Emergent Mechanism

The mechanist is intimately convinced that a precise knowledge of the chemical constitution, structure, and properties of the various organelles of a cell will solve biological problems. This will come in a few centuries. For the time being, the biologist has to face such concepts as orienting forces or morphogenetic fields. Owing to the scarcity of chemical data and to the complexity of life, and despite the progress of biochemistry the biologist is still threatened with vertigo.

—A. Lwoff 1950

That the adoption of the mechanistic view has had profound and far-reaching consequences for the whole of society is an historical fact which gives rise to the most divergent opinions. Some commend it as a symptom of the gradual clarification of human thought. . . . Others, though recognizing the outstanding importance it has had for the progress of our theoretical understanding and our practical control of nature, regard it as nothing short of disastrous in its general influence.

—E. J. Dijksterhuis 1950

In Part II we saw that the rejection of localization and decomposition tends to accompany the rejection of a mechanistic program. Providing a mechanism involves describing distinct components, each of which makes a contribution to the performance of the system. This requires both functional and physical independence. In the simplest cases, these components are thought of as making their contributions independently: nature is simply decomposable and embodies an aggregative organization. In slightly more complicated cases, the components are thought to make their contributions sequentially, or linearly, and to retain an integrity of their own: nature is nearly decomposable. As we saw in Part III, a wide variety of organizations may be revealed by beginning with an assumption of near decomposability. The resulting models may not retain the integrity of the components, but may describe what we have termed an integrated system. In such a system nature is at best only minimally decomposable. If organization becomes even more dominant in explaining the behavior of the system, and we appeal less to different and distinctive functions performed by the components, we reach a point where decomposition and localization in any recognizable form have to be surrendered.

While, historically, forgoing decomposability seemed to require giving up mechanistic approaches, this is not the only possibility. Formal modeling techniques have made it possible to explore the behavior of systems in which the components play very minor functions; the explanation of how the system behaves lies in the way these components are organized. Component behavior can be very simple, given a complex and interactive organization. Such *connectionist* systems represent an alternative way of elaborating a mechanist program without assuming decomposability or near decomposability. In Chapter 9 we briefly explore three cases in which researchers have pursued connectionist models and look at the modes of reasoning that support this alternative conception of mechanism.

We begin with a description of the contributions of John Hughlings Jackson, a late-nineteenth-century neurophysiologist who rejected the localizationist programs of Broca and others. Jackson was no vitalist, and dualism played no significant role in his scheme. His opposition to the localizationist program of neo-Phrenologists thus stood in sharp contrast to that of Flourens. Jackson denied discrete modules in the cortex controlling specific functions, just as Flourens did. Jackson did propose an alternative decomposition of the nervous system into lower and higher levels, but, unlike phrenologists or neophrenologists, these levels were not spe-

lesion data, but, given the resources of the time, it was very difficult to develop into a precise theory of how the nervous system worked.

We are now capable of modeling such systems. Until recently researchers in artificial intelligence have taken a computer that processes information in accord with rules to simulate, and perhaps to realize, thought. A new generation of researchers is exploring how connectionist systems might explain a variety of cognitive phenomena. Instead of viewing cognition as involving the processing of information according to rules, cognitive behavior is seen as the "emergent" product of a system that consists of simple units and controlled by simple learning rules. According to this approach, computers are employed not to realize rules or procedures performed on symbols, but rather to solve the equations that characterize the behavior of systems with large numbers of components. These simulations allow the computer to show how behavior might emerge from the interactions of simple components. These connectionist systems are apparently capable of overcoming some of what have seemed to be the greatest liabilities in more traditional artificial intelligence, and they do so *without* assuming that cognitive tasks can be decomposed into discrete subtasks, or that the system is organized into modules; instead, organization and integration supplant compartmentalization.

Our third example of the emergence of connectionist thinking comes from a quite different domain, work on genetic regulation. In recent years biologists have developed increasingly complex models of gene regulation, positing genes having the function of regulating other genes in a localizationist fashion. In an evolutionary framework, however, this becomes quite problematic. The more complex a system, the stronger the evolutionary pressure would have to be to maintain it in the face of mutations. Stuart Kauffman's recent work is directed toward explaining how evolution can maintain such a regulatory system; however, he has ended up *redescribing* the phenomenon to be explained, in something like the manner of the case discussed in Chapter 8. Kauffman develops an abstract, formal model of a genetic system to simulate the regulatory process, a model remarkably like the connectionist systems developed to simulate cognition. These model genetic regulatory networks function by allowing units to excite and inhibit each other until a stable pattern of activation is achieved. Kauffman equates these stable patterns with cell types. Random perturbations will disturb the system's behavior, but it turns out that while it is nearly impossible for selection forces to maintain a large and complex system against mutation in an arbitrarily defined optimal state, some systems will evolve toward a stable state where mutations

over, it takes tremendous selection pressure to move the system out of these stable states. This suggests an answer to the original problem, but one that requires changing the question. Rather than asking how selection could maintain a complex regulatory system, Kauffman claims that the regulatory system is inherently stable and does not require selection to sustain it. The result of developing a connectionist system to model gene regulations suggests that what was a pressing issue when the system was assumed to be nearly decomposable does not even require explanation.

In Chapter 10 we place the description of theory development we have built using the cases discussed in Chapters 3–8 within a broader context. We have focused on a number of choice-points confronted in the development and elaboration of mechanistic programs, points which describe the kinematics of theory development. The informal flow chart falls short of a fully dynamic model, though it is a useful step in that direction. It is also useful in suggesting the general directions that a realistic account of discovery might take. Following the lead in Chapter 2, our primary focus has been on psychological constraints and, in particular, on the heuristics of decomposition and localization. In examining the cases in Parts II and III, we have explored how these heuristics came into play in shaping research programs. It is clear from the cases discussed in Part III that the heuristics of decomposition and localization underdetermine the result. There are several directions development takes, depending on other contingencies; the heuristics do not operate in isolation.

In a more speculative vein we will identify three general factors that feature, in conjunction with heuristics, in explaining the dynamics of theory development. These include *phenomenological regularities*, *operational constraints*, and *physical constraints*. We have no detailed account of the character of these constraints, nor of their interaction. A fully elaborated description of how these factors, and others still not identified in any detail, figure in the course of theory development lies beyond the scope of this book. Until these factors are detailed we doubt it will be possible to develop a complete normative account of the history of mechanistic research programs, much less a comprehensive normative account of theory development. Having examined some of the factors affecting theory development, though, we can begin to glimpse what a more adequate account of discovery would look like.

"Emergent" Phenomena in Interconnected Networks

1. Introduction: Dispensing with Modules

The more complex localizationist explanations we examined in Part III are still recognizably mechanistic. Tasks involved in performing a function are divided between components, and system behavior is explained by showing how it is accomplished through the combined performance of the component tasks. Although one might prefer explanations in which the component tasks can be thought of as following a linear, sequential order, so that the contributions of each component can be examined separately, natural systems are not always organized in such a manner. Component tasks are often dependent on one another, so we cannot understand the operation of the system by imposing a linear order on it. Cyclic rather than linear organization occurs when the activity of any given component is dependent on a variety of other components that, in turn, depend on it. In integrated systems, the explanation of the behavior of the whole system depends in a *nonlinear* way on the activities of the components and on the modes of interaction found within the system.

To the extent that organization is important in affecting system behavior, a system is nondecomposable. As we discussed in Chapter 2, there is a continuum of cases. At one end are simply decomposable systems for which the major explanatory task is to identify the components and understand their behavior. These include the kinds of cases considered in Chapter 4, in which a single part is held to be responsible for the behavior of the whole system. Toward the other end of the continuum lie cases of integrated systems in which the behavior of the system is largely due to the interaction of the components. In all of these cases, the operations of the components can, nonetheless, be understood in terms of the operations performed by the whole. Conversely, the behavior of the whole is explained in terms of the behaviors of component parts. There are other systems, yet farther out on the continuum, in which localization and decomposition appear to be hopeless, or even misguided. The hallmark of these cases is that, given a principled structural analysis, the activities of the parts seem to be different in kind from, and so far simpler than, those performed by the whole. The parts can be so simple, in fact, that they do

With systems in which the parts do not seem to be performing intelligible subtasks contributing to the overall task, classical mechanistic strategies—and, in particular, decomposition and localization—fall short. What alternatives are there to pursuing a program of mechanistic explanation? In Chapter 5 we considered one possibility: rejecting the mechanistic program and settling for descriptive accounts of behavior. But this is not a strategy for developing an explanation; it is a denial of any explanation. In this chapter we consider an explanatory strategy that abandons localization and decomposition. We leave open whether it constitutes a properly mechanistic approach.

