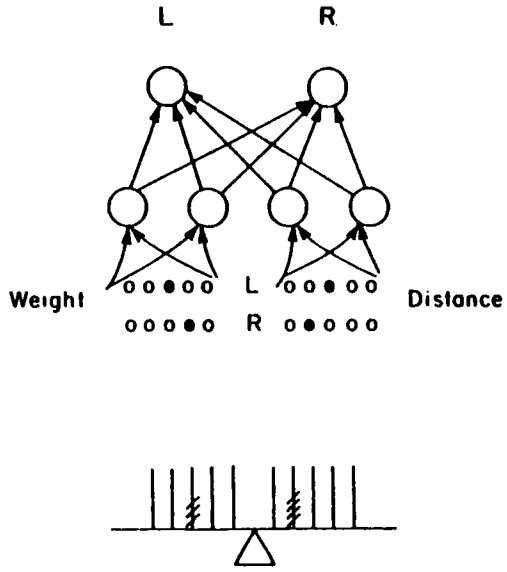

FIG. 4.9 The connectionist network used to study stage transitions by McClelland (1989). Reproduced by permission, Oxford University Press, New York.

of the child than distance does—in the simulations, this amounts to using more frequent presentations of cases in which weight differs on the two sides of the scale but distance is the same, by analogy to children's experience with see-saws, where distance from the fulcrum is relatively fixed by the situation. More recently, I have examined the idea that distance is a more complex relation than weight. Weight is (at least in a fixed gravitational field) a one-place predicate—a property of an individual object. In contrast, distance between two objects is a two-place predicate, a relation between an object and another object—in this case, the weight and the fulcrum. To capture this idea, I used a more complex input, in which the position of the weights and of the fulcrum could vary (McClelland, in press). This made it necessary for the network to learn to compute the more complex distance cue from the simpler position cues. In that case, even if weight and distance vary equally often in the training examples, the network learns the weight cue much more quickly, and there is a long phase that corresponds to Siegler's rule 1, followed eventually by a transition to a stage like Siegler's rule 3, in which weight and distance are both considered.

In the work with the simpler representation (McClelland, 1989), it is easy to examine the time-course of learning about the weight and distance cues, in terms of the strengths of the connection weights into and out of the hidden units (Fig. 4.10). For both cues, we see roughly the same pattern. Initially, the network is insensitive to the cue at both levels. Then there is a relatively steep transition, followed by a levelling off. The transition is more rapid for the weight cue than the distance cue, because of differential frequency of exposure to variations in the two cues. There are actually many empirical results supporting the idea that the acquisition of the distance cue is relatively gradual in children, as it is in the model (e.g. Wilkening & Anderson, 1991). The key point though is the non-homogeneity of the developmental process, both for the weight cue and the distance cue. In both cases, there is an early phase of no overt change, followed by a more rapid transition to a new level.

Of course, this simulation vastly oversimplifies the challenges a learner faces during development. Children do not learn about balance scales in isolation from other kinds of experience with distance and weight information. Distances and weights do not come nicely prepackaged in unit quantities, as they do in the input to the network. Nevertheless, I think the model does capture one important aspect of development, namely its stage-like character in many different domains.

In summary, then, we see that connectionist models can shed some light on the time-course of development. The work does not treat the process as one that arises strictly from environmental influences. Indeed, the model illustrates the general importance of architectural constraints, since it is clear that the success of the model depends on the separate encoding of weight and distance information. Experience is not the only factor that plays a role in shaping development. But experience is the engine that drives development forward, and the work makes it clear how stage-like behaviour can emerge from incremental learning mechanisms.

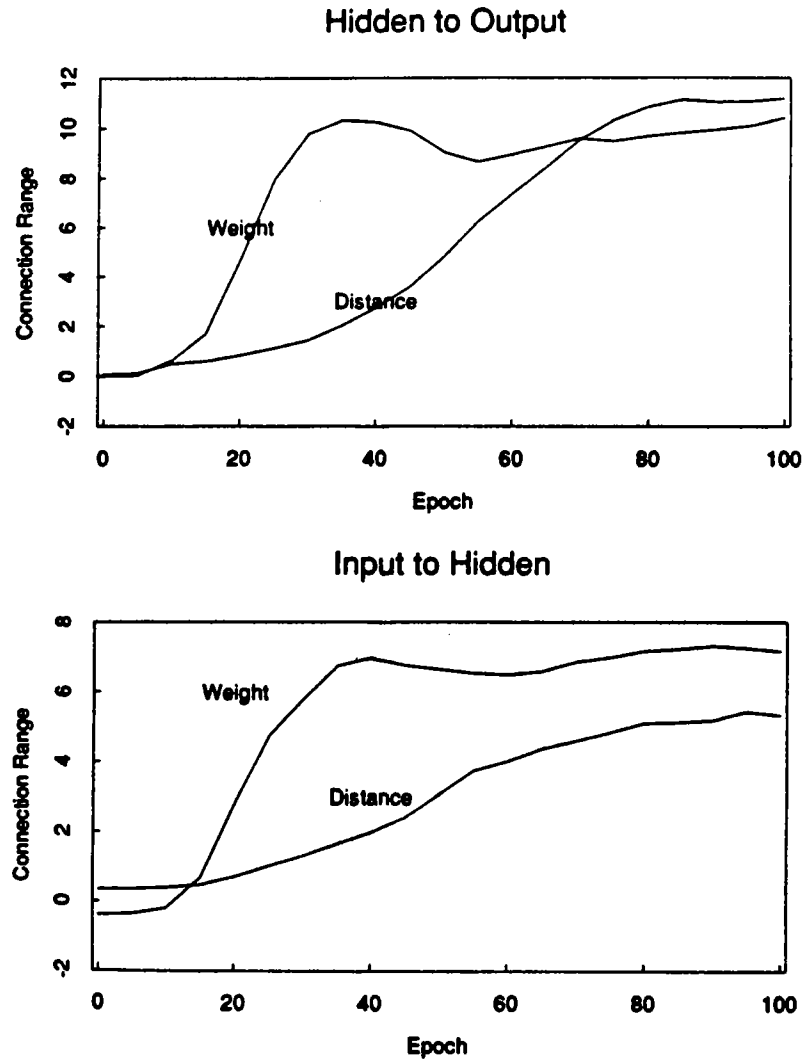

FIG. 14.10 Time-course of acquisition of connection strengths relevant to the role of weight and distance in the balance scale simulation of McClelland (1989). Reproduced by permission, Oxford University Press, New York.

It is worth stopping for a moment to note how much the connectionist approach advances our understanding of the mechanisms of developmental change, relative to proposals advanced, for example, by Piaget. Piaget presented his thinking in