7 Neuroscience

1 Neuroanatomy: The Evolutionary Background

Near the surface of the Earth's oceans, between three and four billion years ago, the Sun-driven process of purely chemical evolution produced some *self-replicating* molecular structures. From the molecular bits and pieces in their immediate environment, these complex molecules could catalyze a sequence of bonding reactions that eventually yielded exact copies of themselves. With respect to achieving large populations, the capacity for self-replication is plainly an explosive advantage. Population growth will be limited, however, by the availability of the right bits and pieces in the molecular soup surrounding, and by the various forces in the environment that tend to break down these heroic structures before they can replicate themselves. Among competing self-replicating molecules, therefore, the advantage will go to those specific molecular structures that induce, not just their own replication, but the formation of structures that protect them against external predations, and the formation of mechanisms that produce needed molecular parts by the direct chemical manipulation of environmental molecules that are unusable directly.

The *cell* is the triumphal example of this solution. It has an outer membrane to protect the intricate structures within, and complex metabolic pathways that process outside material into internal structure. At the center of this complex system sits a carefully coded DNA molecule, or a family of them, the director of the cellular activity and the winner of the competition described. Such cells now dominate the Earth. All competitors have been swept aside by their phenomenal success, save for the residual viruses, which alone pursue the earlier strategy, now as parasitic invaders upon cellular success. With the emergence of the cell, we have what fits our standard conception of *life*: a self-maintaining, self-replicating, energy-using system.

The emergence of conscious intelligence, as one aspect of living matter, must be seen against the background of biological evolution in general. We here pick up the story after it is already well along: after multicelled organisms have made their appearance, close to one billion years ago. Significant intelligence requires a nervous system, and single-celled organisms such as algae or bacteria cannot have a nervous system, since a nervous system is itself an organization of many cells.

The main advantage of being a multicelled organism (a *metazoan*) is that individual cells can be specialized in their biological function. Some can form a tough outer wall, within which other cells can enjoy an environment more stable and more beneficial than the ocean at large. These cloistered cells, in turn, can exercise their own specializations: digestion of food, transport of nutrients to other cells, contraction and elongation to produce diverse movements, sensitivity to key environmental factors (the presence of food or predators), and so on. The result of such organization can be a system that is more durable than any of its parts, and far more likely to succeed in reproducing itself than is any one of its single-celled competitors.

The coordination of these specialized parts requires *communication* between cells, however, and some additional specializations must address this important task. It is no use having muscle cells if their contractions cannot be coordinated to produce useful locomotion, or mastication, or elimination. Sensory cells are useless if their information cannot be conveyed to the motor system. And so on. Purely chemical communication is useful for some purposes: growth and repair is regulated in this way, with messenger cells broadcasting specific chemicals throughout the body, to which selected cells respond. But this is too slow and unspecific a means of communication for many purposes.

Fortunately, cells themselves have the basic features needed to serve as communicative links. Most cells maintain a tiny voltage difference—a *polarization*—across the inner and outer surfaces of their enveloping cell membranes. An appropriate disturbance at any point on that membrane can cause a sudden *de*polarization at that point and, like the collapse of a train of dominoes stood precariously on end, the depolarization will *spread* some distance along the surface of the cell. After this depolarization, the cell gamely pumps itself back up again. In most cases the depolarization pulse attenuates and dies in a short distance, but in others it does not. Conjoin this convenient property of cells with the fact that some cells have extremely elongated shapes—filaments of a meter or more in extreme cases—and you have the perfect elements for a long-distance communication system: specialized nerve cells that conduct electrochemical impulses over long distances at high speed.

Further specializations are possible. Some cells depolarize upon receipt of external physical pressure, others upon changes in temperature, others upon sudden changes in incident light,

and still others upon receipt of suitable impulses arriving from *other* cells. With the articulation of such cells we have the beginnings of the sensory and central nervous system, and we open a new chapter in the evolutionary drama.

The Development of Nervous Systems

The appearance of nervous control systems should not be seen as something miraculous. To appreciate just how easily a control system can come to characterize an entire species, consider an imaginary snail-like creature that lives on the ocean bottom. This species must come partway out of its shell to feed, and it withdraws into its shell only when the creature is sated, or when some external body makes direct contact with it, as when a predator attacks. Many of these creatures are lost to predators, despite the tactile withdrawal reflex, since many are killed at the very first contact. Even so, the species' population is stable, being in rough equilibrium with the population of predators.

As it happens, every snail of this species happens to have a band of light-sensitive cells on the back of its head. In this there is nothing remarkable. Many types of cell happen to be light-sensitive to some degree, and the light sensitivity of these is an incidental feature of our species, a feature that does nothing. Suppose now that an individual snail, because of a small mutation in the coding of its initial DNA, has grown more than the usual number of nerve cells connecting its skin surface to its withdrawal muscles. In particular, it is alone among its conspecifics in having connections from its light-sensitive cells to its withdrawal muscles. Sudden *changes* in the general illumination thus case a prompt withdrawal into its shell.

This incidental feature in this one individual would be of no significance in many environments, a mere idiosyncratic 'twitch'

of no use whatever. In the snail's actual environment, however, sudden changes in illumination are most often caused by *predators* swimming directly overhead. Our mutant individual, therefore, possesses an 'early warning system' that permits it to withdraw safely *before* the predator gets to take a bite. Its chances of survival, and of repeated reproduction, are thus much greater than the chances of its unequipped fellows. And since its novel possession is the result of a genetic mutation, many of his offspring will share in it. Their chances of survival and reproduction are similarly enhanced. Clearly, this feature will swiftly come to dominate the snail population. Of such small and fortuitous events are great changes made.

Further exploitation is easily conceived. If, by further genetic mutation, a light-sensitive surface becomes curved into a hemispherical pit, its selectively illuminated portions will then provide *directional* information about light sources and occlusions, information that can drive directional motor responses. In a chronically mobile creature like a fish, this affords a major advantage, both as hunter and as prey. Once widely distributed, a hemispherical pit can be genetically modified into a nearly spherical pit with only a pinhole opening left to the outside. Such a pinhole will form a literal *image* of the outside world on the light-sensitive surface inside. Transparent tissue can come to cover that pinhole, functioning first as protection, and later as a *lens* for superior images. All the while, increased innervation (concentration of nerve cells) in the 'retina' is rewarded by superior information to conduct elsewhere in the creature's nervous system. By such simple and advantageous stages is the 'miraculous' eye assembled. And this reconstruction is not sheer speculation. A contemporary creature can be found for each one of the developmental stages just cited.

In general, our ongoing reconstruction of the evolutionary history of nervous systems is based on three sorts of studies: fossil remains, current creatures of comparatively primitive construction, and nervous development in the embryos of all creatures. Being so soft, nervous tissue does not itself fossilize, but we can still trace nervous structure in ancient vertebrates (animals with a backbone) from the chambers, passages, and clefts found in the skulls and spinal columns of fossil animals. This is a very reliable guide to size and gross structure, but fine detail is mostly missing. For detail, we turn to the existing animal kingdom, which contains thousands of species whose nervous systems appear to have changed little in the course of many millions of years. Here we have to be careful, since "simple" does not necessarily mean "primitive," but we can reconstruct very plausible developmental 'trees' from such study. Embryological development proves a fascinating check on both studies, since some (only *some*) of any creature's evolutionary history is written in the developmental sequence by which DNA articulates a fertilized egg cell into a creature of the relevant type. Putting all three together, the following history emerges.

The most primitive vertebrates possessed an elongated central *ganglion* (cluster of cells) running the length of the spine, which was connected to the rest of the body by two functionally and physically distinct sets of nerve fibers (figure 7.1). The *somatosensory* fibers brought information about tactile sensations and muscle activity back to this central cord, and the *motor* fibers took command impulses from it out to the body's muscle tissues. The central cord itself functioned to *coordinate* the body's many muscles to produce a coherent swimming motion, and to coordinate such motion with sensed circumstance, to provide flight

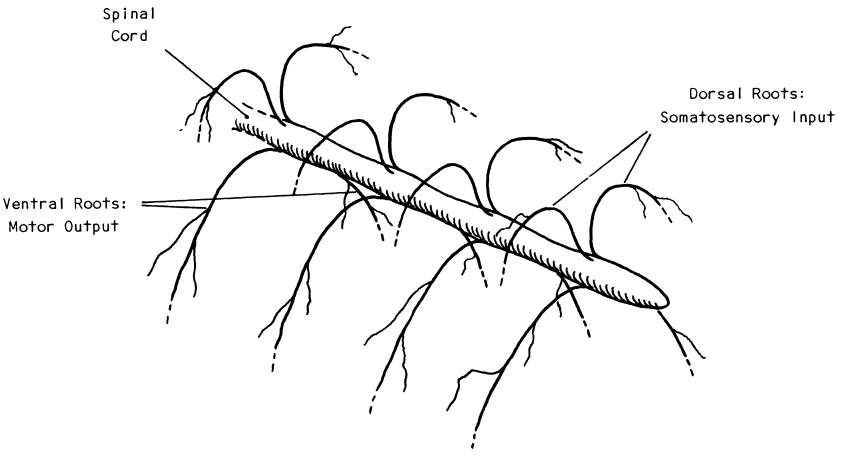

Figure 7.1

from tactile assault or a searching motion to relieve an empty stomach. A simple creature like the modern leech is still an instance of this pattern.

In later creatures this primitive *spinal cord* has acquired an elongation at the front end, with three swellings where the population and density of nerve cells reach new levels. This primitive brain or *brain stem* can be divided into the *forebrain*, *midbrain*, and *hindbrain* (figure 7.2). The nervous network of the small forebrain was then devoted to the processing of olfactory stimuli; the midbrain processed visual and auditory information; and the hindbrain specialized in still more sophisticated coordination of motor activity. The brains of contemporary fishes remain at this stage, with the midbrain the dominant structure.

In more advanced animals such as amphibians and reptiles, it is the forebrain that comes to dominate the brain stem's

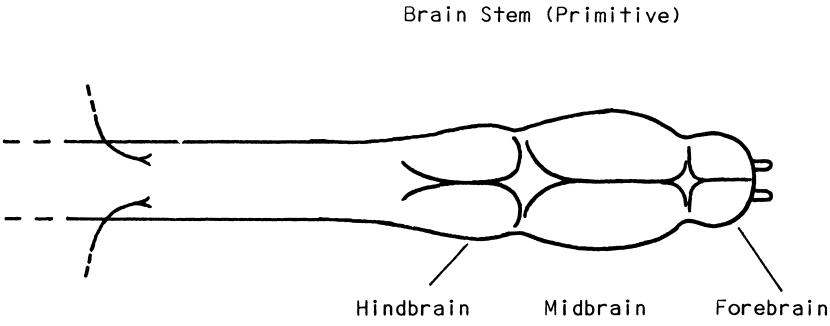

Figure 7.2

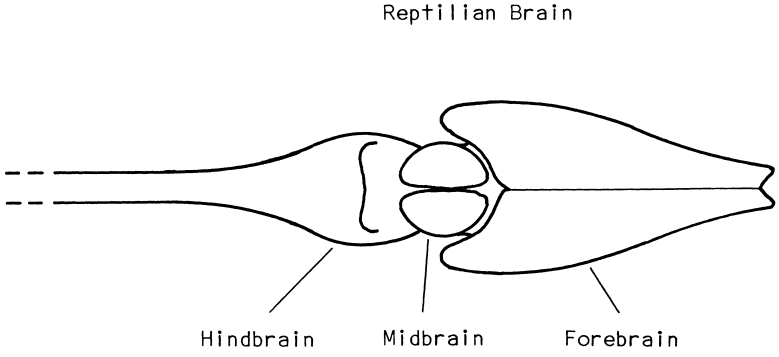

Figure 7.3

anatomy, and to assume a central role in processing all of the sensory modalities, not just olfaction (figure 7.3). In many animals, absolute size also increases, and with it the absolute number of nerve cells in what is already a complex and quasi-autonomous control network. That network had much to do: many dinosaurs were swift bipedal carnivores that pursued distant prey by means of excellent eyesight. A superior control

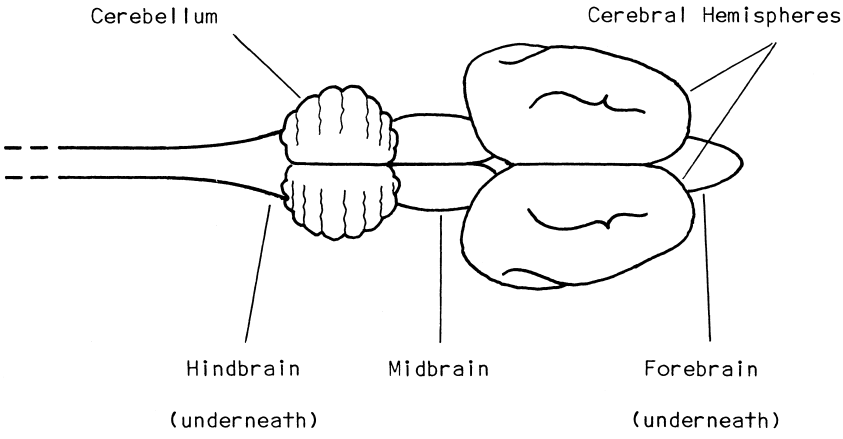

Figure 7.4

system was essential if that ecological niche was to be occupied successfully.

The early mammalian brain displayed further articulation and specialization of the forebrain, and most important, two entirely new structures: the *cerebral hemispheres* growing out each side of the enlarged upper forebrain, and the *cerebellum* growing out the back of the hindbrain (figure 7.4). The cerebral hemispheres contained a number of specialized areas, including the highest level of control for the initiation of behavior; and the cerebellum provided even better coordination of bodily motion in a world of objects in relative motion. The sheer number of neuronal cells in the cerebral and cerebellar cortex (especially in the thin surface at which the cell bodies and their intercellular connections are concentrated) is also strikingly larger than the number found in the more primitive cortex of reptiles. This cortical layer (the classical 'gray matter') is two to six times thicker in mammals.