We examine three cases in the following sections. Hughlings Jackson provides a transitional case. Jackson rejects the localization of Broca and proposes a complex, hierarchical model of the nervous system in which control of tasks was distributed over different neural structures at different levels, with higher-level systems regulating and modulating performance of lower-level systems. Our other two cases come from contemporary research, which use newly developed formal mathematical tools for modeling the behavior of complex systems. These two cases are, respectively, models of cognitive performance and genetic regulation. In these cases performance depends primarily upon the interaction of the components in the system. The components do not perform tasks that would appear in a functional decomposition of the system.

2. Hierarchical Control: Hughlings Jackson's Analysis of the Nervous System

In the late nineteenth century, John Hughlings Jackson developed a hierarchical model of the nervous system that was intended as a repudiation of the kind of localizationist claims advanced by Bouillaud and Broca.[2] These latter "neophrenologists," as we have seen, constructed localized models on the basis of correlations between neurological lesions and pathological symptoms such as the loss of coherent speech. Losses of specific capacities were traced to injury or destruction of specific regions. While such correlations certainly can be found, they are not as precise as would be required to justify the strong conclusions of these localizationists. The clinical syndromes are not simple, and the neurophysiological disorders rarely correlate precisely with the syndromes. For example, aphasia—in at least one of its myriad forms—is commonly described as a deficit affecting the verbal expression of language, but not affecting cognitive abilities. However, these "expressive" aphasias do not affect the whole range of verbal expres-

cal deficits are not neatly circumscribed. Considerable inference and conjecture are required to draw conclusions concerning the locus of the damage. Expressive aphasias are accompanied, as one would expect, by damage to the third frontal convolution; but damage from either external trauma or stroke is not clearly limited or localized.

One response to this lack of precise correlation between cerebral damage and pathological syndromes is to reject the search for a mechanistic explanation. As we saw in Chapter 5, when Flourens failed to substantiate Gall's correlations between craniological structures and psychological capacities, he denied the localizationist program and gave up the search for mechanisms governing the higher cognitive capacities. This was not an option Jackson could accept. He was committed to understanding the neurophysiological operations of the brain, convinced these would explain the associated psychological deficits. This required making sense of how the brain accomplished its operations through the interaction of its parts. As a good clinician, the complexity of the symptoms was always before him. He was led to a different explanatory approach.

Jackson found the key to developing an alternative interpretation of the operation of the nervous system in a different pattern of deficits, this time in epileptic seizures (see Jackson 1884; Melville 1982). Epileptic seizures, or the more extreme cases at least, are generalized, affecting the entire body to varying degrees. Jackson saw three levels of seizures (cf. Jackson 1884, pp. 57ff.). The first and least severe is analogous to a dream state. The second is accompanied by a loss of consciousness. The third leaves the patient comatose. These three levels increase in severity, with the third most encompassing and the first least so. As a consequence of his complex symptomology, Jackson maintained, the epileptic discharge must begin in a region that affects the body as a whole—including the highest levels, such as volition. It must begin in cortical structures. A mild seizure is the analogue to dreaming. A more severe discharge affects the lower levels and also disrupts consciousness. In yet more severe cases, the functioning of more central structures are also disrupted. In an analogous manner, Jackson thought expressive aphasias deprive the patient of the ability to use complex forms of language, but leave the more automatic uses relatively intact. Jackson accordingly maintained that there are "higher" and "lower" uses of language: the intellectual and emotional uses—or, as he later referred to them, the *superior* and *inferior* forms of speech. The former were genuinely expressive of thought; the latter were not. The intellectual, or superior, functions are the most labile, but even when severely impaired, some use of language and some residue of emotional expression are preserved. What these patterns of pathology suggested to

function could be maintained while others were destroyed. And some of the manifestations, he thought, required more development than others; for example, representational speech demanded more sophisticated development than did emotional speech. As he analyzed the syndromes, typically the least-developed mode of the function was maintained, while its most-developed or sophisticated use was lost.

Jackson's appeal to multiple control was influenced by two further sources: associationist psychology and evolutionary theory (see Smith 1982a). As we shall explain, the first source reinforced his repudiation of a localizationist program and his search for a different mode of explanation; the second provided a theoretical basis for his hierarchical model. Jackson derived both of these themes from Herbert Spencer's *Principles of Psychology* (1855, 1872).[3]

The associationist program in psychology assumed a variety of forms throughout its long history.[4] The crucial element in associationism was the idea that complex knowledge was built up from smaller units—ideas, in its classical formulations—through a limited number of general principles of association. Thus, for Locke, "simple ideas" become the atoms of knowledge, which can be held before the mind, compared, and compounded. In Hume's hands the principles of association were more clearly and rigorously circumscribed. He says in his *Treatise* (1888) that the "qualities, from which this association [of simple ideas] arises, and by which the mind is after this manner convey'd from one idea to another, are three, *viz.* RESEMBLANCE, CONTIGUITY in time or place, and CAUSE and EFFECT" (p. 11).

When applied in the neurological realm by Jackson, the associationist program mandated the abandonment of localization, rejecting the faculty psychology upon which the localizationist program of Gall, Bouillaud, and Broca had been built. According to the associationist program as Jackson inherited it, just as complex ideas are produced from simple ideas by the processes of association, so also complex mental operations must be compiled operations composed of simple sensory-motor associations. Any difference between higher and lower cognitive operations will involve simply more associations between lower-level sensory-motor processing, and not a unique cognitive faculty. The cognitive faculties posited by classical localizationist approaches could not be among the basic capacities, and localization would be fruitless (Richardson 1986a). "Will, memory, reason, and emotion," Jackson tells us, "are simply artificially distinguished aspects of one thing, a state of consciousness" (1884, p. 66).

Having abandoned a faculty psychology, Jackson was also committed to abandoning the localization of specific faculties. Whereas Broca had taken the "faculty of articulate language" to be one of many discrete and primi-

so-called 'faculty' of language has no existence" (Jackson 1866, p. 123). Associationism left no room for organology, and without organology there was no room for classical cerebral localization.

The evolutionary commitment provided the positive dimension to Jackson's account of neuropathologies. As he said at the outset of his seminal lectures, "We shall be very much helped in our investigations of diseases of the nervous system by considering them as reversals of evolution, that is, as dissolutions" (Jackson 1884, p. 45; cf. Jackson 1882). Evolution here was not the simple "descent with modification" defended by Darwin; rather, the evolutionary views of Herbert Spencer are what inspired Jackson. Unlike Darwin, Spencer embraced an orthogenetic, progressive view of the evolutionary process. Evolution was the expression of an inherent tendency toward development from lower to higher forms, or from less to more complex forms of organization. Evolution, according to this vision, was "a passage from the most simple to the most complex" and "from the automatic to the voluntary" (Jackson 1884, p. 46). The culmination and natural result of evolution was consciousness. Yet these later, higher forms were not to be constructed de novo; they were, rather, further modifications of the basic plan provided by the lower, less complex form.[5] This would typically involve the imposition of a control mechanism on the early products of evolution, allowing utilization of these devices for more complex functions.

When he turned to the pathological syndromes, Jackson proposed that the nervous system was organized in a hierarchical manner, with different levels representing different stages of evolutionary development.[6] The higher levels are more recent and are connected with volition and consciousness. The lower levels are more primitive and are connected with habit and reflex. Each higher level, Jackson thought, must work through the lower levels. Thus, in animals arising later in phylogeny, newly evolved neural structures would arise that would modify and regulate the performance of brain components that had emerged earlier. At the base of this hierarchy are parts of the nervous system that directly respond to sensory stimuli or control motor output. These basic mechanisms are specialized and task-specific, each representing some specific movement of a specific part of the body. Here is the only point in the hierarchy where we can find neural structures able to work independently to perform a function, and, hence, the only point where we might try to decompose the operation of the whole in terms of component operations. The kinds of operations performed by these components, though, are not the sort that would suffice for localizing *cognitive* capacities. These mechanisms, ac-

nents for planning or reasoning about actions. At the middle level we have motor centers that are less specialized and less task-specific. Each center represents some complex movement, compounded of simpler movements represented directly by the lower level. These middle-level centers effect associations between lower-level components. These centers are less organized in that they are less automatic and more receptive to modification by experience. Part of what is critical about these systems, however, is that they do not work independently, not representing within themselves the information they require to perform their tasks. They achieve their effects by regulating, modifying, and integrating the operation of the lower-level components. Motor centers at the highest level each represent to some degree the entire body or movements of it. Their function is to coordinate complex movements. But, once again, these are not independent structures; they are mechanisms for regulating lower-level structures. Jackson summarizes this tripartite view and offers an anatomical interpretation:

> The lowest motor centres are the anterior horns of the spinal cord, and also the homologous nuclei for the motor cranial nerves higher up. . . . The lowest centres are the most simple and most organised centres; each represents some limited region of the body indirectly, but yet most nearly directly; they are representative. The middle motor centres are the convolutions making up Ferrier's motor region [just anterior to the central sulcus]. These are more complex and less organised, and represent wider regions of the body doubly indirectly; they are re-representative. The highest motor centres are convolutions in front of the so-called motor region. . . . [They] are the most complex and least organised centres, and represent widest regions (movements of all parts of the body) triply indirectly; they are re-re-representative. (1884, p. 53)

What higher units do, from Jackson's perspective, is coordinate what is directly represented in distinct lower-level components and regulate the activities of these components. This is what Jackson means by re-representing or re-re-representing what is already represented at the lower level.[7]

Schematically (Figure 9.1) we may conceive of (S_{11}, S_{12}, . . . , S_{18}) as specialized organs at the lower level. Each controls a single movement, and loss of the organ results in paralysis of the corresponding part of the body. Loss of S_{11} would mean the loss of a specific behavior—perhaps the ability to move an arm, or to flex a finger. Similarly, (S_{21}, S_{22}, . . . , S_{24}) occupy the second level and re-represent the movements represented at the lower level. These involve an intermediate level of integration and coordination. Each exerts some control over a variety of movements,

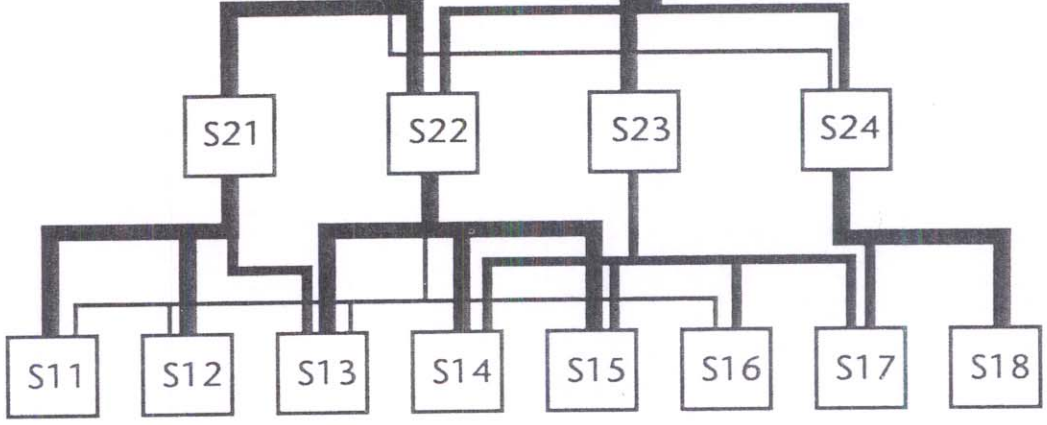


Figure 9.1. A Schematic Representation of a Control Hierarchy of the Sort Proposed by Hughlings Jackson (1884). Higher-level units exert a broader control, while lower-level units are more specific. Units at level 1 (S_{11}, . . . , S_{18}) control specific behaviors. Units at levels 2 (S_{21}, . . . , S_{24}) and 3 (S_{31}, S_{32}) coordinate the more specific behaviors of level 1. Breadth of the lines is meant to correspond to the strength of the connections; for example, S_{31} is strongly tied to S_{11} through S_{15}, but only weakly connected to S_{17} or S_{18}.

though the amount of causal influence they exert over units at the lower level is variable: S_{21} coordinates activity of three more specialized units (S_{11}, S_{12}, and S_{13}). Loss of S_{21} would entail a loss in the coordinated activity, but would not result in any paralysis, because each of these units can be activated by other higher-level units. Units at the highest level (S_{31} and S_{32}) produce the highest level of integration, but in so doing they exercise the least specific influence. They coordinate and integrate the activities controlled at lower levels; in Jackson's terms, they re-re-represent the movements represented at the lowest level and re-represented at the intermediate levels, and their degree of causal influence is also variable.

This hierarchical structure reflects our evolutionary heritage. By preserving the lower-level structure and function that was present in their evolutionary ancestors, higher, more developed organisms reflect the evolutionary history of the species; they recapitulate their evolutionary history. This preservation is most clearly revealed when the higher levels of the nervous system are destroyed or damaged. Jackson insists that symptoms will be both positive and negative. On the negative side there will be a loss of capacities: Some aphasics lose the ability to comprehend spoken speech; others lose the ability to speak. Epileptics may lose consciousness. This would mean that epileptics or aphasics would suffer a loss of the coordination and fine modulation of lower-level functions, but would not

under higher-level control. Jackson insists, "Disease only produces negative mental symptoms answering to the dissolution. . . . [All] elaborate positive mental symptoms . . . are the outcome of activity of nervous elements untouched by any pathological process" (1884, p. 46). What would be left following the destruction of higher centers would be the more automatic, reflexive forms of behavior typical of lower organisms. Jackson refers to this process as the *dissolution* of the nervous system and claims to see it exhibited in the pathological syndromes such as aphasias. As evidence for this interpretation of the neuropathological syndromes, he points to the nature of the symptoms, which are typically negative. As we noted earlier, Broca's patient, Tan, was still capable of some vocalizations, but these were limited, simple, and automatic utterances such as oaths or simply "Tan." This is typical: voluntary applications of speech suffer more dramatically than do more automatic ones, and the latter suffer more than the applications used in emotional expression. With damage to the cerebral lobes, Jackson reasoned, the higher, voluntary uses of language would be lost and the lower functions would accordingly come to predominate. The patient would lose the ability to "convey propositions" through symbols, but would still have the ability to relate feelings.[8] This, from Jackson's perspective, reflects the maintenance of lower-level motor control despite the loss of higher levels of control over speech.

In one respect Jackson has simply repartitioned the nervous system into functional units, offering a different decomposition. His division into units crosscuts the division of the organologists. But the differences are more far-reaching than that would suggest. The localizationists divide the brain *horizontally* into a number of processing components, each operating at roughly the same level and in relative independence of one another. Each component has its own specific function. Jackson partitions the system *vertically*, with higher-level components operating on and modulating the behavior of the lower-level components. Only at the lowest level do we have specialized and independent modules. Moreover, the higher levels cut across the lower levels in their mode of operation, so we do not have a neat division of the cerebral lobes into regions responsible for regulating single, lower-level processes.

As we explained in Chapter 6, localization requires there to be a physical analysis corresponding to a projected cognitive organization. Even on Jackson's view, basic motor control is localized in this sense: there are physically discrete regions of the brain exercising limited control over specific behaviors. But Jackson's approach does not allow localization of cognitive capacities, and the more complex these capacities are the less localization will make sense. It is in this sense that his approach is non-

3. Parallel Distributed Processing and Cognition

Cognitive science is one area of research in which the decomposition of complex tasks into component tasks has been widely applied in recent decades. For the most part this has had a top-down orientation, with little attention to the details affecting realization. At its most extreme it has been antagonistic to physical details. The dominant metaphor is that cognition is information processing. The primary approach has been to analyze complex information processing tasks into simpler information-processing tasks and to seek simple mechanical realizations for the most basic tasks. As William Lycan, somewhat colorfully, portrays the program of research,

> We explain the successful activity of one homunculus . . . by positing *a team* consisting of several smaller, individually less talented and more specialized homunculi—and detailing the ways in which the team members cooperate in order to produce their joint or corporate output. (1987, p. 40)

In Lycan's hands it is homunculi as far down as psychology can see. Even the simplest component tasks in such cognitive theories remain ones of transforming information. In some accounts the information is understood as being represented in the system in symbolic structures. Processing is the transformation of symbolic structures into other symbolic structures; the symbols, in turn, are construed as semantically interpretable—that is, as referring to objects and having associated meanings (Fodor 1975). The resulting explanations describe the overall task in terms of operations that are intelligible given the semantic interpretation. For example, in a computer program that plays chess there may be operations that propose possible moves, and other operations that project the resulting board positions and evaluate them. The representations are of pieces and positions in a space defined by the board and the rules of chess. The operations are the legal moves. Since the symbols may refer to actual pieces and the operations to actual moves, we can readily understand how the program goes about playing the game.

This approach to cognitive behavior is commonly known as the *symbolic* approach, or, sometimes, as one relying on *rules and representations*. It is not the only possible strategy. When cognitive science was in its early infancy, another approach briefly emerged and then lay fallow. This competitor focused on networks composed of simple entities, supposedly similar to neural units, which exchanged activations and inhibitions (Ro-

functions without operations on formally represented symbols. Some network models, such as Selfridge's Pandemonium, employed a homuncular representation of the overall task, decomposing it into significant subtasks. Other systems, such as Rosenblat's Perceptron, did not. A Perceptron takes in one pattern of activations and outputs another, with each input unit linked to an output by weighted connections. The network approach nearly disappeared with the publication of Minsky and Papert's *Perceptrons* (1969), which attempted to establish inherent limitations in network models. Recently, however, the network approach has reemerged and goes by such names as *connectionism*, *parallel distributed processing*, or *neural networks*. Whereas the symbolic approach epitomizes a variation on decomposition and localization, network models present a significant alternative program, proposing to explain cognitive functions without employing decomposition and localization.