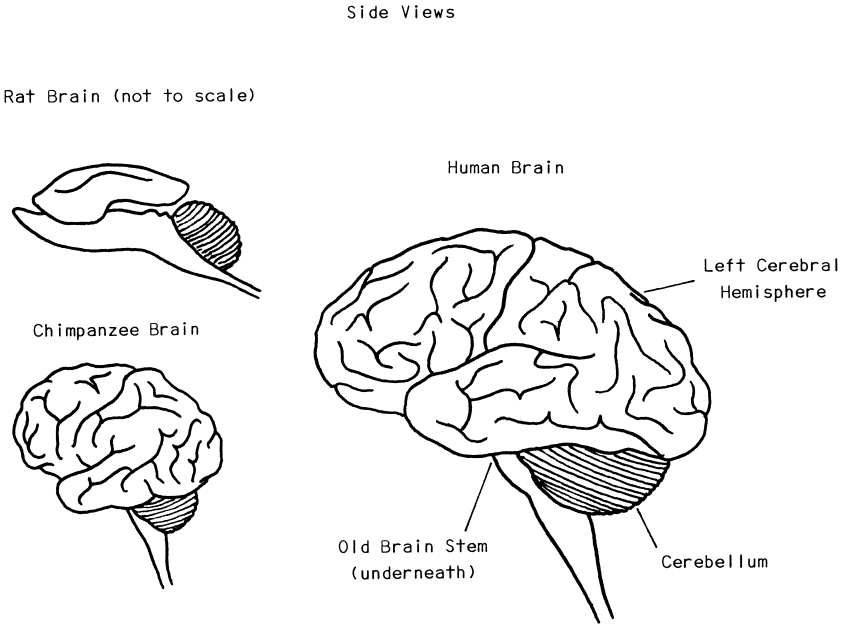

Figure 7.5

In typical mammals, these new structures, though prominent, are not large relative to the brain stem. In primates, however, they have become the dominant features of the brain, at least to the casual eye. And in humans, they have become enormous (figure 7.5). The old brain stem is now barely visible under the umbrella of the cerebral hemispheres, and the cerebellum is also markedly enlarged, compared to what other primates display. It is difficult to resist the suspicion that what distinguished us from the other animals, to the extent that we are distinguished, is to be found in the large size, the dense interconnections, and the unusual cognitive properties of the human cerebral and cerebellar hemispheres.

Suggested Readings

Bullock, T. H., R. Orkland, and A. Grinnell. *Introduction to Nervous Systems*. San Francisco: Freeman, 1977.

Sarnat, H. B., and M. G. Netsky. *Evolution of the Nervous System*. Oxford: Oxford University Press, 1974.

Dawkins, Richard. *The Selfish Gene*. Oxford: Oxford University Press, 1976.

2 Neurophysiology and Neural Organization

A The Elements of the Network: Neurons

Structure and Function

The elongated impulse-carrying cells referred to earlier are called *neurons*. A typical neuron has the physical structure outlined in figure 7.6: a treelike structure of branching *dendrites* for inputs, and a single *axon* for conveying outputs. (The axon is folded for purely diagrammatic reasons.) This structure reflects what appears to be the neuron's principal function, namely, the

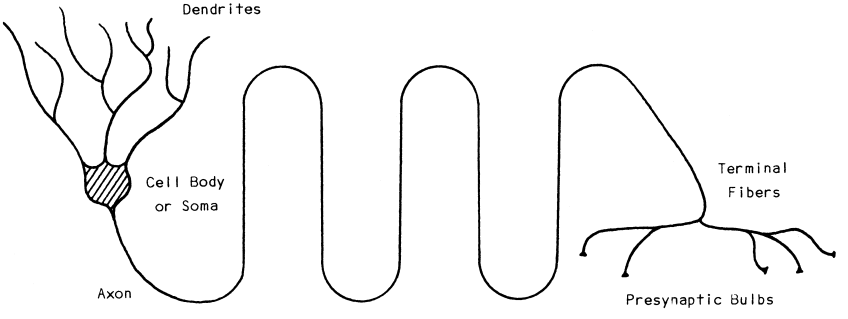

Figure 7.6

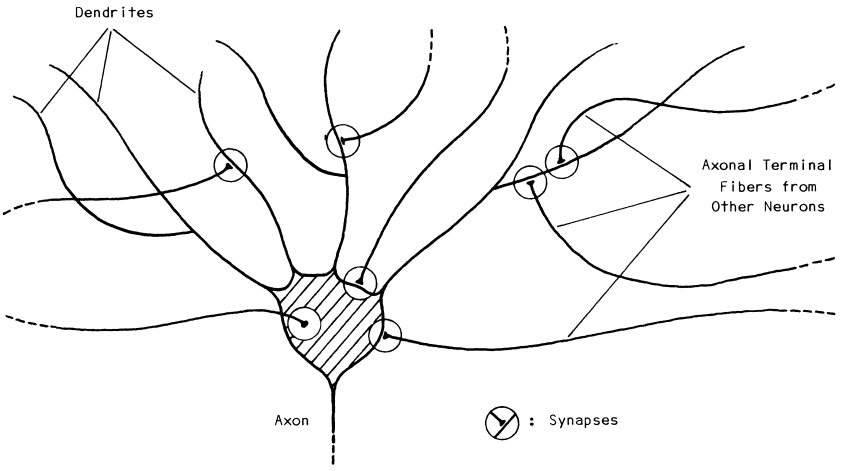

Figure 7.7

integration of inputs from many other cells. Typically, the axons of a great many other neurons—usually in the thousands—make contact either with the dendrites of the receiving neuron, or with the cell body itself. These tiny connections are called *synapses*, and they allow the events in one cell to influence the activity of another, indeed, of many others (figure 7.7).

The influence is achieved in the following ways. When a depolarization pulse—called an *action potential* or *spike*—runs all the way down its length to its many presynaptic endings, its arrival causes the terminal end-bulbs to release a chemical called a *neurotransmitter* across the tiny 'synaptic cleft' separating the arriving axon from the receiving dendrite. Depending on the nature of the bulb's characteristic neurotransmitting chemical, and on the nature of the chemical receptors that receive it on the opposite side of the cleft, the synapse is called either an *inhibitory* or an *excitatory* synapse.

In an inhibitory synapse, such a cross-synaptic transmission causes a slight *hyper*polarization or *raising* of the affected neuron's electric potential. This makes it *less* likely that the affected neuron will undergo a sudden depolarization of its own membrane and fire off a spike along its own axon.

In an excitatory synapse, by contrast, the chemical transmission across the cleft causes a slight *de*polarization of the affected neuron, inching its electric potential downward toward the critical minimum point where it suddenly collapses entirely, initiating its own output spike down the length of its own axon. An excitatory synaptic event therefore makes it *more* likely that the affected neuron will fire.

Putting the two factors together, each neuron is the site of a competition between 'fire' and 'don't fire' inputs. Which side wins is determined by two things. First, the relative distribution of excitatory and inhibitory synapses matters greatly—their relative numbers, and perhaps their proximity to the main cell body itself. If one kind predominates, as often it does, then the deck is 'stacked' for that neuron, in favor of one response over the other. (In the very short term, these many connections are a relatively stable feature of each neuron. But new connections do grow, and old ones are lost, sometimes on a time scale of mere minutes; hence the functional properties of a given neuron are somewhat plastic.)

The second determinant of the receiving neuron's response is the sheer temporal frequency of inputs from synapses of each kind. If 2,000 inhibitory synapses are each active only once per second, while 200 excitatory synapses are each active a busy 50 times per second, then the excitatory influences will predominate and the neuron will fire. After repolarization, it can fire again, and again, with a significant frequency of its own.

It is well to keep in mind the relevant numbers here. A typical neuronal *soma* (= central cell body) will be almost buried under a layer of several hundred synapsing end-bulbs, and its dendritic tree may enjoy synaptic connections with several thousands more. As well, neurons pump themselves back up to resting potential again in rather less than 1/100th of a second; hence they can sustain spiking frequencies of up to 100 hertz (= 100 spikes per second), or even more. Evidently, a single neuron is an information processor of considerable capacity. Inevitably, neurons are likened to the logic gates in the CPU of a digital computer. But the differences are more intriguing than the similarities. A single logic gate receives simultaneous inputs from no more than *two* distinct sources; a neuron receives simultaneous inputs from well in excess of a thousand. A logic gate emits outputs at a constant, metronomic frequency, 10^6 hertz, for example; whereas a neuron's output varies continuously between 0 and 100 spikes per second. Logic-gate output is and must be rigidly coordinated with that of other gates; neuronal outputs are not thus coordinated. The function of a logic gate is the transformation of binary information (sets of ONs and OFFs) into further binary information; the function of a neuron seems more plausibly to be the transformation of sets of spiking *frequencies* into further spiking *frequencies*. And last, the functional properties of a logic gate are fixed; those of a neuron are decidedly plastic, since the growth of new synaptic connections and/or the enhancement of their individual strengths, on the one hand, or the loss of old synaptic connections or the reduction of their individual strengths, on the other, can *change* the input/output function of the neuron dramatically. These changes are induced, in large measure, by the prior activities of the cell itself.

If neurons are information-processing devices, as almost certainly they are, their basic mode of operation is therefore very different from that displayed in the logic gates of a digital CPU. This is not to say that systems of the latter, suitably programmed, could not simulate the activities of the former. Presumably they could. But we need to know rather more about the plastic functional properties of neurons, and we need to take into account much more about their myriad interconnections, before we can successfully simulate their collective activity.

Types of Neurons

An initial classification finds three kinds of neurons: *motor* neurons, *sensory* neurons, and a large variety of *interneurons* (that is, all the rest). Primary motor neurons are found almost exclusively in the spinal cord, and are defined as those neurons whose axons synapse directly onto a muscle cell. The axons of the primary motor neurons are some of the longest in the nervous system, extending from deep within the spinal cord, out the *ventral roots* (see figure 7.1) between the spinal vertebrae, and on out the limbs to the most distant peripheral muscles. Motor neurons secure graded muscle contraction by two means: the spiking frequency of individual motor neurons, and the progressive recruitment of initially quiescent neurons that innervate the same muscle.

Sensory neurons come in greater variety, and they are conventionally defined as those whose input stimulus is some dimension of the world outside the nervous system. For example, the rod and cone receptor cells of the retina are very tiny, with little axon to speak of, and no dendrites at all. They synapse immediately onto more typical neurons in a layer right next to them. Their job is solely to transform received light into

discriminatory synaptic events. The somatosensory cells, by contrast, are as long as the motor neurons. Their axons project from the skin and muscles into the spinal cord by way of the *dorsal roots* (see figure 7.1), and they find their first synapses deep within the spinal cord. Their job is to convey tactile, pain, and temperature information, and information about muscle extensions and contractions—that is, the ever changing positions of the body and its limbs. Other sensory cells have their own idiosyncrasies, dictated by the nature of the physical stimulus to which they respond.

The central interneurons also come in a great variety of shapes and sizes, though they all seem to be variations on the same theme: multiple dendritic inputs and a single axonal output. Most, called multipolar cells, have many dendritic branches emerging directly from the cell body. Some, such as the Purkinje cells of the cerebellum, have extraordinarily extensive and bushy dendritic trees. Others enjoy only sparse dendritic extensions. The axons of many neurons project across the entire brain, synapsing at distant points within very different neuronal populations. Others make merely local connections among extended concentrations of neurons whose axons project elsewhere.

These densely populated layers of heavily interconnected neural bodies are called *cortex*. The outer surface of each cerebral hemisphere is one large sheet of cortex, heavily folded upon itself like crumpled paper so as to maximize the total area achieved within the limited volume of the skull. The brain's interneural connections are at their heaviest within this folded area. The surface of the cerebellum is also cortex, and functionally specialized cortical 'nuclei' are distributed throughout the brain stem. These show as gray areas in brain cross-sections. The

remaining white areas contain the axonal projections from one cortical area to another. Which brings us to the matter of the brain's overall organization.

B The Organization of the Network

Seeking organization in a network as complicated as the human brain is a difficult business. Much structure has emerged, but as much or more remains either hidden, or functionally opaque, or both. One can explore the large-scale structure of neuronal interconnections by using special *stains* that can be injected into a neuron so as to diffuse down its axon all the way to its terminal synaptic end-bulbs. If we wish to know where the axons of a stained cortical area actually project to, successive cross-sections of the postmortem brain will reveal both the path that those stained axons take through the comparatively colorless volume that contains them, and the region of their ultimate terminus in some new population of neurons. This technique (*Golgi stains*, named after their original inventor) has revealed the major interconnections between the various cortical areas of the brain, the 'superhighways' that involve many thousands of axons strung out together. Knowing their locations does not always reveal their functions, however, and the smaller neuronal highways and byways constitute a horizon of ever-shrinking detail that defies attempts at complete summary.

With microscopes, thin sections, and a variety of further staining techniques, the microarchitecture of the brain begins to emerge. Cerebral cortex, for example, reveals six distinct neuronal layers, distinguished by the density of the neuronal populations within them, by the types of neurons that they contain, and by the proprietary (usually short) connections they make with the other cortical layers. Interneuronal communication is

evidently quite extensive, both within layers and across them. The details are complex and obscure, and the point of this particular arrangement remains mostly obscure, but we cling to what order we do discover, and use it to try to find more. As it happens, the six-layered architecture just mentioned is not entirely uniform across the cerebral cortex: the thickness and density of certain layers is diminished or exaggerated in certain areas of the cortical surfaced. Tracing areas of *identical* architecture, and noting their boundaries, has led us to identify about fifty distinct cortical areas, known as *Brodmann's areas* after the microscopist who mapped them out, areas that are the same across all normal humans.

Are these areas of any further significance? Indeed they are, both in their functional properties and in their more distant axonal connections. A few salient cases will now be outlined.