It should help to develop the contrast more fully. The return to a cognitive psychology—that is, one acknowledging internal cognitive processes—after decades of dominance by behaviorism was inspired by two developments: the introduction of the digital computer, and Chomsky's (1957) defense of a generative grammar. The focus on formal systems in which symbolic structures are manipulated according to sets of rules is critical in both of these domains. Computers are often construed as list-processing machines, wherein lists of symbols coded in digital form are manipulated either by the hardware's configuration or by instructions stored in other lists of symbols that constitute the program. Chomsky's linguistics proposed to represent the set of grammatical sentences available in a natural language in terms of a finite structure. This would come in a set of rules that operate recursively on sequences of symbols, transforming these sentential representations while preserving grammaticality. Chomsky subsequently attempted to extend his analysis as a general account of human language. The basic idea has now been generalized into a wide-ranging set of models of human cognition in which the information to be processed is represented symbolically, and rules are introduced to manipulate these symbols. From our vantage point, what is important is that these rules and representations embody an attempt to account for an overall performance of the cognitive system by decomposing that task into simpler tasks.

One of the widely employed types of design for achieving cognitive performance within symbolic systems is the production system (Anderson 1983; Newell 1973). In production systems information is encoded symbolically in working memory (a variety of information can be encoded at any given time). The rules are conditionals: the *antecedent* specifies a test

gram assumes that mental activity can be decomposed into a set of operations, each of which is governed by a set of rules operating on symbolic information. It has produced cognitive models successful in accounting for many features of human cognition; moreover, for many years it seemed the only option in developing an account of cognition.

A growing number of cognitive scientists have begun to explore the possibility that what *appears* to be rule-following behavior might only be *described* by these rules and that it might be feasible to explain cognitive performance without mirroring the rules in the mechanisms that explain that performance. The reemergence of network models, and the mounting evidence that more complex network architecture can perform tasks that previously seemed beyond the ability of the networks,[9] have led to a reexamination of the view that cognitive processes must be decomposed according to a rule. In this line of research, then, the decomposition of the symbolic approach is lost.

To illustrate why a network architecture is a significant alternative to the symbolic architecture, and how it promises to explain cognitive performance without engaging in the sort of decomposition that has been characteristic of the symbolic approach, it will be useful to describe briefly some common features of connectionist architecture and present two simple examples of connectionist systems. The basic components of a connectionist system are simple processing units able to assume varying states of activation and connections of varying strengths through which they can excite or inhibit others. There is tremendous variability in the network designs currently being considered, but what is common to all is that the units are simple and their interactions are critical. Activations can either be limited to discrete values (for example, 0 and 1), or can vary within a continuous range (say, from 0 to 1 or -1 to $+1$). Units can be connected with each other in various ways. They can be layered so that units within one layer project only to units in adjacent layers. Or, the network can be completely interconnected. Typically the input to a unit is determined by summing over the products of the activations of other units connected to it, and the weights associated with the connections between those units and the target unit. In Figure 9.2, the input to unit A is determined by multiplying the activation of B, C, and D by the weights connecting them to unit A, and then adding those values; this yields a net input to unit A of .24. The activation of A may then be determined by a variety of formulas. Some take the prior activation of A into account; some do not. The functions, moreover, may involve either linear or nonlinear functions in determining the activation level of A.

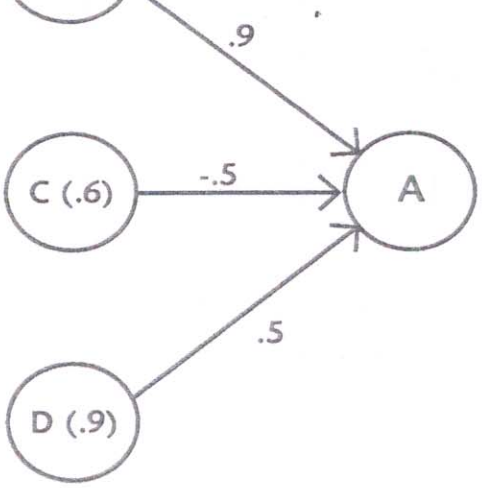


Figure 9.2. A Simple Connectionist System. In this case the input to A is an aggregative function of the activation levels of B, C, and D and their connection strengths to A.

One important source of variation arises in determining how the weights associated with the various connections are decided. They can be preset, but much of the interest in these networks stems from their ability to alter their own performance by changing the weights. This is done by using a variety of algorithms. Operations for changing these weights model learning in the networks; accordingly, the algorithms used are commonly referred to as *learning rules*. The simplest learning rules are variations on a suggestion by D. O. Hebb (1949). These are basic associationistic principles, requiring an increase in connection strength between units when they are simultaneously excited. For example, one variation might require increasing the strength of the connection between units in proportion to the product of their activation levels.[10]

In addition to determining the actual mechanics of a network, a researcher must also specify how these networks are to be construed as performing a cognitive function. This is usually spoken of as providing an *interpretation* for the activities of the network. There are two general approaches to interpreting the activity of such systems. The simplest is to let each unit represent a hypothesis or a goal. Since each unit has its own representational function, this approach is referred to as *localist*. The more complex, but potentially more interesting, interpretation is to treat a particular pattern of activation over an ensemble of units as serving the representational role. In such systems one pattern of activation over a set of units may receive one interpretation, while another pattern over the *same* set of units receives another. Since it is the pattern of activations that determines the interpretation, this approach is referred to as *distributed*.

The crucial move common to all network models is that of explaining cognitive performance without casting the explanation in terms of rules

overall cognitive task. Network models do account for the cognitive performance, but often they do so without providing an explanation of component operations that is intelligible in terms of the overall task being performed. The network is a cognitive system; the components are not. The result is that we do not explain how the overall system achieves it performance by decomposing the overall task into subtasks, or by localizing cognitive subtasks.

In order to show how network models provide an explanation that does not involve decomposing the overall task, we will briefly describe two simulations. The first involves a two-layer network learning to recognize patterns.[11] The overall process is one that could figure either in basic perception or in categorizing objects already perceived. The network consists of eight input and eight output units, with each input unit connected to each output unit (see Figure 9.3). The activation (a_j) of an output unit j is the sum over the products of the input activations for each of the i units, and the weights of the connections linking them to the output units (w_{ji}): $a_j = \Sigma a_i w_{ji}$. The input arrays can be viewed as representing objects belonging to four different categories. We will suppose arbitrarily that these are cup, bucket, hat, and shoe. Table 9.1 shows the input patterns that correspond to a prototypical instance of each category, and the target output patterns that the network is trained to approximate. The target outputs can be thought of as the system's names for the four categories. For example, the input array for a prototypical bucket is $\langle -1, -1, +1, +1, +1, -1, -1, -1\rangle$, and the output array for "bucket" would be $\langle -1, -1, -1, -1, +1, +1, +1, +1\rangle$.

The goal of the network is to learn the association between the input array and the "name." In training the network, the actual inputs and target outputs were distorted by a randomly chosen amount between -0.5 and 0.5 to capture the fact that we do not always encounter prototypical objects or undistorted names. Thus, where the pattern designated for the prototypical cup is $\langle -1, -1, -1, -1, +1, +1, -1, -1\rangle$, an actual input on one trial might be $\langle -0.76, -0.89, -1.21, -1.01, +1.33, +0.99, -0.65, -0.92\rangle$. The network was trained over fifty epochs, or training sets. During each epoch the network received a distorted version of each input and its corresponding distorted target output. The network used the *delta rule*, which adjusts the weights (w_{ji}) of each connection leading to each output unit on each trial by an amount proportional to the product of the difference between the actual output (a_j) and the target output (t_j) and the activation of the relevant input unit (a_i) on that trial. This can be represented as $\Delta w_{ji} = .0125\,(t_j - a_j)\,a_i$. If the target output for U_5 is -1.0, and

Object:	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
Cup	−1	−1	−1	−1	+1	+1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
Bucket	−1	−1	+1	+1	+1	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1	+1
Hat	−1	+1	+1	+1	−1	+1	−1	+1	−1	+1	−1	+1	−1	+1	−1	+1
Shoe	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

Table 9.1. Prototypical Inputs and Outputs for Two-Layer Pattern-Recognition Network with Eight Units per Layer.

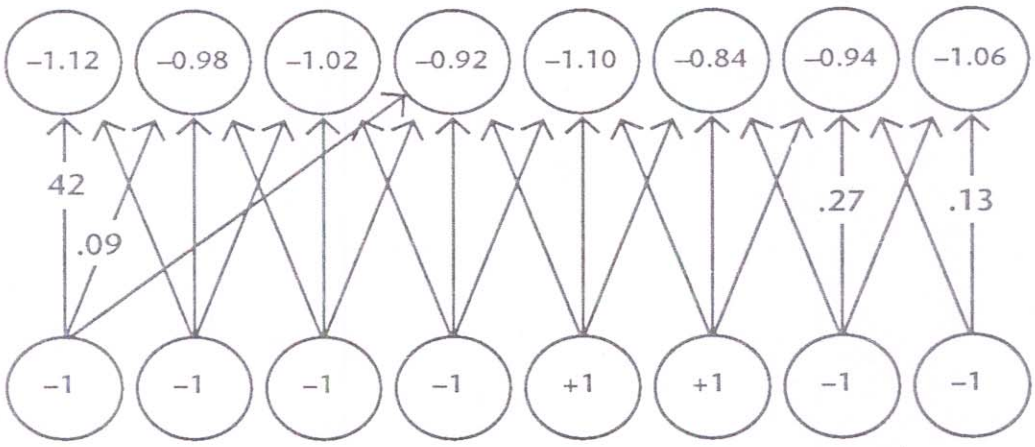


Figure 9.3. A Two-Layer Pattern-Recognition Network. All units at the lower level are connected to each unit at the higher level, with no intervening levels. For simplicity, not all connections or connection weights are shown.

its actual output is -0.8, then the above input array would adjust the weights of the connections to *U5* from the input units up or down by the following amounts: $\langle +.00190,\ +.00223,\ +.00303,\ +.00253,\ -.00333,\ -.00248,\ +.00163,\ +.00230\rangle$. In this instance the rule increases the strength of the connection weights from input units having the same sign as the target output, and decreases the rest.

After training through the full fifty epochs, the network was tested on three different types of input: the actual prototype of each category, an instance randomly distorted in the way described above, and an input for which the sign of one of the input units of the prototype was reversed. The test inputs are detailed in Table 9.2. When presented with an actual prototype, or with a distorted version of the input, the outputs were all within .5 of the target output. Even when given a pattern in which one of the eight input values was reversed in sign from the prototype, the system produced outputs that were positive or negative as appropriate in all but one case.

	Test inputs								Outputs						
Object	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	U_1	U_2	U_3	U_4	U_5	U_6	
Cup	–1.00	–1.00	–1.00	–1.00	+1.00	+1.00	–1.00	–1.00	1.12	–0.98	–1.02	–0.92	–1.10	–0.84	–0
	–0.76	–0.51	–0.82	–1.11	1.47	0.82	–0.83	–0.90	–0.81	–0.90	–0.71	–0.83	–0.77	–0.72	–0
	–1.00	–1.00	–1.00	–1.00	+1.00	*–1.00*	–1.00	–1.00	–0.86	–1.39	–0.85	–1.41	–0.26	–0.78	–0
Bucket	–1.00	–1.00	+1.00	+1.00	+1.00	–1.00	–1.00	–1.00	–0.99	–1.06	–0.98	–0.96	0.91	0.94	0
	–1.00	–0.54	1.34	0.63	0.98	–0.59	–1.24	–0.81	–1.06	–0.81	–1.03	–0.68	0.63	1.00	0
	–1.00	–1.00	*–1.00*	+1.00	+1.00	–1.00	–1.00	–1.00	–0.98	–1.24	–0.96	–1.22	0.30	0.06	0
Hat	–1.00	+1.00	+1.00	+1.00	–1.00	+1.00	–1.00	+1.00	–0.91	0.96	–0.87	1.05	–0.84	1.06	–0
	–1.18	0.62	1.20	0.87	–1.21	1.38	–1.02	1.48	–1.07	1.11	–1.01	1.22	–1.12	1.10	–1
	–1.00	*–1.00*	+1.00	+1.00	–1.00	+1.00	–1.00	+1.00	–1.20	0.38	–1.14	0.49	–0.74	0.87	–0
Shoe	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	0.99	0.94	1.05	1.07	0.93	1.03	0
	1.42	1.44	0.64	1.31	0.72	1.24	1.03	1.19	1.20	1.28	1.25	1.39	0.81	1.00	0
	–1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	0.13	0.75	0.21	0.85	0.38	1.18	0

Table 9.2. Test inputs and corresponding outputs following 50 training epochs, based on target outputs in Table 9.1. In the tes system was presented with three test inputs for each of the four items. The first row for each item gives the prototypical input an output. The second row shows randomly permuted prototypes and the responses to them. And the third row gives the prototype, sign of one input unit reversed (indicated by italics). With one exception (indicated in boldface), the outputs are at least of the ri

is something that connectionist networks do well, it is also something quite difficult for symbolic systems. Consider how one would design a symbolic system to perform this task. The input pattern for an object is simply an arbitrary array of activations produced in units of the system. In more realistic situations, this array would be produced by a set of feature detectors in the visual system, and the values in the array would represent values on these features. For example, a positive value on an input unit might indicate the presence of a feature or the activation (confidence) level of the corresponding feature detector. The values of the output units would likewise represent features of either the word or the mental representation of the object. A symbolic system would begin with encodings of the features of the input and seek to develop rules that would produce the symbolic representation of the output. The rules would have to specify what set of features would constitute, for example, a shoe. Since there is significant variability between shoes, the rules would have to specify the various combinations of features in the input that should still result in recognizing the object as such. The attempt to produce such rules has overwhelmed AI investigators.[12]

A connectionist approach, by contrast, does not set out to identify rules. Rather, the connections in the network are allowed to adjust during the training phase until the network can efficiently distinguish the objects in the domain. For the network to accomplish this, it must have structure. This is found in the connections, which serve the function of rules in a symbolic system. We sometimes can interpret the connections as providing rules for how to identify, say, balls as opposed to shoes on the basis of features. It is important to recognize how these rules are obtained. They are *not* developed by specifying the conditions under which an object of a given sort is present. Instead, the rules result from the network's discovering the correlation between features and objects. Each weight represents the reliability of a specific feature as an indicator within the class of objects. Thus, in the array given in Table 9.1, a positive value for I_7 is a good indicator for a shoe, while a positive value for I_2 tells us less (since it is ambiguous between hats and shoes). The learning rule allows the system to adjust dynamically to find a set of weights that achieves the best fit, given the class of inputs. What is important for understanding the contrast is that, except in the limiting case in which the weight to an output unit is 0, each input feature contributes something to the net output, and generally none is individually sufficient to determine the output. Thus, no decomposition into meaningful subtasks is needed.

Two-layer networks of the sort we have just described are able to learn to compute many relations between inputs and outputs, but there are

AND and inclusive OR, it cannot compute a function such as exclusive or (XOR). A two-layer network is unable to compute a function for which the target outputs cannot be linearly separated.

Two steps are required to surmount this sort of difficulty. The first is to insert one or more layers of units between the input and output layers. The second is to use a nonlinear activation function such as the logistic function $a_i = 1 / (1 + e^{-net(i)})$, where *net(i)* is the sum of the products of input activations and connection weights. The second network we will describe uses multiple layers and the logistic activation function. Our interest in multilayer networks, however, is not due simply to their greater computational power. The intermediate layers of units—generally referred to as *hidden units*—can be viewed as performing component processing steps: they process information received from earlier layers and then provide this new information to units in later layers for further processing. But if the network can learn, then the determination of the computations performed by the hidden units, as well as what the hidden units represent, are dynamically determined by the system. Moreover, the information represented by these units does not precisely correspond to the type of information that would normally be employed in a symbolic decomposition of the overall task.

This is clearly seen in a network designed by Hinton (1986), which learns information about the two isomorphic family trees, one English and one Italian, shown in Figure 9.4. The information in these trees can be encoded in 104 simple relational propositions of the form $\langle person_1$ R $person_2\rangle$ (for example, Colin has mother Victoria). Hinton constructed a network containing 36 input units and 24 output units. Twenty-four of the input units stand for the individuals in the two families and are the values that can be assigned for "$person_1$." The other 12 input units give the possible values for R. The twenty-four output units also represent the individuals, though this time as the possible values for "$person_2$." In Figure 9.5, modified from Hinton (ibid.), we show the two sets of units representing individuals (*P1* through *P24*) and the twelve indicating geneological relations (*R1* through *R12*). Because a single unit stands for a person or a relation, this is an instance of a local representation. In this system, in addition, there are three layers of hidden units. The first layer consists of 12 units, 6 of which (*H1* through *H6*) receive inputs from the 24 units coding for $person_1$, while the other 6 (*H7* through *H12*) receive inputs from the 12 units specifying the relationship. The second layer also consists of 12 units, which receive inputs from all 12 units in the first hidden