Sensory Projections within the Brain

As mentioned earlier, the primary somatosensory neurons enter the spinal cord via the dorsal roots, and find their first synaptic connections with neurons in the cord. Those neurons conduct their information up the spinal cord all the way to the thalamus in the forebrain, where they synapse onto neurons in an area called the *ventral thalamic nucleus*. Finally, these neurons project in turn to the cerebral hemispheres, into a cortical area neatly defined by three connecting Brodmann's areas. This overall area is now known as the *somatosensory cortex*. Damage to various parts of it produces a permanent loss of tactile and proprioceptive awareness of various parts of the body. Moreover, subtle electrical stimulation of neurons in this area produces in the subject vivid tactile sensations 'located' in specific parts of the body. (Brain surgery to correct various problems with this part of the cortex

has provided the occasional opportunity for such probing, and since subjects can be wholly conscious during brain surgery, they can report the effects of such stimulations.)

In fact, the somatosensory cortex constitutes what is called a *topographic map* of the entire body, since the spatial arrangement of anatomically specific neurons is a systematic *projection* of the original anatomical areas themselves. Each hemisphere represents the opposite one-half of the body. The cross-section of one hemisphere in figure 7.8 illustrates the point. The distorted creature represents the areas of the cortex devoted to the body part next to it, and the variations in size represent the relative numbers of cortical neurons devoted to inputs from that part. This diagrammatic creature is called "the somatosensory homunculus."

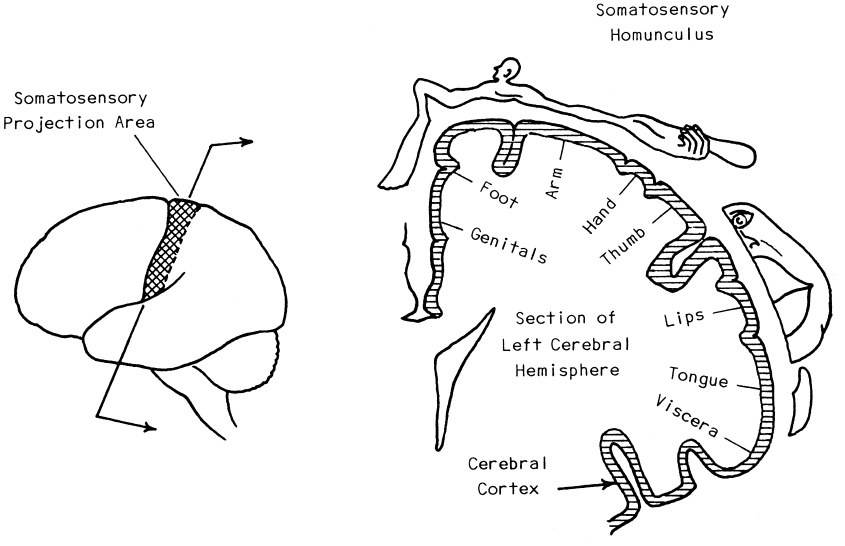

Figure 7.8

The organization and function of the *visual* system also makes contact with the Brodmannian structure of the cerebral cortex. If we look at the eye, we find that right next to the primary rods and cones of the retina, there is an interconnected layer of small neurons that performs some initial processing before synapsing onto the long *ganglion cells* at the back of the retina whose longish axons constitute the familiar *optic nerve*. The optic nerve projects its axons to an important thalamic area, deep within the brain, called the *lateral geniculate nucleus* or *LGN*. The cells here constitute a topographic *map* of the retinal surface, although it is metrically distorted in that the *fovea*, the physical functional center of the retina, is very heavily represented (i.e., it is magnified relative to the retinal periphery). The neurons in the LGN finally project to one of Brodmann's areas on the rearmost surface of the cerebral hemispheres called the *primary visual cortex*, and it embodies a topographic projection of the retina, with each hemisphere representing one-half of the retinal surface. But rather more is going on in the visual cortex, and in its precortical processing, than occurs in the somatosensory system, and the visual cortex represents rather more than just an area of retinal stimulation. Subpopulations of visual neurons turn out to be specialized, in their responses, to highly specific features of the visual information. A cell early in the processing hierarchy is sensitive only to brightness *differences* within its 'receptive field' (= the retinal subarea to which it is sensitive). But a higher cell, to which these early cells project, may be sensitive only to lines or edges of a particular *orientation* within its receptive field. Cells higher still are sensitive only to lines or edges that are *moving* in a particular direction. And so on. The impression of a *cumulative* information-processing system is impossible to escape.

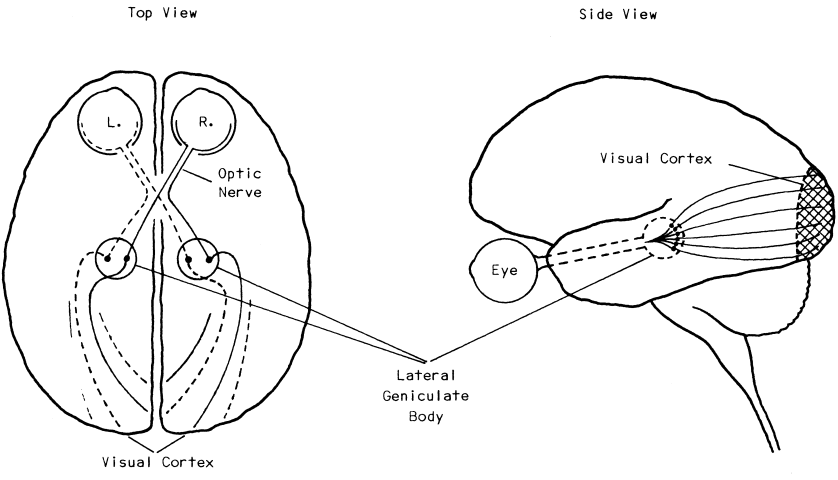

Figure 7.9

Further microstructures promise to explicate features of binocular vision—in particular, the sophisticated *stereopsis* or three-dimensional vision possessed by humans. Since the two human eyes see the external 3D world from two slightly different spatial perspectives, the two images of that world, formed on the right and left retinas, are subtly *different* from one another, and those differences embody detailed information about the differences in the relative *distances*, from the viewer, of the various objects portrayed in those two images. Intriguingly, those differences (in the right and left images) are faithfully preserved all the way up to the unitary surface of the primary visual cortex, where they overlap one another perfectly where *distant* objects are concerned, but show increasing *failures of mutual registry* in the case of perceived objects that are progressively *closer* to the viewer. This failure of mutual registry is called *binocular*

rivalry, and, on the face of it, is a recipe for nothing but visual confusion.

However, there is a population of neurons spread all across the visual cortex, each one of which is sensitive to precisely such local left/right representational disparities where they occur, *and* to the magnitude of those disparities. The collective behavior of those 'stereo cells' thus *marks* the appropriate areas of the visual cortex as representing an object that is progressively *closer* to the viewer as the relevant disparity increases. The result, as you know well, is the visual awareness of external objects at a variety of different *distances* from you. Their collective behavior gives you 3D vision.

This specific neuronal arrangement has been modeled (by this author) in an *artificial neural network* that also receives slightly disparate left/right retinal images as inputs, and gives coded visual areas as outputs, areas appropriately coded for external objects located at various distances from the viewer. It produces exactly the same input/output behavior displayed by the disparity-computing program discussed in the section on AI (see again pp. 181–182, and view once more the initially opaque stereo-pair of images). As realized in a human or animal brain, however, a biologically realistic neural arrangement of this kind will produce the desired cognitive output—a vivid 3D image—hundreds of times *faster* than does the programmed computer, because the local conformities-and-disparities (many *hundreds of thousands* of them) are all being responded to *simultaneously*, rather than one-by-one and in laborious sequence, as a serial computer is doomed to respond. As you look again at the stereo-pair of images on p. 182, you can see the hidden 3D structure within an instant of your 'fusing' the two images. The

AI program, by contrast, recovered that hidden structure only slowly and gradually, not all at once.

Here we have a sterling illustration of *why* the biological brain typically performs its cognitive functions so much more swiftly than do the serial-processing computers designed to simulate those very same input/output functions. The brain is making systematic use of a computing technique called *parallel distributed processing* or *PDP*, for short. This particular expression deliberately highlights the fact that the retinal-disparity information cited earlier is *distributed* across the entire population of neurons carrying information *to* the visual cortex, and the further fact that all of that distributed information is processed simultaneously or *in parallel* across that entire neuronal population. The relevant processing is done 'all at once', and so the process is *completed* in milliseconds. Unlike the computer program cited, you see that 3D structure in an instant, rather than as the end result of a long sequential process.

Motor Projections Outward

Just in front of the somatosensory cortex, on the other side of a major cleft, is another of Brodmann's areas known as the *motor cortex*. This area is also a clear topographic map, this time of the body's *muscle* systems. Artificial stimulation of these motor neurons ultimately produces movement in the body's muscles—those corresponding to the specific area of the map that was stimulated. This metrically deformed map or 'motor homunculus' is displayed in figure 7.10.

This is only the beginning of the functional story, of course, since motor control is a matter of well-orchestrated *sequences* of muscle contractions—sequences that cohere, moreover, with

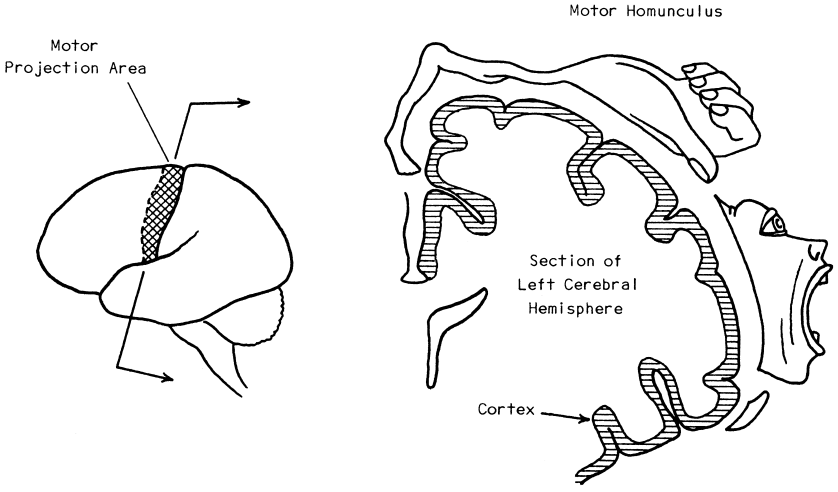

Figure 7.10

the body's perceived environment. Accordingly, the motor cortex has axonal projections, not just to the cord and thence to the body's muscles, but to the cerebellum and the basal ganglia, and it receives *reciprocal* projections from both, primarily through the centrally located thalamus, which we already know to be a source of sensory information. The motor cortex is therefore a highly integrated part of brain activity, and though some of its output goes more or less directly to the cord—to provide independent control of fine finger movements, for example—much of it goes through intricate processing in the cerebellum and the lower brain stem before entering the spinal cord.

We must think of the brain's cortical output here as a sort of high-level 'fine tuning' of motor capacities that are more basic still, since the neuronal organization within the spinal cord

itself is sufficient to produce basic locomotion in most vertebrates. A familiar example of this is the headless chicken whose body runs around aimlessly for several seconds after it has been slaughtered. Even small mammals whose brains have been substantially removed will display smooth locomotor activity upon gentle electrical stimulation of the spinal cord. We have here a reflection of just how *old*, evolutionarily speaking, the capacity for vertebrate locomotion is: it was first perfected when primitive vertebrates enjoyed little more than a spinal cord. The progressive additions to that basic neuronal machinery all survived because they added some useful fine tuning of, or intelligent guidance to, that initial capacity. The motor cortex is merely one of the later and higher centers in an extensive hierarchy of neuronal motor controls. These extend from the simple and cord-centered 'reflex arcs'—such as will withdraw a hand from a hot stove—up to the highest centers, which formulate abstract, long-term plans of action.

Internal Organization

The brain *monitors* the extra-neural world through one's primary sensory organs, of course. But in the process, it also monitors many aspects of its own internal operations. And the brain exerts *control* over the extra-nervous world. But it also exerts control over many aspects of its own internal operations. The internal projections among distinct parts of the brain are rich and extensive, and they are crucial to its proper functioning. A good example is the existence of 'descending control' mechanisms. In our earlier discussion of the visual system I did not mention that the visual cortex at the rear of the brain not only receives axonal projections from the LGN, as shown in figure 7.9, it also sends massively many axonal projections from its

own neurons *back* to the LGN, where the optic nerve originally terminates. What this means is that, via these 'descending' pathways, the visual cortex can exert an influence on the LGN to *modulate* what is being sent upward, perhaps to highlight certain features of the original retinal input, or to suppress others. We have here the elements of some real-time *plasticity* in the brain's processing activities, such as the capacity for directing attention and focusing relevant resources. Descending control pathways (as opposed to upward, always upward) are especially prominent in the visual system and in the auditory system, which must process speech, but they are common throughout the brain.

Between the sensory areas of the cortex here discussed, and other sensory areas similarly connected to the other organs of one's sensory periphery, there remains a great deal of highly active brain. The large so-called association areas, between the various types of sensory cortex, are not well understood, and neither are the large frontal areas of the cerebral hemispheres, though it is plain from cases of brain damage that these frontal areas are implicated in emotion, drive, and the capacity for planned action.