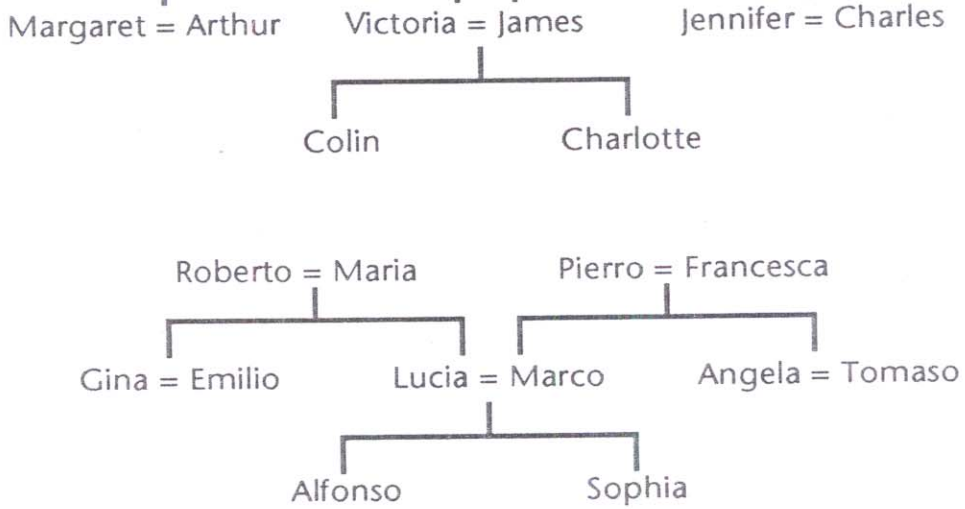


Figure 9.4. Two Isomorphic Family Trees. Hinton (1986) examines a connectionist network designed to deal with information concerning family relationships learned from the two isomorphic family trees depicted here.

layer. The final layer of hidden units consists of 6 units. The fact that the 24 units coding for person$_1$ must feed their information through a bottleneck of 6 hidden units forces the network to find a distributed representation of the different individuals that captures whatever information about the individuals the network requires to complete its task. The same principle operates with respect to the 6 hidden units receiving input from the 12 units specifying the relationship.

The network was trained to identify the correct person$_2$ when given person$_1$, and the relationship for 100 of the 104 relational propositions using the back-propagation algorithm as the learning rule.[13] After 1500 cycles of training,[14] Hinton's network not only learned to complete all 100 training propositions, but was also able to generalize to the four remaining propositions.[15] How the network accomplished this is significant. We can determine the representational function assumed by the first set of hidden units (6 are connected to the person$_1$ units, and 6 to the relationship units) by examining the weights between the inputs and this layer. In Figure 9.6 the 6 hidden person$_1$ units and the weights coming into them from the 24 person units are displayed. White boxes indicate positive weights on a connection; black boxes indicate negative weights. The size of the boxes represents the strengths of these connections. It is clear that the hidden units have extracted useful information about the individuals, despite the fact this information was never explicitly provided to the network. Thus,

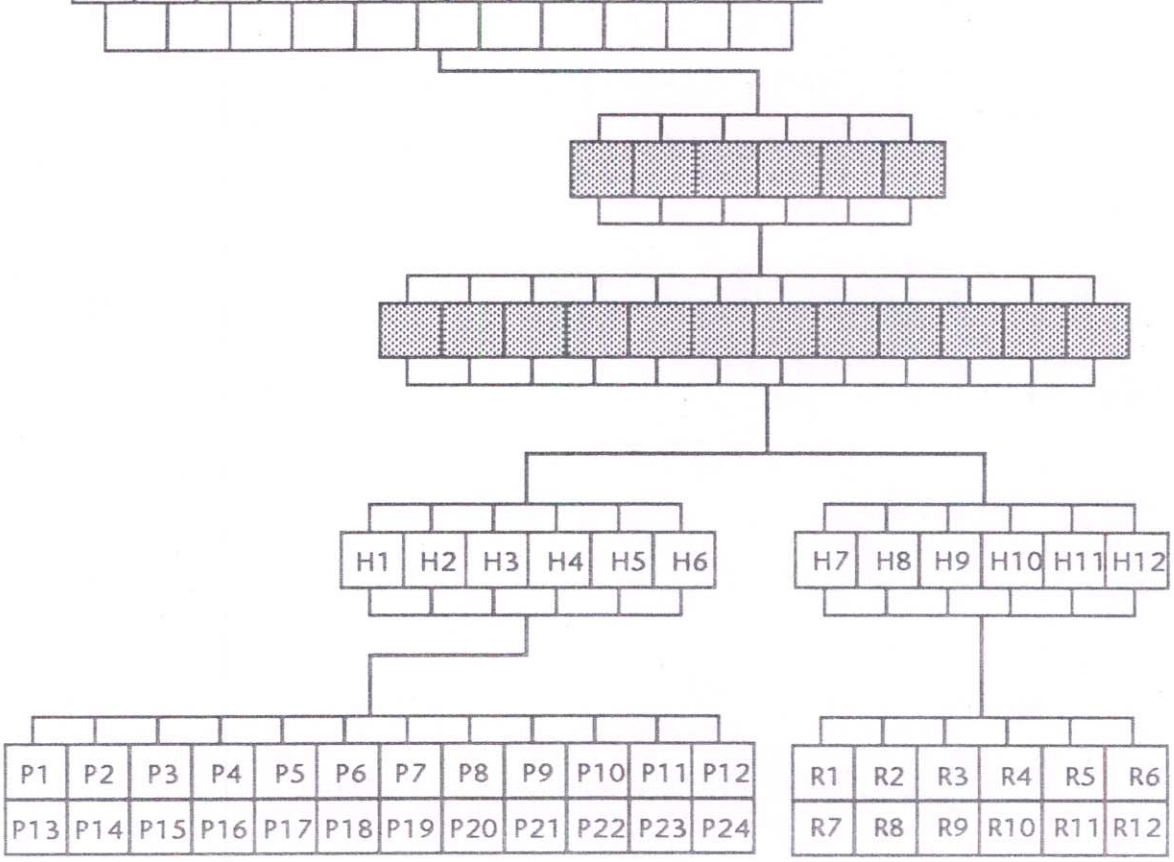


Figure 9.5. Five-Layer Network. A schematic representation of the network used by Hinton, with three layers of hidden units. P1 through P24 represent the individuals in Figure 9.4. R1 through R12 represent geneological relationships. The existence of a bottleneck forces the network to utilize a distributed representation.

unit 1 and (to a lesser degree) unit 5 identify family membership, while the remaining units are all generally indifferent to family memberships. Units 2 and 3 appear to encode the individual's generation (unit 2 responds most positively to older members, unit 3 to younger members). And units 4 and 6 seem to be representing membership in the two root families in each tree (unit 4 favoring the right side of the two trees, unit 6 the left).

Two comments are in order. First, it is the network, not the theorist, that determines what information to represent in the hidden units. Second, while it is often possible, as in this case, to label the hidden units in terms of what they represent, these labels are approximate; typically, it is difficult, and sometimes impossible, to fix these labels. For example, if unit 1 does indeed represent the English branch and unit 5 the Italian, they do not treat all members of the respective families equally. Penelope

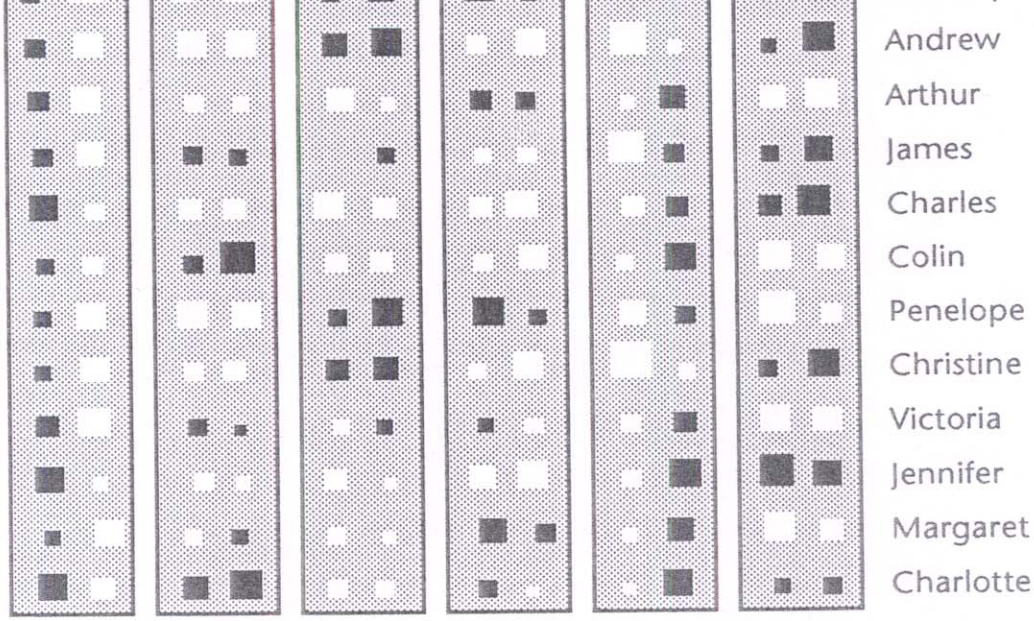


Figure 9.6. Six Hidden Units with Inputs from Twenty-Four-Person Units. Each node in the 6 hidden units represents one of the 24 individuals. The right-hand column in each of 1 through 6 represents a British individual, as indicated. The left-hand column within each represents the corresponding Italian family member (see Fig. 6.4). White boxes indicate a positive weight, so that, for example, inputs from any British name will tend to produce positive activation on all the units within the right-hand column of unit 1, while inputs from an Italian name will not. Black boxes indicate a negative weight, and box size indicates the level of activation. (From Hinton 1986.)

is treated by unit 1 as more of an English person than Charles, and Francesca is treated by unit 5 as more Italian than Emilio. In the six units encoding relationships (see Figure 9.7), unit 10 seems to represent the sex appropriate for the target person, but for others, such as units 9 and 12, it is difficult to specify what information is represented. Thus, while sometimes we are able to assign labels to the information-processing activities of the hidden units, these labels are partial and approximate. Moreover, the particular information-processing task a unit carries out may not be one we can describe at all. Thus, while we can construe multiple-layer networks as decomposing the information-processing task, the decomposition is not one advanced by the theorist and perhaps not even one the theorist can describe, at least not in the vocabulary in which the overall task is defined.

Connectionist models explain performance without explicitly or necessarily decomposing that performance into intelligible subtasks. Insofar as cognitive scientists accept connectionist simulations as explaining cogni-

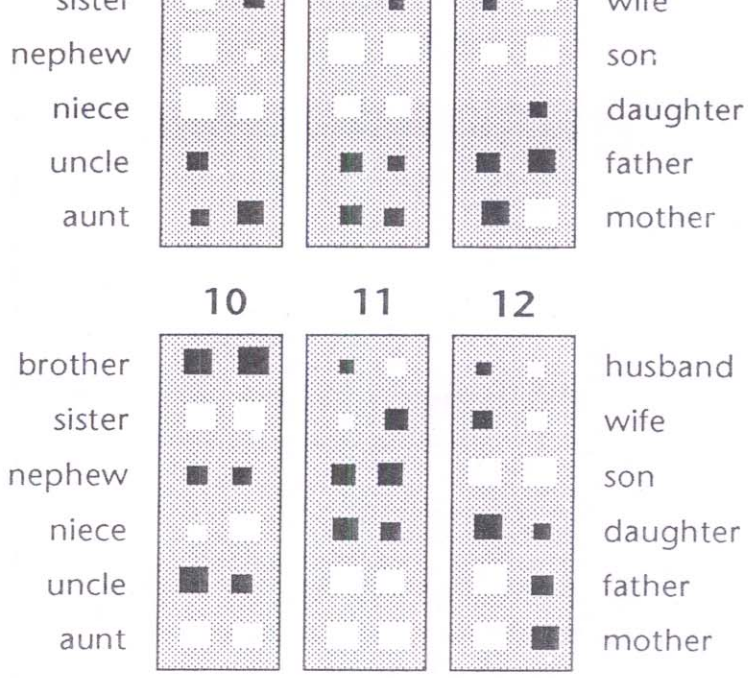


Figure 9.7. Units Recording Relationships and the Apparent Representation from Hidden Units. Projection from the 12 input units that represent relationships to the 6 in the second layer. In this case the representational functions are far less evident. (Also from Hinton 1986.)

tive performance, they are making a significant break with the decompositional strategy of traditional mechanism. It is still too early to determine how successful the connectionist strategy will be. It may be that connectionist approaches will only be useful for modeling low-level cognitive tasks such as visual perception and will fail in domains of reasoning and linguistic performance (cf. Fodor and Pylyshyn 1988; Pinker and Prince 1988). It is not important that we take a stand on the ultimate viability of connectionism as a framework for cognitive theorizing in order to make our main point: connectionism represents a break with traditional mechanism, pointing toward a different category of models and employing an alternative strategy for developing them. This alternative emphasizes systems whose dynamic behavior corresponds to the activity we want to explain, but in which the components of the system do not perform recognizable subtasks of the overall task. As with the case of Jackson's model, the decomposition of the system fails to correspond to cognitive organi-

the specific tasks performed by the components. We have abandoned decomposition and localization.

4. Distributed Mechanisms for Genomic Regulation

Stuart Kauffman (1986; forthcoming, chs. 9–13) has developed a network model for examining the structure and origin of genomic regulatory systems. His work can be understood as directed, in part at least, to the question of how genetic regulatory systems can be maintained in the face of genetic mutation and recombination.[16] One result of research on the mechanisms of gene expression in recent decades has been the recognition of complex sets of genes that regulate the expression of other genes. The activity of regulatory genes is probably relatively specific, but what is indicated by this work is a fairly complex network of genes connected either directly or indirectly; as a consequence, any mutations or transpositions affecting regulatory genes would alter the expression of the genetic system as much as—indeed, more than—would mutations in nonregulatory genes. The changes in these genes should have complex ramifications downstream. The more complex and interactive the regulatory system is, the more unstable and delicate it would seem to be, as it would offer more loci where the effects of mutation could have widespread effects.

Kauffman claims that unrealistically high selection pressures would be required to counter the mutation rates and maintain a regulatory system, once it becomes sufficiently complex. To give a definite form to the claim, he develops models with an arbitrarily defined set of "correct" connections and takes fitness to be a function of the frequency of such connections. If we assume a fixed mutation rate and a basal fitness of 0, then the class of fitness functions is given by $W_x = (G_x/T)^a_1$, where T is the total number of regulatory connections in the network, and G_x is the number of the "correct," or "good," connections. The value of a corresponds to three ways fitness can vary with changes in the frequency of good connections: If $a = 1$, then fitness falls off linearly as the frequency of bad connections increases. If $a > 1$, the fitness function is concave, falling off steeply at first and then leveling off. If $a < 1$, the fitness function is convex, falling off slowly at first and then more rapidly. If we suppose the mutation rate is constant, then, Kauffman reasons, because fitness is inversely proportional to T, increasing the number of regulatory connections will mean a decrease in the significance of selection; that is, increasing T will decrease the absolute fitnesses of alternatives and thereby decrease their absolute-fitness differences.[17] Mutation pressures will then be more likely to over-

someone with a more conventional viewpoint would perceive the complexity as an adaptation maintained in the face of selection and would confront the problem of identifying suitable forces, from Kauffman's vantage point the stability of the regulatory system may not be something that requires a special explanation in terms of selection. Kauffman suggests that, in fact, the genome is spontaneously self-organizing. To support this suggestion he again examines statistical features of systems with multiple interconnections. He shifts from seeking a highly localized explanatory unit to exploring the power and structure of a distributed system. Decomposition and localization would assume that one could identify discrete genetic units responsible for specific characteristics of the system, including regulatory control. Accordingly, some researchers have tried to identify specific connections between regulator genes and the genes regulated. Kauffman looks instead at the patterns of interactions between genes, not at the specific effects of any genes. In many cases, he says, the self-organizing pattern of connections in the genome may be the critical determinant of the genome's behavior, rather than any gene or set of genes that must be maintained by selection.

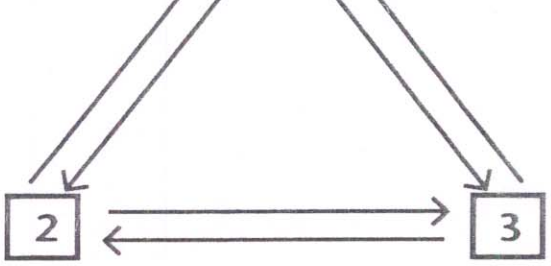
Kauffman's model networks have a fixed number of "genes," treated as nodes in the model, with a set number of random connections between them. These connections provide the vehicle for one gene to regulate the behavior of another. Kauffman then analyzes the properties of such networks. The ratio of the number M of connections to the number N of genes affects the expected number of genes any one gene potentially influences directly. This in turn affects (1) how many steps are required for a gene to communicate to all that it regulates and (2) the chances of a given gene receiving feedback, via a loop, from itself. In networks where M exceeds N, connected circuits typically form: some genes have regulatory influence on most others, many genes lie on feedback-loops and eventually receive activation from themselves. Just beyond the point where the ratio of M to N exceeds 1.0, both the average radius of effect of each such gene and the mean length of feedback-loops between genes increase. If M then becomes much larger than N, however, both the average radius and mean length of feedback-loops fall as more and more genes are connected directly.

Kauffman's model regulatory networks (forthcoming, ch. 5) rely on simple Boolean operations. That is, he confines himself to networks where each unit (gene) is limited to one of two values (on, off) and in which there are deterministic transitions according to Boolean operations. He focuses

as *canalizing* and says that the majority of known genes in bacteria and phage are governed by such Boolean functions (1986, p. 174).[18] He then shows that even when such systems are confronted with random alterations—that is, mutations—they will exhibit highly ordered regulatory behaviors. There will be a spontaneous, natural order to the system.