There is a hypothesis that makes rough sense of these areas, of their cognitive functions and of their specific axonal connections with other areas. Consider figure 7.11. The cross-hatched areas are the areas of *primary* sensory cortex: somatosensory, visual, and auditory. The three vertically striped areas next to each are called *secondary* sensory cortex. Neurons in the primary cortex project their axons to neurons in the secondary cortex, for all three sensory modalities, and these secondary neurons are responsive to more complex and abstract features of the original sensory input than are the neurons in the primary cortex. Secondary cortex projects its axons, in turn, into the

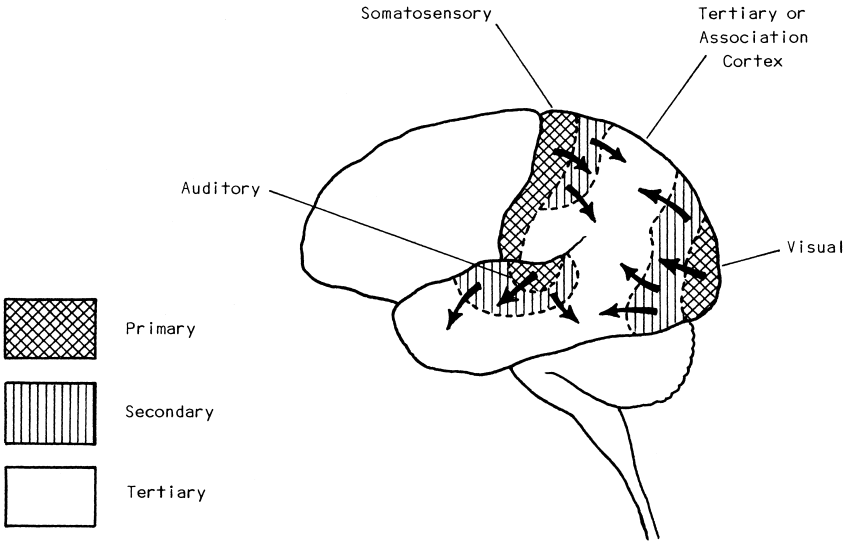

Figure 7.11

unshaded areas, called *tertiary* or *association* cortex. Neurons in the association cortex are responsive to still more abstract features of the original sensory input, but here we find a mixture of cells, some responsive to visual input, some to auditory input, some to tactile input, and some to combinations of all three. It would appear that the brain's most abstract and integrated analysis of the sensory environment takes place in the association cortex between the several sensory areas.

From this rear or 'sensory' half of the brain, several axonal superhighways project to the frontal or 'motor' half of the brain, into what we may call the tertiary motor areas. This is the unshaded frontal area in figure 7.12. This area appears to be responsible for the formation of our most general plans and intentions. Neurons here project into the secondary motor

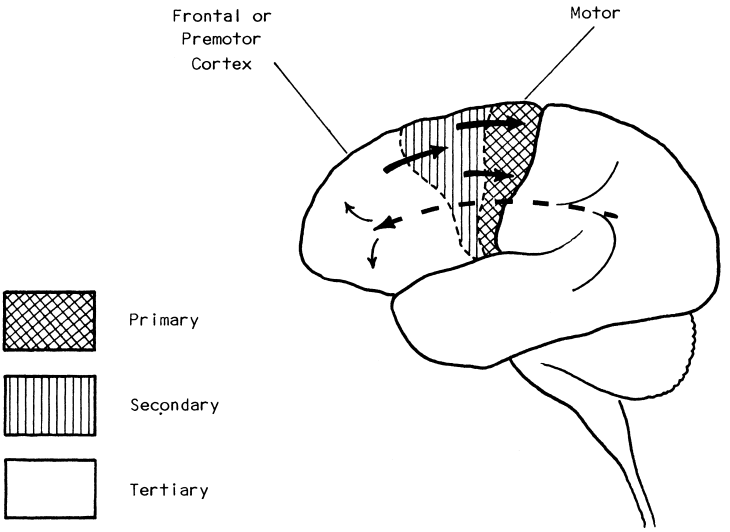

Figure 7.12

cortex, which appears to be the locus of more specifically conceived plans of action. This area projects finally to the primary motor cortex, which is responsible for the highly specific physical motions of the various parts of the body.

This general hypothesis is consistent with the neuroarchitecture of the brain, with its overall capacities as a sensorily guided controller of bodily behavior, and with detailed studies of the highly specific cognitive deficits produced by isolated lesions at various sites within the brain. Damage to the extreme frontal lobe, for example, leaves the victim unable to conceive of, or to distinguish in a caring fashion between, alternative possible futures beyond the most immediate and simple practical matters.

The preceding sketch of the global organization of the brain represents the classical view, but the reader should be warned

that it presents a provisional and oversimplified picture. Recent studies indicate that distinct topographic maps of the retina, for example, are scattered throughout the cortical surface, and enjoy distinct projections from the LGN, or from elsewhere in the centrally located thalamus. The hierarchical system of topographic maps discussed earlier, which culminates in the 'secondary visual cortex' toward the rear of the brain, is thus only one of several parallel systems, each processing different aspects of visual input. The 'classical' system for vision may be the dominant one, but it has company, and all of these systems interact. Similar complexities attend the 'somatosensory cortex', which emerges as only one of several parallel systems processing different types of somatosensory information: light touch, deep pressure, limb positions, pain, temperature, and so forth. Sorting out the functional differences between these distinct maps and tracing their functional interconnections is a job that has only begun. As that information continues to emerge, our appreciation of the intricate and occasionally unsuspected achievements of our perceptual systems must grow in equal measure.

One further area of intrigue is worthy of mention, not because it is large, but because it is the ultimate target of a hierarchy of axonal projections from many and varied areas of the cerebral cortex. The smallish *hippocampus* is located at the back end of the limbic system, a forebrain structure just under the great cerebral hemispheres. If we trace the inputs to the hippocampus back to their origins, against the flow of incoming information, we fairly quickly implicate the entire cerebral cortex. Damage to the hippocampus, it emerges, blocks the transfer of information from short-term into long-term or permanent memory. Victims of such damage live in a nightmare world of *no* memories reaching longer than a minute or so into the past, save only

for their original long-term memories, of those ever more distant events, entrenched before the damage to the hippocampus occurred.

It is natural to think of the brain as something which is interposed between the peripheral sensory nerves and the peripheral motor nerves, something controlled by the former and in control of the latter. From an evolutionary perspective, this makes sense, at least in the early stages. But with the brain at the level of articulation and self-modulation found in mammals and most especially in humans, a certain *autonomy* has crept into the picture. Our behavior is governed as much by our past learning, and by our current plans for the future, as by our current perceptions. And through self-directed learning, the long-term development of the brain's internal organization is to a substantial extent under the control of the brain itself. We do not, by this means, *escape* the animal kingdom, but we have become its most creative and unpredictable members.

Suggested Readings

Hubel, D. H., and T. N. Wiesel. "Brain Mechanisms of Vision." *Scientific American* 241 (3) (September 1979): a special issue devoted to the various brain sciences.

Bullock, T. H., R. Orkland, and A. Grinnell. *Introduction to Nervous Systems*. San Francisco: Freeman, 1977.

Kandel, E. R., and J. H. Schwartz. *Principles of Neural Science*. New York: Elsevier/North-Holland, 1981.

Shepherd, G. M. *Neurobiology*. New York: Oxford University Press, 1983.

Sherman, S. M. "Thalamic Relays and Cortical Functioning." *Progress in Brain Research* 149 (2005): 107–126.

3 Neuropsychology

Neuropsychology is the discipline that attempts to understand and explain psychological phenomena in terms of the neurochemical, neurophysiological, and neurofunctional activities of the brain. We have already seen some tentative but intriguing neuropsychological results in the preceding section: how the hierarchical structure of the visual system permits us to discriminate selected features from a scene, how interleaved retinal representations on the cortical surface make stereo vision possible, and how the overall organization of the cerebral cortex makes it possible for highly processed sensory information to guide the formulation and execution of general plans of action.

Unfortunately, the greater portion of the data traditionally available to neuropsychology derives from cases of brain damage, brain degeneration, and chemical disequilibrium. What we understand best is the neural basis of *abnormal* psychology. Brain tissue can be physically disrupted by invasive objects; it can be crushed by growing tumors or fluid pressure; it can starve and atrophy from localized loss of blood supply; or it can be selectively destroyed by disease or degeneration. Depending on the specific *location*, within the brain, of the lesion produced by any of these means, very specific losses in the victim's psychological capacities typically result.

Such losses may be minor, as with an inability to verbally identify perceived colors (lesions to the connections between the secondary visual cortex and the secondary auditory cortex of the left hemisphere). Or they may be more serious, as with the permanent inability to recognize individual faces, even those of family members (lesions in the association cortex of the right hemisphere). And they can be devastating, as with the

total and permanent loss of speech comprehension (lesions to the secondary auditory cortex of the left hemisphere), or the inability to lay down long-term memories (bilateral damage to the hippocampus).

Using postmortem examinations, and other diagnostic techniques such as the various types of modern brain scans, neurologists and neuropsychologists can find the neural correlates of these and hundreds of other losses in cognitive and behavioral function. By such means we can slowly piece together a *functional map* of the brain. We can come to appreciate the functional specializations and the functional organization of the brain in a *normal* human. This information, in conjunction with a detailed understanding of the neuroarchitecture and microactivity of the relevant areas, can lead to a real understanding of how our cognitive activities are actually produced. Recall our earlier glimpse into feature extraction and stereopsis in the visual system. Once we know where to look for them, we can start to find specific neural structures that explain the specific features of the cognitive capacity at issue. Overall, there is cause for much optimism here, even though our ignorance still dwarfs our understanding.

The functional sleuthing just described requires caution in two respects. First, the simple correlation of a lesion in area x with loss of some cognitive function F need not mean that area x has the function F. It means only that some part of area x is typically involved in some way in the execution of F. The key neural structures that sustain F may be located elsewhere, or they may not be localized at all, being distributed over large areas of the brain.

Second, we must not expect that the functional losses and functional localizations that we do find will always correspond

neatly with cognitive functions represented in our common-sense psychological vocabulary. Sometimes the deficit is difficult to describe, as when it involves a global change in the victim's personality, and sometimes its description is difficult to credit. For example, some lesions produce a complete loss of awareness, both perceptual and practical, of the *left half* of the victim's universe, including the victim's own body. (This condition is called *hemi-neglect*.) A victim will typically dress only the right side of his body, and even deny ownership of his left arm and leg. Other lesions leave the victim able to write lucid, readable prose, but *un*able to read and understand what she or anyone else has written, even though her vision is wholly normal. (This condition is called *alexia without agraphia*.) Lesions different yet again leave the victim wholly 'blind', in the sense that his visual field has disappeared and he insists that he cannot see anything at all; and yet he can 'guess' the direction where a small light has been placed somewhere in front of him with an accuracy approaching 100 percent. (This condition is called *blind-sight*.) Still other lesions, to the entire primary visual cortex, for example, leave the victim genuinely and utterly blind, but, for a time after the onset of that condition, the victim perversely insists that she *can* see perfectly well, as she stumbles around the room confabulating lame excuses for her clumsy behavior. (This condition is called *blindness denial*.)

These cases are surprising and confusing, relative to the default conceptions of folk psychology. How could one possibly be blind and not know it? See with no visual field? Write freely but not read a word? Or sincerely deny ownership of arms and legs obviously attached to oneself? These cases violate entrenched expectations. But we cannot expect that our current

folk psychology represents anything more than one stage in the historical development of our self-understanding, a stage the neurosciences may help us to transcend.

Beneath the level of structural damage to our neural machinery, there is the level of chemical activity and chemical abnormalities. The reader will recall that transmission across each tiny synaptic junction itself is a critical element in all neural activity, and that such transmission is chemical in nature. Upon receipt of an axonal impulse or 'spike', the axon's end-bulb releases a chemical called a *neurotransmitter* that swiftly diffuses across the synaptic cleft so as to bind with the chemical receptors waiting on the far side. This binding leads to the breakdown of the neurotransmitter chemical, and the breakdown products are subsequently taken up again by the end-bulb for resynthesis and reuse.

Evidently, anything that frustrates, or exaggerates, these subtle chemical activities will have a profound effect on neural communication and on collective neural activity. This is precisely how the many psychoactive drugs work their effects. The various types of neurons make use of distinct neurotransmitters, and different drugs have different effects on their activities, so there is room here for a wide variety of effects, both chemical and psychological. A drug may block the synthesis of a specific neurotransmitter; or bind to its receptor sites, thus blocking its normal effects; or it may block the reuptake of its breakdown products, thus preventing its resynthesis. On the other hand, a drug may enhance synthesis, increase receptor sites, or accelerate the reuptake of breakdown products. Alcohol, for example, is an antagonist to the action of *noradrenaline*, an important neurotransmitter, whereas the amphetamines positively enhance its activity, producing the very opposite psychological effect.

Most important, extreme doses of certain of the psychoactive drugs produce symptoms that closely resemble those of the major forms of mental illness—depression, mania, and schizophrenia. This suggests the hypothesis that these illnesses, as they occur naturally, involve the same or closely similar chemical abnormalities as are artificially produced by these various drugs. Such hypotheses are of more than purely theoretical interest because if they are true, then the naturally occurring illness may well be correctable or controllable by a drug with an exactly opposite neurochemical effect. And thus it seems to be, though the situation is complex and the details are confusing. *Fluoxetine* controls chronic depression, *lithium salts* control mania, and *chlorpromazine* controls schizophrenia. Imperfectly, it must be said, but the qualified success of these drugs lends strong support to the idea that the victims of mental illness are the victims primarily of sheer chemical circumstance, whose origins are more metabolic and biological than they are social or psychological. If so, this fact is important, since better than 2 percent of the human population has a significant encounter with one of these three conditions at some point in their lives. If we can discover the nature and origins of the complex chemical imbalances that underlie the major forms of mental illness, we may be able to cure them outright or even prevent their occurrence entirely.

Suggested Readings

Kolb, B., and I. Q. Wishaw. *Fundamentals of Human Neuropsychology*. San Francisco: Freeman, 1980.

Hecaen, H., and M. L. Albert. *Human Neuropsychology*. New York: Wiley, 1978.

Gardner, H. *The Shattered Mind*. New York: Knopf, 1975.

4 Cognitive Neurobiology

As its name implies, cognitive neurobiology is an interdisciplinary area of research whose concern is to understand the specifically cognitive activities displayed by living creatures. It has begun to flower in recent years, for three reasons.

First, there has been a steady improvement in the *technologies* that allow us to explore the microstructure of the brain and to monitor our ongoing neural activities. Modern electron microscopes give us an unparalleled access to the details of brain microstructure, and various nuclear technologies allow us to image the internal structure and neural activity of living brains without invading them or disrupting them at all. Second, research has benefited from the appearance of some provocative general *theories* about the function of large-scale neural networks. These theories give a direction and a purpose to our experimental efforts; they help tell us what are the useful questions to ask of Nature. And third, modern *computers* have made it possible for us to explore, in an efficient and revealing way, the functional properties of the highly intricate structures that recent theories ascribe to our brains. For we can model those structures within a computer and then let the computer display how they will behave under various circumstances. We can then test these computer-generated behavioral predictions against the behavior of real brains in comparable circumstances.

In this section we will take a brief look at two of the central questions of cognitive neurobiology. How does the brain *represent* the world? And how does the brain perform *computations* over those representations? Let us take the first question first, and let us begin with some phenomena entirely familiar to you.

How does the brain represent the color of a sunset? The smell of a rose? The taste of a peach? Or the face of a loved one? There is a simple technique for representing, or *coding*, such external features that is surprisingly effective, and can be used in all of the cases mentioned, despite their diversity. To see how it works, consider the case of taste.

Sensory Coding: Taste

On one's tongue, there are four distinct kinds of chemically sensitive receptor cells. (There are recent indications of a fifth type, but for simplicity's sake I'll leave this aside.) Cells of each kind respond in their own peculiar way to any given substance that makes contact with them. A peach, for example, might have a substantial effect on one of the four kinds of receptor cell, a minimal effect on the second kind, and some intermediate level of effect on the third and fourth kinds. Taken altogether, this exact *pattern* of relative stimulations constitutes a sort of neural 'fingerprint' that is uniquely characteristic of peaches.

If we name the four kinds of cells *a*, *b*, *c*, and *d*, respectively, then we can describe exactly what that special fingerprint is, by specifying the four levels of neural stimulation that contact with a peach actually produces. If we use the letter S, with a suitable subscript, to represent each of the four levels of stimulation, then the following is what we want: $<S_a, S_b, S_c, S_d>$. This literal *list* of excitation levels is called a *sensory coding vector* (a vector is just an ordered list of numbers, or magnitudes). The important point is that there is evidently a *unique* coding vector for every humanly possible taste. Which is to say, any humanly possible taste sensation is just a pattern of stimulation levels

across the four neural channels that convey news of these activity levels away from the mouth and to the rest of the brain.

We can graphically display any given taste by means of an appropriate point in a 'taste-space', a space with four axes, one each for the stimulation level in each of the four kinds of sensory taste cell. Figure 7.13 depicts a space in which the positions of the various tastes are located. (However, in this diagram, one of the four axes has been suppressed, since it is hard to draw a 4D space on a 2D page.) What is interesting immediately is that subjectively similar taste-sensations turn out to have very

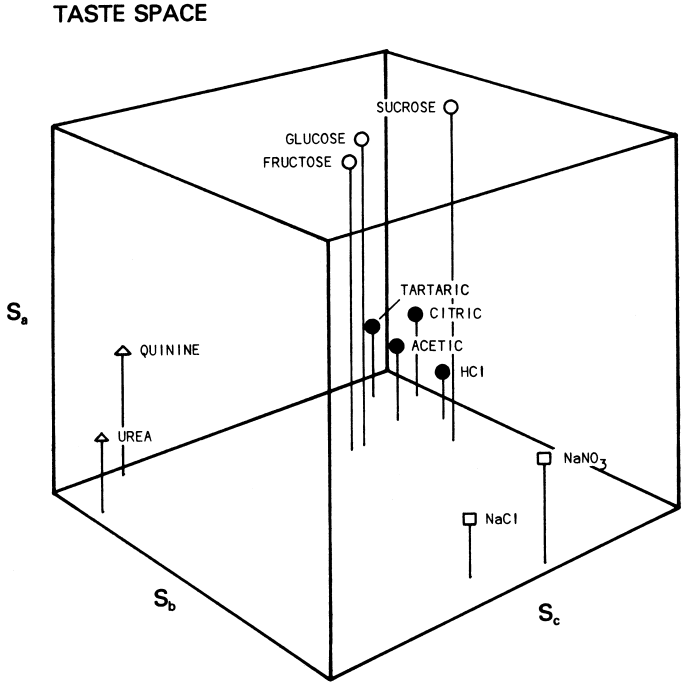

Figure 7.13

similar coding vectors. Or what is the same thing, their proprietary points in taste-space are very *close together*. You will notice that the various types of 'sweet' tastes all get coded in the upper regions of the space, while sundry 'tart' tastes appear in the lower center. Various 'bitter' tastes appear close to the origin of the space (the 'bitter' axis is the one we dropped), and 'salty' tastes reside in the region to the lower right. The other points in this space represent all of the other taste sensations it is possible for humans to have. Here there is definite encouragement for the identity theorist's suggestion (chapter 2.3) that any given sensation is simply identical with a set or pattern of spiking frequencies in the appropriate sensory brain area.

Sensory Coding: Color

A somewhat similar story appears to hold for color. There are three distinct types of color-coding neurons distributed uniformly throughout cortical area V4, just downstream from the primary visual cortex. These three types of cells are ultimately driven by the wavelength-sensitive cells in the retina, via a clever tug-of-war arrangement involving the axons between the two cell populations. (I'll spare you the details.) Here also, a (*three*-dimensional) neuronal activation space, embedded in area V4, displays simultaneous activation-levels across those three types of cells for each small area of the visual field, an activation space for each of the possible colors perceivable by humans. Figure 7.14 portrays that space, and you will notice that it contains a special double-coned or spindle-shaped *subvolume*, within which *all* of the familiar objective colors are systematically placed according to their unique *similarity* (i.e., proximity) and *dissimilarity* (i.e., distance) relations to all of the other objective colors. Orange, for example, is tucked closely

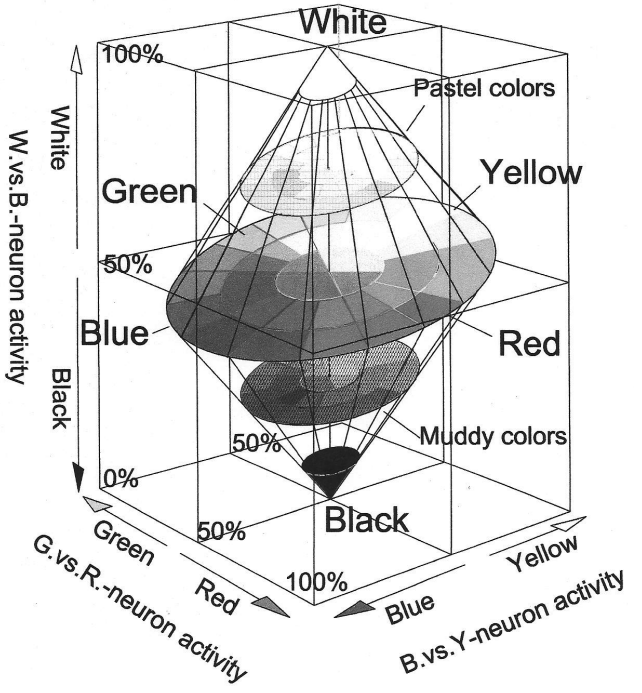

Figure 7.14

between red and yellow, as you would expect, while green is a maximal distance from red, as is blue from yellow, black from white, and so forth. This neuronal coding system recreates, in complete detail, the internal qualitative structure of human phenomenological color space, as displayed in introspection. One might even say that it *explains* it, especially since it predicts, with equal accuracy, the qualitative character of the many thousands of possible *after images* one can induce in the human visual system by temporarily fatiguing the neurons involved. Indeed, it even predicts the weird qualitative characters of

certain unusual visual activation-vectors *outside* the central spindle of the familiar objective colors. That is, it correctly predicts the qualitative characters of sensations you have never even had before. Evidently, phenomenological qualia are not quite so inaccessible to physical theory as was originally advertised. (For an accessible summary of these results, with color diagrams to help produce the relevant after-images, see the article by P. M. Churchland [2005], in the suggested readings at the end of this section.)

Sensory Coding: Smell

The olfactory system appears to involve at least six or seven, and perhaps many more, distinct kinds of receptors. This suggests that smells are coded by a vector of activation levels or spiking frequencies with at least six or seven different elements. This allows for a great many distinct combinations of individual spiking frequencies, and hence for a great many distinct smells. Let us suppose that a bloodhound, for example, has seven different kinds of olfactory receptors and can distinguish between thirty different levels of stimulation within each type (rather more than we can). On these assumptions, we must credit the bloodhound with an overall 'smell space' of 30^7 or 22 *billion* discriminable positions! No wonder dogs can distinguish any one person from among millions, by smell alone.

All of this—the story on taste, color, and smell—must provide encouragement for the identity theorists, who claim that our sensations are simply identical with a signature set of stimulation levels (spiking frequencies) in the appropriate sensory pathways. For as the preceding material shows, neuroscience is successfully reconstructing, in a systematic and revealing way, the various features of, and the relations between, our subjective

sensory qualia. This is the same pattern that, during the nineteenth century, motivated the scientific claim that light is simply identical with electromagnetic waves of suitable frequencies. For within the emerging theory of electricity and magnetism, we could systematically reconstruct all of the familiar features of light. And also some unfamiliar ones, such as the existence of infrared and ultraviolet light, outside of the range of normal human vision.

Sensory Coding: Faces

Among humans, it is *faces* that get distinguished with great skill, and a recent theory say that faces are also handled by a vector-coding strategy. For each of the various elements of a human face to which we are sensitive—nose length, width of mouth, distance between eyes, squareness of jaw, etc.—suppose there is a devoted neural pathway whose level of stimulation corresponds to the degree to which the perceived face displays that particular element. A particular face, therefore, will be coded by a unique vector of stimulations, a vector whose elements correspond to the visible elements of that face.

If we guess (because we do not know) that there are perhaps ten different facial features to which a mature human is selectively sensitive, and if we suppose that we can distinguish at least five different levels within each feature, then we must credit humans with a 'facial space' of at least 5^{10} (about 10 million) discriminable positions. Once again, it is small wonder we humans can distinguish any face, from among millions, by sight alone.

The faces of people who are close relatives, of course, will be coded by vectors with many of the same or similar vector elements. By contrast, people bearing no facial resemblance to each

other will be coded by quite disparate vectors. As well, a person with a supremely *average* face will be coded by a vector where all of its activational elements are in the middle of the relevant range of variation. And someone with a highly *distinctive* face will be coded by a vector that has several elements at an extreme value. Interestingly, the human brain boasts a smallish area downstream from the visual cortex, called the *fusiform gyrus*, whose injury or destruction produces in the victim an inability to recognize or discriminate between human faces. Here, we may postulate, are human faces coded.

An *artificial* neural network (modeled in a conventional computer), with three layers of vector coding connected by a suitable spray of intervening axonal connections, has already been constructed and taught to discriminate among a dozen distinct human faces. That is, after repeated exposure to these various faces and progressive adjustments of its thousands of synaptic connections, it will accurately reidentify the same individual across distinct photographs of that same individual. Its first neuronal layer corresponds to the retina, of course, and its third layer corresponds to our fusiform gyrus. But like our own retina, its first or sensory layer contains no special cells that are automatically and devotedly sensitive to elemental features of human *faces*, as we found with sensory cells for taste, color, and smell. This network had to *learn* which abstract features of faces would allow it to code the faces on which it was trained so as to discriminate them reliably. Withal, it did learn and it fell into the familiar practice of coding its perceived faces with signature activation vectors across its middle population of neurons. Its final layer learned to respond to those signature activation vectors with a code for the name, and even the gender, of the specific person recognized.

Sensory Coding: The Motor System

The virtues of vector coding are especially apparent when we consider the problem of representing a very complex system, such as simultaneous position of the thousands of muscles in one's body. You have a constant and continuously updated sense of the overall posture or configuration of your body in space. And a good thing, too. To be able to effect any useful movements at all, you must know where your limbs are starting from. This goes for simple things like walking, just as much as for complex things like ballet or basketball.

This sense of one's bodily configuration is called *proprioception*, and it is possible because each and every muscle in the body has its own nerve fiber constantly sending information back to the brain, information about the contraction or extension of that muscle. With so many muscles, the total bodily coding vector reaching the brain will plainly have not three elements, or ten, but something over a thousand elements. But that is no problem for the brain: it has billions of fibers with which to do the job.

Output Coding

While we are talking about the motor system, you might notice that vector coding can be just as useful for directing *motor output* as it is for coding sensory input. When a person is engaged in any physical activity at all, the brain is sending a cascade of distinct messages toward every muscle in the body. But those messages must be collectively well organized if the body is to do anything coherent: every muscle must assume just the right level of contraction or extension if they are to make the body assume the position intended.

Neural Computing

As we have seen, stimulation vectors are a beautifully effective means of representing things as various as tastes, colors, faces, and complex limb positions. Equally important, it turns out that they are part of a very elegant solution to the problem of high-speed *computing*. If the brain uses vectors to code various sensory inputs, and also various motor outputs, then it must somewhere be performing computations so that those inputs are in some way *guiding* or *producing* those motor outputs. In short, it needs some systematic arrangement to transform its various sensory input vectors into appropriate motor output vectors.

As it happens, large segments of the brain have a microstructure that is ideally suited to performing transformations of precisely this kind. Consider, for example, the schematic arrangement of axons, dendrites, and synaptic connections portrayed in figure 7.15. Here the input vector, <*a*, *b*, *c*, *d*>, is conveyed along the four horizontal input axons, one axon for each of the four letters. That is, at any moment, each axon is conducting an incoming train of spikes with a certain frequency. And as you can see, each axon makes a total of three synaptic connections, one for each of the three vertical cells. Altogether, that makes $4 \times 3 = 12$ synapses.

The receiving cell emits a train of spikes down its own output axon, a train whose frequency is a function of the total excitation that the various inputs have collectively produced in that cell. Since all three of the receiving cells do this, their collective output is obviously another vector, a vector with three elements. Plainly, our little neural network will transform any incoming 4D vector into an outgoing (and quite different) 3D vector.