As a simple illustration, consider the three-unit system depicted in Figure 9.8: each unit sends activation to the other two, and the response of each to the incoming signals is a Boolean AND or OR operation. The value of unit 1 is + if the values of both units 2 and 3 were + on the previous cycle; otherwise it assumes a − value. Unit 1 is governed by an AND function. Units 2 and 3 will record a + value if any other unit assumed a + value in the previous cycle; otherwise they will assume a − value. Units 2 and 3 are governed by an OR function. Both functions are canalyzing, as unit 1 will go to − if either unit 2 or unit 3 is − on the previous cycle, and units 2 and 3 will go to + if either of the other units are + on the previous cycle. A system with three Boolean nodes has eight possible patterns of activation. Assuming the network is synchronously updated—that is, a pattern at time t completely determines the new pattern to be found at $t + 1$—then, because there are a finite number of states, and deterministic transitions, the system will inevitably encounter cycles. Once it enters a cycle, it will repeat it indefinitely. These cycles are what Kauffman calls *dynamical attractors* or *attractor state cycles*. For example, the alternation between $\langle - - + \rangle$ and $\langle - + - \rangle$ is a stable cycle. Which cycle a given network will settle into depends totally upon the starting point.

Kauffman suggests that Boolean nets have a number of properties that make them suggestive for understanding genomic regulation. If we limit our attention to cases in which $M/N > 1$ and to relatively large systems in which N is on the order of 10^4 or 10^5, then these cases will also find dynamical attractors. It turns out that such systems have parallels with the behavior of eukaryotic cells. For example, the number of cell types in an organism and the number of attractor state cycles in a network are both roughly the square root of the number of constituent genes. More importantly, both turn out to be very stable in the face of mutations. Kauffman has carried out computer simulations in which mutations occur as random alterations in the wiring diagrams of a population of 100 networks. Given appropriate values for M and N, the stability in the model networks is so great that 90% of possible perturbations in these networks leave no net change on the overall stable state of the network. Even when units are deleted from the network, only 10–15% of other units alter their pattern



+	−	+
−	+	+
−	−	−

2	3	1
+	+	+
+	−	−
−	+	−
−	−	−

1	3	2
+	+	+
+	−	+
−	+	+
−	−	−

Figure 9.8. A Simple Boolean Network with Three Nodes. In this case we have three nodes influencing each other. Unit 1 computes an "and" function, assuming a + value at step n + 1 if and only if both units 2 and 3 assume a + value at step n. Units 2 and 3 compute "or" functions, assuming a + value at step n + 1 if and only if at least one of the other two units assume a + value at step n. The system will exhibit simple cycles, or attractors; for example, both the case where all units are on and the case where all are off are stable, or attractor, states.

of expression. The reason for the stability of such systems and their ability to migrate to only a few alternative states is that a majority of the units in the system settle into fixed activation states that do not alter as the system cycles, thereby producing large *forcing structures* which inhibit propagation from other units through the system. Only the remaining isolated networks are capable of undergoing change, so they alone determine the range of variability in the system.

If the genome consists of large ensembles of genes mutually influencing and regulating one another, then there will be a stable, inherent order to the system independent of the influence of selection. The networks of genes will tend to be stable, independently of selection. As Kauffman puts it, given "statistical theories of the expected structure and behavior" of genomic regulatory networks, then those aspects of structure and behavior that would be predicted constitute "typical or generic properties of the ensemble of genomic regulatory networks" (forthcoming, ch. 11).

terization of the null state, evolutionists are prone to attribute too much power to natural selection, viewing it as capable of generating almost any possible genetic state that is highly adaptive. In contrast, Kauffman contends that the null state has quite powerful properties which selection is generally powerless to overcome:

> A general implication of [this class of models] is that a sufficiently complex genetic regulatory wiring diagram will approach arbitrarily close to the typical organizational properties of the unselected system. Thus, for sufficiently complex genomic systems, predications from the typical properties to be expected in the absence of further selection to those actually found in organisms would be reasonably accurate. (1986, p. 180)

Kaufmann's vision of the genetic regulatory system is close in spirit and method to connectionist models of psychological functioning. In both, it is ultimately the statistical pattern of the connections in the system, and not the jobs performed by specific units in the system or outside the system, that is critical to the behavior of the system. One respect in which the case examined by Kauffman is interestingly different from that discussed by connectionists is that, on the basis of his network model of the genetic regulatory system, Kauffman argues for a need to revise the question of what needs to be explained. What *appeared to* require explanation was the ability of selection to maintain the regulatory system. If Kauffman is correct, however, this will turn out not to be an issue. To the degree the regulatory system is stable, it is so *independently* of selection; the stability turns out to be due to self-organizing features of the genetic network itself. What requires explanation through selection is not how the system maintains its stability, but how it can be transformed from the stable state.

5. Conclusion: Mechanistic Explanations without Functional Decomposition and Localization

We have briefly described three cases in which researchers have pushed beyond classical mechanistic views. In Part III we saw that traditional mechanism, guided by localization and decomposition, explains why a system behaves as it does in terms of the behavior of its individual components. Even the explanations of complex systems with a variety of integrated circuits still make a major appeal to the contributions of specific modules in the system. In the cases sketched in this chapter, by contrast, this strategy is abandoned; instead, the approaches attempt to show how the properties of the system emerge simply as a result of the connectivity

surprising feature of these networks is that the pattern of connections results in systemic properties that would not be anticipated by focusing on the contributions of component units.

While network models are not classical mechanistic models, there is still a clear sense in which they are mechanistic. The behavior of the system is a product of the activities occurring within it. All the components are simple mechanical units, and their interactions are all characterized in simple mechanical terms. If the models are well motivated, then component function will at least be consistent with physical constraints. The difference is that what is important in determining the behavior of the system in a network model is not the contribution of the parts, but their organization. In simpler network models the parts are interchangeable; indeed, they are typically simple on/off units. Their role is to excite or inhibit the activities of other units in the system. The connections within the system determine the patterns of behavior that are observed in the system. There is clearly no case here for abandoning a mechanistic perspective. Nonetheless, these systems defy the approach to mechanism that we charted in earlier chapters, because these systems are neither decomposable nor even minimally decomposable, and systemic functions cannot be localized. Whether we are interested in cognitive capacities or genomic regulation, analyzing the components of the system in isolation throws no light on the phenomenon under investigation. One can only produce the phenomenon in the whole (or nearly whole) system. Analytic techniques that focus on the behavior of individual components through excitatory or inhibitory studies will fail; moreover, synthetic approaches are not liable to reveal component structure or organization. As a consequence, localization and decomposition break down with network systems.

Connectionist systems thus defy some of our traditional tools for studying natural systems, for these tools rely on being able to decompose the system, work on components singly, and then build up again to understand the whole. It should not surprise us that there are such connectionist systems when we recall the fallibility of heuristics. For many problems localization and decomposition have been highly successful, either in giving accurate accounts of how certain systems work, or accounts that constitute good approximations. When natural systems are not nearly, or at least minimally, decomposable, then those heuristics will lead us astray. Human cognition and genetic regulation may turn out to be processes for which localization and decomposition fail. Insofar as connectionist systems are intrinsically parallel systems, it is not possible to understand

parallel system serially, characterizing the transformations the system undergoes, the operating principles and architecture are fundamentally different. To understand the behavior of these systems it is necessary to understand that a multitude of such changes occur simultaneously.[19]

The development of network models is likely to alter our conception of machines and mechanism in radical ways. In particular, it may alter our understanding of the respects in which properties of a mechanistic system can be said to be *emergent* (cf. Bechtel and Richardson 1992). Advocates of emergentism have maintained that certain kinds of systems are capable of giving rise to radically new properties not present in the components of the system. Such appeals to emergence have struck many as mysterious, others as trivial. On the one hand, it is too easy to see that when a certain degree of complexity is reached in a system, new properties will appear: a square has properties none of the component line-segments have. On the other hand, emergence is vacuous unless we have some account of how the organization matters. In network systems we can understand how emergent properties appear without waxing mysterious. In calling the systemic properties of network systems emergent, we mark a departure from the behavior of simpler systems and indicate that traditional mechanistic strategies for understanding network systems may simply fail. But the behavior of the system is not unintelligible or magical; it follows from the nature of the connections between the components within the system.

The emergent behavior of connectionist systems may be mysterious in another sense, however: We may not be able to follow the processes through the multitude of connections in a more complex system, or to see how they give rise to the behavior of the system. We may fail in the attempt to understand such systems in an intuitive way. To quote Simon once again, complex systems that are not hierarchical and decomposable "may to a considerable extent escape our observation and understanding" (1969, p. 219).