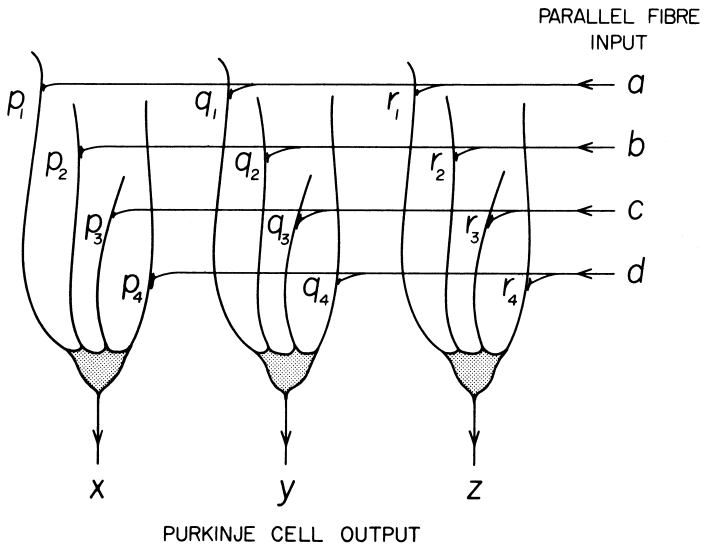

Figure 7.15

What determines the nature of the overall computation or transformation is of course the distribution of *sizes* or *weights* among the various synaptic connections. When we specify the distribution of synaptic weights in a system of this sort, we have specified the character of the transformation it will perform on any incoming activation vector. We have specified the *computation* that the network will perform.

A Real Example: The Cerebellum

The vector-transforming system of figure 7.15 is just a schematic sketch, highly simplified for purposes of illustration. But the same connection configuration is repeated, with local variations, again and again throughout the brain, especially within its widespread *gray matter*, the subvolumes of the brain where

the neuronal bodies and their many dendrites are heavily concentrated. An especially impressive instance of the relevant pattern of connections is found in the *cerebellum* of all mammals. Figure 7.16 depicts a tiny section of the cerebellar cortex, and you can see how the many Mossy-fiber input axons conduct their spiking frequencies, first, into the tiny Granule cells in the shaded area, and second, out to the many *parallel fibers*, each one of which makes *multiple* synaptic connections with many of the unusually bushy *Purkinje cells* awaiting their incoming messages. Each Purkinje cell then sums the activities thus induced in it, and emits an appropriate train of spikes down its

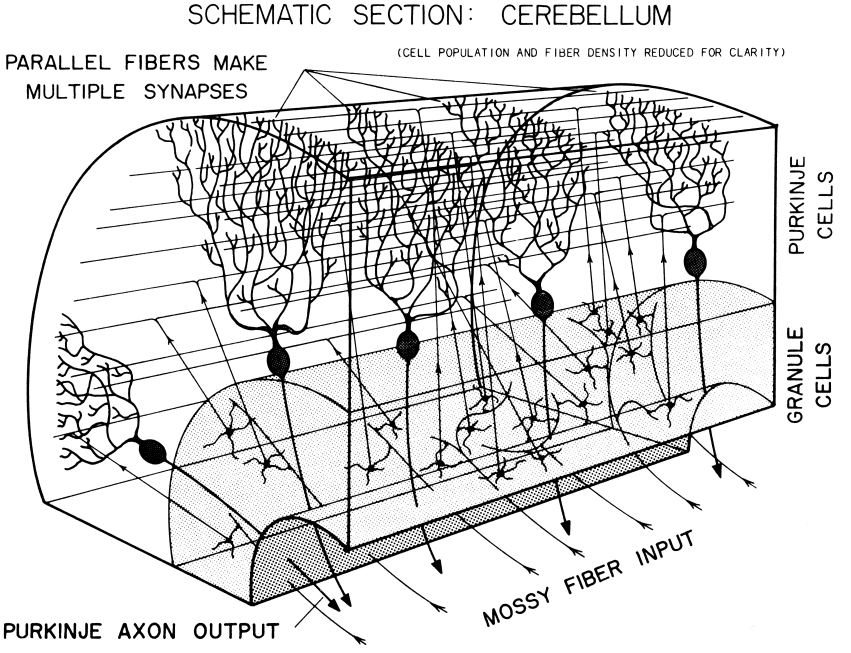

Figure 7.16

own axon as output. The assembled activity levels in the entire set of Purkinje axons constitute the cerebellum's output vector.

What is impressive about this particular example is the unusually bushy nature of the Purkinje cells' dendritic trees, the large number of such cells, and the massive number of parallel fibers that make multiple synaptic contacts with them. The 'connection matrix' portrayed in figure 7.14 had only *twelve* transforming synaptic connections, but the connection matrix of a real neural population, like this one, would have many *hundreds of thousands*, perhaps even *millions* of such connections. Its computational potential would thus be vastly greater than the cartoon network of figure 7.14.

We must remember also that the output vector of such a gargantuan processor can be conveyed by its output axons to make systematic synaptic contact with the waiting dendrites of a *second* population of distinct neurons elsewhere in the brain, which synaptic connections collectively form a *further* computational matrix, one with its own transformational/computational concerns. This downstream matrix may lead in turn to a third neuronal population, and that to a fourth, and so on. This is in fact how the brain is wired together, overall, and one can now begin to appreciate the incredible computational power of such an iterated and massively populated system. This is what makes our brain the most complex physical system for many light-years in every direction.

There are three more important features to notice about a 'computing' system of the general kind here displayed. First, it is highly resistant to minor damage and scattered cell death. Since it is made up of many hundreds of thousands of synaptic connections, or even more, each one of which contributes only a tiny amount to the overall transformation of the incoming

vectors, the loss or corruption of a few hundred connections here and there will change the network's global behavior hardly at all. It can even lose many thousands of connections entirely, so long as they are scattered randomly throughout the network, as happens with the gradual death of neurons in the natural course of aging. The quality of the network's computations will therefore slowly and gracefully *degrade*, rather than suddenly collapse. This welcome feature is called *functional persistence*. The CPU of your desktop computer does not have this feature, but you do.

Second, and just as important, a massively parallel computing system of this kind will perform any given vector-to-vector transformation, no matter how many elements are involved, in a figurative *instant*. Because each synapse performs its individual 'calculation' more or less *simultaneously* with every other synapse, the overall matrix of connections performs the relevant transformation *all at once*, rather than in laborious sequence, as with a conventional serial/digital computer. The time taken to perform any given global transformation is therefore *independent* of the size or complexity of the transformation involved. This is the principal reason why the biological brain can outperform a high-speed electronic computer on so many typical cognitive tasks: those tasks regularly involve the transformation of high-dimensional vectors by very large synaptic matrices. Evolution stumbled upon a winner when it stumbled upon what has come to be called *parallel distributed processing*.

Third, and perhaps most important of all, such networks are functionally modifiable. In technical parlance, they are *plastic*. They can change their transformational or computational capacities simply by changing some or all of their constituent synaptic weights. This is important since it must be possible for

the system to *learn* to perform the required transformations in the first place. We are not born with our adult perceptual, conceptual, and practical skills. The brain acquires them only slowly, and in stages. Learning is a complex process we are only beginning to penetrate, but, for all creatures and at the most basic level, learning appears to consist in the gradual *adjustment* of the myriad *weights* of the synaptic connections that make up the vector-transforming matrices here under discussion. This gradual 'tuning' of our diverse synaptic matrices is driven by our unfolding *experience* of the world at large. It is driven by the spatial and temporal *structures* that the objective world repeatedly displays to us as we navigate its many challenges. By way of these synapse-adjusting procedures, those objective structures end up being *mapped* in the space of possible activation-vectors across the brain's neuronal populations. *Similar* external structures produce activation-vectors that are *close* to each other in the space of possible activation vectors. And *dissimilar* external structures produce activation vectors that are *far apart* in the space of possible activation vectors. In this way does the brain get a lasting grip on the external reality that embeds it.

In summary, neural networks of the massively parallel sort at issue are computationally powerful, damage-resistant, fast, and modifiable so as to represent the structure of the world. Nor do their virtues end here, as we are about to see in the next section.

Suggested Readings

Llinás, R. "The Cortex of the Cerebellum." *Scientific American* 232, no. 1 (1975).

Bartoshuk, L. M. "Gustatory System." In *Handbook of Behavioral Neurobiology*, vol. 1: *Sensory Integration*, ed. R. B. Masterton. New York: Plenum, 1978.

Pfaff, D. W. *Taste, Olfaction, and the Central Nervous System*. New York: Rockefeller University Press, 1985.

Hardin, C. *Color for Philosophers: Unweaving the Rainbow*. Indianapolis: Hackett, 1988.

Churchland, P. M. "Chimerical Colors: Some Phenomenological Predictions from Cognitive Neuroscience." *Philosophical Psychology* 18, no. 5 (2005). Reprinted in P. M. Churchland, *Neurophilosophy at Work* (Cambridge: Cambridge University Press, 2007).

Churchland, P. S. *Neurophilosophy*. Cambridge, MA: MIT Press, 1986.

5 AI Again: Computer Models of Parallel Distributed Processing

In the late 1950s, very early in the history of AI, there was considerable interest in artificial 'neural networks', that is, in hardware systems physically modeled on the biological brain. Despite their initial appeal, these first-generation networks were shown to have serious practical limitations, and they were quickly eclipsed by the techniques of 'program-writing' AI. These latter, though also successful at first, have since proved to have severe limitations of their own, as we saw at the end of chapter 6, and recent years have seen a rebirth of interest in the earlier approach. The early limitations have been transcended, and artificial neural networks are finally beginning to display their real potential.

Artificial Neural Networks: Their Structure

Consider a network composed of simple neuronlike units, connected in the fashion displayed in figure 7.17. The bottom-most units may be thought of as sensory neurons, as they are directly stimulated by the environment outside the system. Each of these bottom units emits an output signal along its own 'axon',

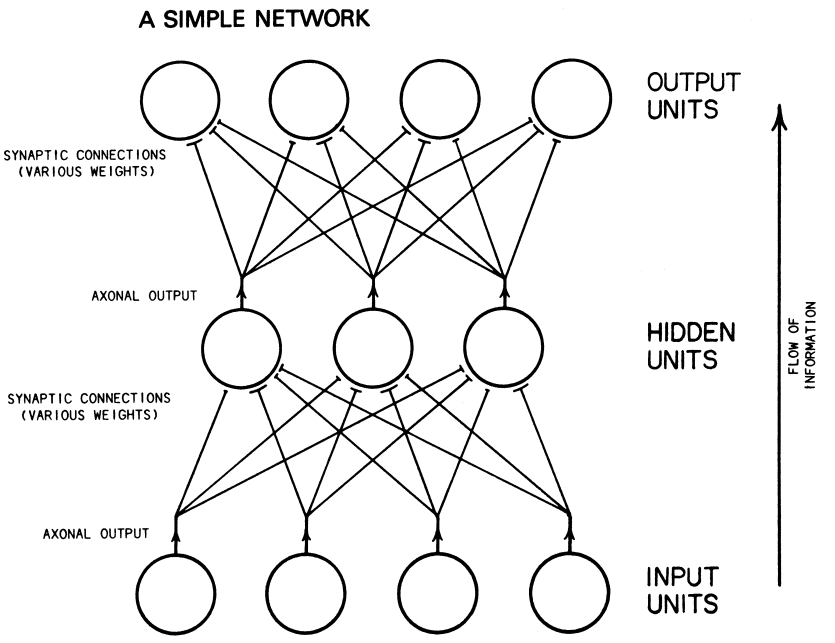

Figure 7.17

an output signal whose strength is a function of the sensory unit's level of perceptual stimulation. That axon divides into a number of 'terminal end branches', and a copy of its output signal is thus conveyed to each and every neuronal unit at the second or middle level. These middle units are often called the *hidden units* (because they are 'hidden away' between the top and bottom layers), and the reaching axonal end-branches make a variety of 'synaptic connections' with each of them. Each connection has its own strength or *weight*, as you might expect.

You can see already that the bottom one-half of this system is just another vector-to-vector transformer, much like the

neural matrices discussed in the previous section. (The only difference is that here it is the arriving *axons* that do the necessary branching, rather than a receiving tree of *dendrites*, as before. But this is just a diagrammatic convenience, not a real functional novelty.) If we stimulate the bottom units, the overall pattern of activity-levels we induce therein (i.e., the input vector) will be conveyed upward toward the hidden units. As it arrives, it gets transformed by the intervening matrix of synaptic connections, and by the summing activity within each of the hidden units. The result is a set or pattern of activation-levels now across the hidden units: another activation vector, although this time the vector has only three elements instead of the original four.

This three-element vector serves in turn as an input vector to the top half of the overall system. The axons reaching up from the hidden units make a spray of synaptic connections, of various weights, to the final units at the topmost level. These are the so-called output units, and the overall set of activation-levels finally induced in them is what constitutes the 'output' vector for the entire network. The upper half of that network is therefore just another vector-to-vector transformer, just like the bottom half, save that it converts a three-element vector into a vector with four elements.

Following this general pattern of connectivity, we can clearly construct a network with any desired number of input units, hidden units, and output units, depending on the size of the vectors that need processing. And we can begin to see the point of having a *two*-tiered arrangement if we consider what such an iterated network can do when confronted with a real-life problem. The crucial point to remember is that we can progressively *modify* the synaptic weights within the overall system, so

as to implement any input-vector to output-vector transformation that we might desire.

Perceptual Recognition: Learning from Examples

Our sample real-life problem is as follows. We are the command crew of a submarine, whose mission will take us into the shallow waters of an enemy harbor, a harbor whose bottom is sprinkled with explosive mines, mines equipped with metal detectors primed to set off the mine whenever a large metal object, such as our submarine, gets to within a certain distance of it. We need to give these mines a wide berth, and we can at least detect them from a safe distance with our sonar system, which sends out a brief pulse of sound and then listens for a returning echo in case the pulse bounces off some solid object lying on the harbor bottom. (How long the return takes tells us how far away the object is.) Unfortunately, a sizeable *rock* will also return a sonar echo, an echo that is indistinguishable, to the casual ear, from a genuine mine echo (figure 7.18).

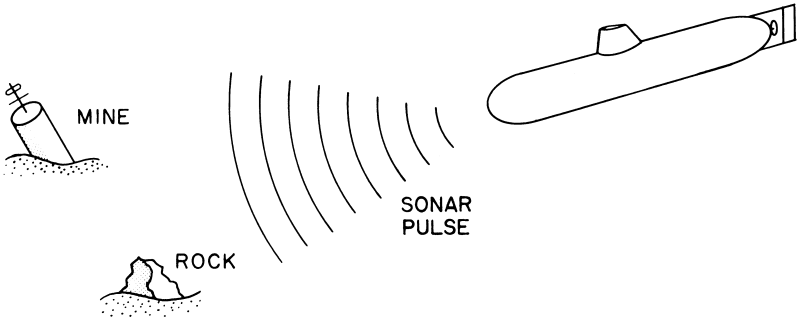

Figure 7.18

This is frustrating, because the harbor bottom is also well sprinkled with largish *rocks*. The situation is further complicated by the fact that the mines come in various shapes and lie in various orientations relative to the arriving sonar pulse. And so do the rocks. So the echoes returning from each type of object also display considerable variation within each class. On the face of it, our situation looks hopelessly confused.

How might we prepare ourselves to distinguish the explosive-mine echoes from the benign-rock echoes, so that we may undertake our harbor intrusion in confidence? As follows. We first assemble, on recording tape and while still in our home port area, a large set of sonar echoes from what we already *know* to be genuine mines of various types and in various positions, mines we have ourselves placed on the ocean bottom so as to examine their sonar-reflecting properties. We do the same for rocks of various kinds, and of course we keep careful track of which echoes are which. We end up with, say, fifty samples of each.

We then put each recorded echo through a 'frequency analyzer', a simple device which yields up information of the sort displayed at the extreme left of figure 7.19. This just shows *how much* sound energy the echo embodies at *each* of the many sound frequencies that make it up. It is a way of getting a 'signature profile' of any given echo. All by itself, this analysis doesn't help us much, since the collected profiles still don't seem to display any obvious uniformities or regular differences among the one hundred echoes we have carefully recorded.

But now let us bring a neural network into the picture. (See the rightmost part of figure 7.19. This is a simplified version of a network originally explored by Gorman and Sejnowski. Note

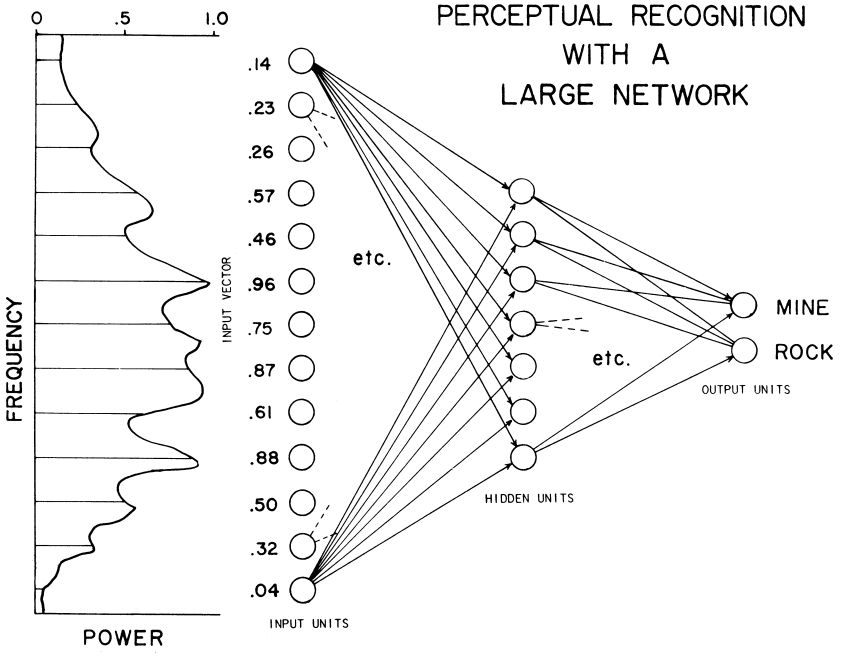

Figure 7.19

that it has been tilted on its side, relative to figure 7.17, again for purely diagrammatic reasons.) This network is organized along the same lines as the simple network of figure 7.17, but it has fully 13 input units, 7 hidden units, 2 output units, and a total of 105 synaptic connections. The activity levels of each unit vary between zero and one. Remember also that the synaptic weights of this system can be adjusted to whatever values may be needed to solve our diagnostic problem. But of course we do not *know* which values are needed! So, at the beginning of this experiment, each connection is given a randomly small

weight, either excitatory or inhibitory, fairly close to zero. The specific transformation that the network performs for us in this condition is therefore unlikely to solve our original problem. But we proceed as follows.

We take a mine echo from our store of samples, and we use the frequency analyzer to sample its energy levels at the thirteen specific frequencies. This gives us the input vector, which has 13 elements. We then enter this vector into the student network by stimulating each of its 13 input units by an appropriate amount, as indicated in figure 7.19. The vector is propagated swiftly forward through the two-stage network, and it produces a two-element activation vector across the output units. What we would *like* the network to produce (we should be so lucky) is the output vector <1, 0>, which is our conventional output vector coding for a *mine*, since it was a mine echo we entered in the first place. But given the random configuration of synaptic weights throughout the network, that correct output would be a miracle. Most likely the network produces some random output vector nowhere near <1, 0>, such as <.49, .51>, which tells us next to nothing.

But maybe we can tweak our network to do a *little* bit better than this. To do so, we calculate, by simple subtraction, the *difference* between the output vector we actually got and the vector we wanted. This is a measure of the network's *error*. We want to reduce that error, if possible. So we fiddle with the network's 105 weights, one by one, as follows.

We begin by focusing on the *first* of those many weights. We *raise* the value of that weight by a tiny amount, and then reenter our original mine echo at the input layer, in order to see what difference this tiny weight-adjustment makes in the output vector that gets produced. In particular, we ask, does it produce

an output vector *closer* to the desired output vector of <1, 0> (closer than our disappointing <.49, .51>), if only by a small amount? If so, we leave our student weight with this new and slightly better value. If not, we then try *lowering* the weight of that connection, in order see if *that* adjustment will yield a better output vector. If it does, we fix that lowered value as the new weight of that connection. If neither fiddle improves things, we leave the weight as we found it.

We then move on to the second connection in our population of 105, and repeat the same probing procedure, in hopes of making some small improvement by fiddling with *that* connection. If either raising or lowering its weight improves the output vector for the same input, then we keep that fortunate adjustment, and move on to the network's third synaptic weight. In this way do we proceed through all 105 of the network's synaptic weights, adjusting each so as to purchase some tiny improvement on its 'judgment' concerning our mine- vector. This is a tedious procedure, to be sure, but remember that our artificial network is modeled within a standard digital computer, and so we can program that computer to do all of this testing and adjusting on its own, and much more swiftly than we can.

The result is a network that performs *slightly* better than our randomly configured original, but alas, only slightly. But of course, we have 'retuned' it with only a single echo in mind—our opening sample mine-echo. We still have 99 *further* echoes (i.e., input activation vectors) waiting for *their* turn to guide this retuning process. And so we program our computer to repeat the systematic weight-adjusting procedure described above for each and every one of those sample echoes.

We do this many times, for all 100 echoes (or rather, the programmed computer does). This is called *training up the*

network. Somewhat surprisingly, the result is that the set of synaptic weights gradually relaxes into a final configuration where the network gives a <1, 0> output vector (or close to it) when and only when the input vector is from a mine; and it gives a <0, 1> output vector (or close to it) when and only when the input vector is from a rock.

The first remarkable fact in all of this is that there *is* a configuration of the network's synaptic weights that allows the system to distinguish fairly reliably between mine echoes and rock echoes. Such a configuration exists because it turns out that there is a rough internal pattern or abstract organization that is characteristic of mine echoes as opposed to rock echoes. And the trained network has managed (finally) to lock onto that rough pattern.

If, after successfully training up the network in the fashion described, we examine the activation-vectors across the *hidden* units produced by each of the two salient kinds of input-unit stimulations, we find that such vectors already form two entirely disjoint classes, even at that intermediate level. Consider, if you will, an abstract 'vector-coding space', a space with seven dimensions, one each for the activity levels of each hidden unit. (Think of this space along the lines of the abstract sensory coding spaces in figures 7.13 and 7.14. The only difference is that the present space represents the activity patterns of cells farther along in a processing hierarchy.) Any 'mine-like' vector occurring across the hidden units falls into a large subvolume of the larger space of *possible* hidden-unit vectors. And any 'rock-like' vector falls into large but *distinct* (i.e., non-overlapping) subvolume of that abstract seven-dimensional space.

What these hidden units are doing in a trained network is successfully to code some fairly abstract structural features of

mine echoes—features they all have, or all at least approximate—despite their superficial diversity. And it does the same for rock echoes. It does all of this by slowly finding a set of synaptic weights that magnifies those structural features, and minimizes the noise due to the inevitable variation across the recorded samples, a set of weights that produces disjoint classes of hidden-unit coding vectors for each.

Given success of this sort at the level of the hidden units, what the right-hand half of the trained network does is just transform any hidden-unit mine-like vector into something close to a <1, 0> vector at the output level, and any hidden-unit rock-like vector into something close to a <0, 1> vector at that final level. In short, the final layer slowly learns to distinguish between the two salient subvolumes of the hidden-unit vector space. Vectors close to the center of either subvolume—these are the 'prototypical' examples of each type of vector—produce a clear and unambiguous verdict at the output level. By contrast, hidden-unit vectors close to the boundary dividing the two subvolumes produce a much less decisive response: a <.4, .7> perhaps. The network's 'guess' at a rock in this case is thus not very 'confident'. But such graded responses can be useful even so.

A very important by-product of this procedure is the following. If the network is now presented with entirely *new* samples of rock echoes and mine echoes—samples it has never heard before—its output vectors will categorize them correctly straight off, and with an accuracy that is only negligibly lower than the accuracy now shown on the 100 recorded samples on which it was originally trained. The new samples, novel though they may be, also produce vectors at the level of the hidden units that fall into one of the two distinguishable subvolumes of that space. In short, the 'knowledge' that the system has acquired—

due to its considerable training—*generalizes* reliably to new cases. (Or, it will if our original training set of echoes is truly representative of the two classes of echoes at issue.) Our system is finally ready to probe the enemy harbor. We just feed it the sonar returns there encountered, and the trained network will give us an informed verdict on whether or not we are approaching an enemy mine.

What is interesting here is not the proposed military application of the device described, although that was indeed a part of its origins. What is interesting, rather, is that such a *simple* brain-like system can perform the sophisticated recognitional task described. That a suitably adjusted network will do this job at all is the first marvel. The second marvel is that there exists a rule or procedure that will successfully *shape* the network into the necessary configuration of weights, even if it starts out in a random configuration. That procedure makes the system learn from the 100 sample echo-vectors we provided it, plus the sequential errors that it produces at the output units. This process is called *automated learning by the back-propagation of errors*, and it is relentlessly efficient. For it will regularly find order and structure, within a set of instructional examples, where initially you or I would see only chaos and confusion. This learning process is an instance of what is called *gradient descent*, because the configuration of weights in the student system can be seen as sliding down a meandering slope of ever-decreasing output errors until it enters the narrow region of a lowest valley, at which the performance-errors get closer and closer to zero. (See figure 7.20 for a simplified representation of this process.)

Training up the network on many sample echoes may take many hours, or even days, but once the system is trained, it will

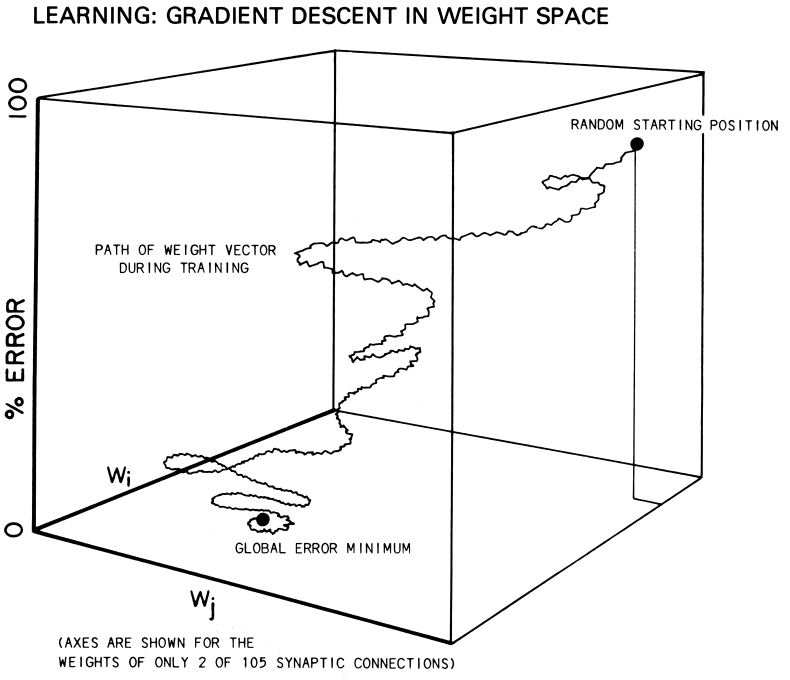

Figure 7.20

yield up a verdict on any given sample in only a few seconds. The computer in which our network is being modeled must calculate, *separately* and in *sequence*, each of the local transformations performed at each of the 105 'synaptic' connections, but since it is a modern computer with a fast CPU, it does this fairly swiftly. You can see, however, that if it could perform all of the calculations for the hidden-unit connections *simultaneously*, as would happen in a genuinely biological neural network, then that overall computation would be fully $(13 \times 7 =)$ 91 times faster still! And so for every other layer. The genuinely *parallel*

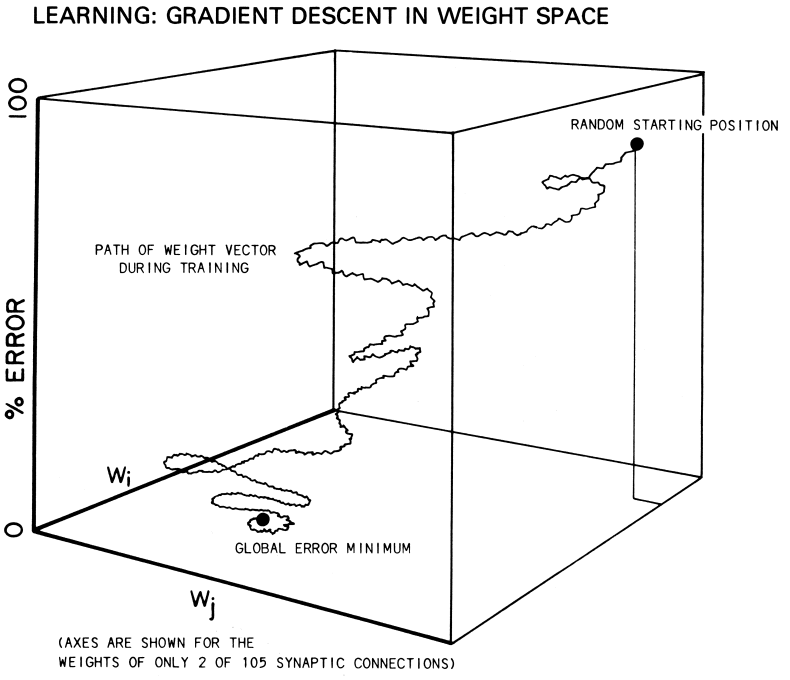

Figure 7.20

yield up a verdict on any given sample in only a few seconds. The computer in which our network is being modeled must calculate, *separately* and in *sequence*, each of the local transformations performed at each of the 105 'synaptic' connections, but since it is a modern computer with a fast CPU, it does this fairly swiftly. You can see, however, that if it could perform all of the calculations for the hidden-unit connections *simultaneously*, as would happen in a genuinely biological neural network, then that overall computation would be fully $(13 \times 7 =)$ 91 times faster still! And so for every other layer. The genuinely *parallel*

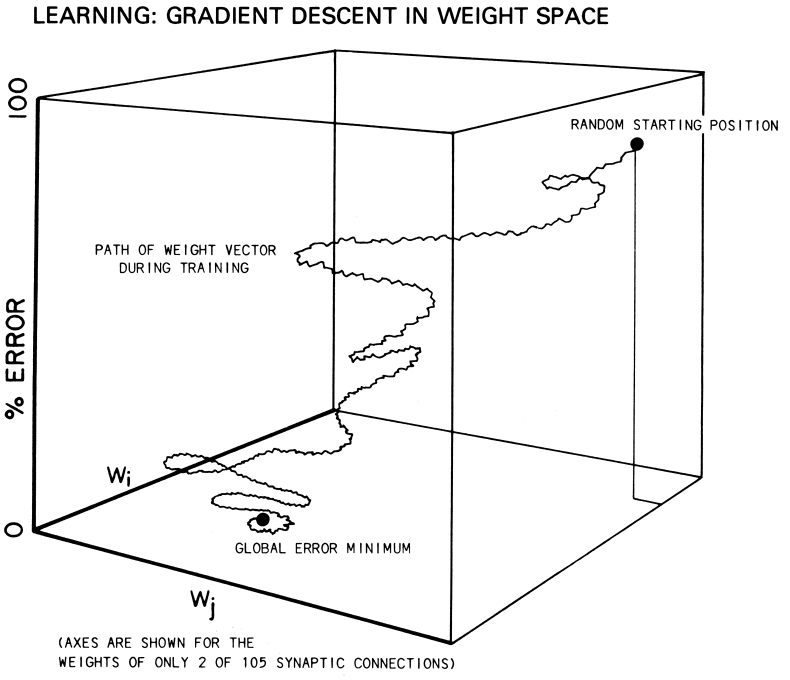

Figure 7.20

yield up a verdict on any given sample in only a few seconds. The computer in which our network is being modeled must calculate, *separately* and in *sequence*, each of the local transformations performed at each of the 105 'synaptic' connections, but since it is a modern computer with a fast CPU, it does this fairly swiftly. You can see, however, that if it could perform all of the calculations for the hidden-unit connections *simultaneously*, as would happen in a genuinely biological neural network, then that overall computation would be fully $(13 \times 7 =)$ 91 times faster still! And so for every other layer. The genuinely *parallel*

distributed processing displayed in biological neural networks remains a stunning advantage, one that only gets larger as the networks at issue get larger. Since, as we saw, the brain regularly displays connections matrices with *millions* of synapses, all doing their jobs simultaneously, the speed advantage over a conventional computer becomes overwhelming. No wonder your brain is so much smarter than a desktop computer on real-world cognitive tasks.

Further Examples and General Observations

I have focused closely on the rock/mine network in order to provide some real detail on how a parallel network does its job. But the example is only one of many. If mine echoes can be recognized and distinguished from other kinds of sounds, then a suitably trained network of this general kind should able to recognize the various *phonemes* that make up English speech and not be troubled at all by the wide differences in the character of people's voices, as traditional AI programs are. Truly effective speech-recognition is thus now within reach.

Nor is there anything essentially auditory about the talents of such networks. They can be 'trained up' to recognize complex visual features just as well. A recent neural network can tell us the 3D shape and orientation of a smoothly curved two-dimensional surface given only a grayscale photo of the surface in question. That is, it solves the traditional problem in visual psychology of how we can divine 'shape from shading', which all of us do effortlessly.

Nor is there anything essentially perceptual about their talents. They can be used to produce interesting motor output just as easily. A very simple network has been produced that will direct an artificial creature's two-jointed arm to reach out and

grab an object that appears at a specific point within its visual field. This is a simple case of 'sensorimotor coordination'. And a rather larger network has already learned to solve, for example, the problem of converting printed text (as input) into audible speech (as output). That is, it has learned to *read aloud* as motor output (through a voice-synthesizer, of course) given printed text as input (through a visual scanner, of course). This is Sejnowski and Rosenberg's celebrated network called NETtalk. It uses a vector-coding scheme for input letters, another vector-coding scheme for output phonemes, and it was gradually taught to perform the appropriate vector-to-vector transformations. In plain English, it learned to pronounce printed words. And it does so without being given any *rules* to follow whatsoever. This is no mean feat, especially given the wanton irregularities of standard English spelling. The system must learn not just to transform the letter "a" into a certain sound. It must learn to transform "a" into one sound when it occurs in "save," into another when it occurs in "have," and into a third when it occurs in "ball." It must learn that "c" is soft in "city," but hard in "cat," and something else again in "cello." And so on and so on, as we all learned in grade school.

Initially, of course, it does none of this. When fed printed text in its untrained state, its output vectors produce, through its sound synthesizer, nonsense babbling rather like a baby's: "nananoonoo noonanaah." But each of its erroneous output vectors is analyzed by the standard computer that is monitoring the process. The network's many synaptic weights are adjusted according to the 'back-propagation of error' procedure discussed above. And the quality of its babbling slowly improves as we feed it many examples of printed English text. After only ten hours of training on a sample of 1,000 words, it produces,

if a little clumsily, coherent, intelligible speech given arbitrary English text as input. And it does so without explicit rules being represented anywhere within the system.

Are there any limits to the transformations that a parallel network of this general kind can perform? Current opinion among workers in the field is that there are no obvious theoretical limits, since the new networks have important features that the early networks of the late 1950s did not have. For example, the axonal output signal produced by any neuronal unit need no longer be a straight or 'linear' function of the level of activation within the unit itself. In the current generation of artificial networks, it typically follows a kind of S-curve. This simple wrinkle allows a network to compute, at least approximately, what are called *nonlinear* vector-transformations, and this broadens dramatically the range of cognitive problems it can handle.

Equally important, the new networks have one or more layers of 'hidden' units intervening between the input and output levels, where the early networks had only an input layer connected directly to an output layer. The advantage of the intervening layer(s) is that, within that layer, the system can explore possible features that are not explicitly available in the input vectors, as we saw in the mine/rock network. This process is iterable, and it allows a many-layered network to dig progressively more deeply into the structural complexities implicit in the sensory input domain.

You can now appreciate why artificial networks of the kind at issue have captured so much attention. Their microstructure is similar in many respects to that of the brain, and they display at least some of its hard-to-simulate functional properties.

How far does the analogy go? Is this really how the brain might work? I don't know. But the research program sketched

above is plainly worth pursuing. However, let me close this section by addressing an obvious *defect* in the examples proposed so far. The problem is the learning procedure typically used to train the networks: the *back-propagation of errors* procedure. It is highly effective, to be sure, but it cannot be how the biological brain learns. For one thing, that artificial procedure requires that someone or something know *at the outset* what the network's final behavior *should be*. For only then can we identify, and quantify, the network's *error* at any stage of its training. But biological creatures in the real world possess no such information. (If they did, they wouldn't *need* to learn!) Back-propagation is an example of *supervised* learning. But learning in the real world is almost always *un*supervised.

A second defect with that artificial procedure is that, even if the brain were somehow given the sorts of systematic error-reports that it requires, the brain has no way to convey that information, bit by bit, to the specific synaptic connection that is responsible for it, so as to increase, or decrease, its weight appropriately. The learning brain, apparently, must use a learning procedure quite different from the back-propagation of carefully determined errors.

And so, it seems, it does. A process called *Hebbian learning* appears to be the primary determinant of synaptic change in the *biological* brain. It was initially proposed in the middle of the last century by the psychologist D. O. Hebb, but we have only slowly come to appreciate its full potential for sculpting the weight configuration within any network.

The basic idea is that a given synapse will progressively increase its weight when and only when the arrival of a strong signal from its own axonal end-branch chronically *coincides* with a high level of activation in the receiving neuron. Since

above is plainly worth pursuing. However, let me close this section by addressing an obvious *defect* in the examples proposed so far. The problem is the learning procedure typically used to train the networks: the *back-propagation of errors* procedure. It is highly effective, to be sure, but it cannot be how the biological brain learns. For one thing, that artificial procedure requires that someone or something know *at the outset* what the network's final behavior *should be*. For only then can we identify, and quantify, the network's *error* at any stage of its training. But biological creatures in the real world possess no such information. (If they did, they wouldn't *need* to learn!) Back-propagation is an example of *supervised* learning. But learning in the real world is almost always *un*supervised.

A second defect with that artificial procedure is that, even if the brain were somehow given the sorts of systematic error-reports that it requires, the brain has no way to convey that information, bit by bit, to the specific synaptic connection that is responsible for it, so as to increase, or decrease, its weight appropriately. The learning brain, apparently, must use a learning procedure quite different from the back-propagation of carefully determined errors.

And so, it seems, it does. A process called *Hebbian learning* appears to be the primary determinant of synaptic change in the *biological* brain. It was initially proposed in the middle of the last century by the psychologist D. O. Hebb, but we have only slowly come to appreciate its full potential for sculpting the weight configuration within any network.

The basic idea is that a given synapse will progressively increase its weight when and only when the arrival of a strong signal from its own axonal end-branch chronically *coincides* with a high level of activation in the receiving neuron. Since

that receiving neuron will typically display a high level of activation precisely *because* a considerable number of its *other* synaptic connections, from various other neurons earlier in the connective hierarchy, are *also* bringing a strong signal at the very same time, what we are looking at is a process that increases the synaptic weight of *all* and only the specific connections involved in this collective and simultaneous chorus. That is, if some specific cadre of neurons within the *preceding* layer sends a set of strong signals to our receiving neuron *all at the same time*, then all of its synaptic connections *from* that specific cadre will have their weights increased slightly as a result. And if that same cadre *repeatedly* stimulates our receiving neuron in this collective fashion, then the relevant *family* of synaptic weights will all go up substantially and permanently. The result is that our receiving neuron has become preferentially sensitive to (i.e., reacts most strongly to) a specific *pattern* of activation-levels within the preceding layer of neurons—specifically, the pattern of the excited cadre—and it does so because that simultaneous pattern has happened *more frequently* in the network's experience than most of the other possible patterns.

In this way can a single neuron, downstream from the sensory layer, become a reliable indicator of some frequently repeated aspect of the network's ongoing experience of the world. And it will do so without any prescient supervision, at the outset, concerning how the network *should* behave in response to that world. Hebbian learning is a process that manages to pull out, all by itself, the dominant patterns or structures that are displayed in the creature's own experience. It doesn't need a teacher, beyond the world itself.

Hebbian learning has other virtues as well, such as the capacity to make a 'recurrent' network (that is, a network with

'backward' axonal projections as well as purely 'forward' ones, as in all of the examples explored above) selectively sensitive to certain prototypical *sequences* or *temporal patterns* displayed in its experience, such as the ongoing gait of a walking human, or the flapping motion of a flying bird, or the repeating arc of a bouncing ball. And a good thing, too, since recognizing salient patterns unfolding in time is at least as important to a cognitive creature as is recognizing salient structures in space. I'll spare you the details of how this specific capacity is achieved, but they are part of a fertile story that continues to unfold. The fact is, AI, Cognitive Science, and Neuroscience are now interacting vigorously. They are now teaching each other, a process from which everyone will profit.

A final observation. According to the style of theory we have here been exploring, it is high-dimensional neuronal activation-vectors that constitute the most important kind of representation within the brain. And it is vector-to-vector transformations that form the most important kind of computation. If this is correct, then it gives some substance to the earlier suggestion of the eliminative materialist (chapter 2.5) that the concepts of Folk Psychology need not, and perhaps do not, capture the dynamically significant states and activities of the mind. The elements of cognition, as sketched in the preceding pages, have a character unfamiliar to common sense. Perhaps we should actively expect that, as our theoretical understanding here increases, our very conception of the phenomena we are trying to explain will undergo significant revision as well. This is a common pattern throughout the history of science, and there is no reason why the cognitive sciences should prove any exception.

Suggested Readings

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. "Learning Representations by Back-Propagating Errors." *Nature* 323 (Oct. 9, 1986): 533–536.

Sejnowski, T. J., and C. R. Rosenberg. "Parallel Networks That Learn to Pronounce English Text." *Complex Systems* 1 (1987).

Churchland, P. S., and T. J. Sejnowski. *The Computational Brain*. Cambridge, MA: MIT Press, 1992.

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

Churchland, P. M. *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA: MIT Press, 2012.