1 Introduction

The goal of science is not to open the door to everlasting wisdom, but to set a limit on everlasting error.
Galileo, in *Galileo*, by Bertolt Brecht

1 Core Questions

Bit by experimental bit, neuroscience is morphing our conception of what we are. The weight of evidence now implies that it is the *brain*, rather than some nonphysical stuff, that feels, thinks, and decides. That means there is no soul to fall in love. We do still fall in love, certainly, and passion is as real as it ever was. The difference is that now we understand those important feelings to be events happening in the physical brain. It means that there is no soul to spend its postmortem eternity blissful in Heaven or miserable in Hell. Stranger yet, it means that the introspective *inside*—one's own subjectivity—is *itself* a brain-dependent way of making sense of neural events. In addition, it means that the brain's *knowledge* that this is so is likewise brain-based business.

Given what is known about the brain, it also appears highly doubtful that there is a special nonphysical module, the *will*, operating in a causal vacuum to create voluntary choices—choices to be courageous in the face of danger, or to run away and fight another day. In all probability, one's decisions and plans, one's self-restraint and self-indulgences, as well as one's unique individual character traits, moods, and temperaments, are all features of the brain's general causal organization. The self-control one thinks one has is anchored by neural pathways and neurochemicals. The mind that we are assured can dominate over matter is in fact certain brain patterns interacting with and interpreted by other brain patterns. Moreover, one's *self*, as apprehended introspectively and

represented incessantly, is a brain-dependent construct, susceptible to change as the brain changes, and is gone when the brain is gone.

Consciousness, almost certainly, is not a semimagical glow emanating from the soul or permeating spooky stuff. It is, very probably, a coordinated pattern of neuronal activity serving various biological functions. This does not mean that consciousness is not real. Rather, it means that its reality is rooted in its neurobiology. That a brain can come to know such things as these, and in particular, that it can do the science of itself, is one of the truly stunning capacities of the human brain.

This list catalogues but a few of the scientific developments that are revolutionizing our understanding of ourselves, and one would have to be naive to suppose that things have "gone about as far as they can go." In general terms, the mind-body problem has ceased to be the reliably tangled conundrum it once was. During the last three decades, the pace of discovery in neuroscience has been breathtaking. At every level, from neurochemicals to cells, and onwards to the circuit and systems levels, brain research has produced results bearing on the nature of the mind (figures 1.1 and 1.2). Coevolving with neuroscience, cognitive science has probed the scope of large-scale functions such as attention, memory, perception, and reasoning both in the adult and in the developing infant. Additionally, computational ideas for linking large-scale *cognitive* phenomena with small-scale *neural* phenomena have opened the door to an *integration* of neuroscience, cognitive science, and philosophy in a comprehensive theoretical framework.

There remain problems galore, and the solution to some of these problems will surely require conceptual and theoretical innovation of a magnitude that will surprise the pants off us. Most assuredly, having achieved significant progress does *not* imply that only mopping-up operations remain. But it does mean that the heyday of unfettered and heavy-handed philosophical speculation on the mind has gone the way of the divine right of kings, a passing that has stirred some grumbling among those wearing the mantle of philosopher-king. It does mean that know-nothing philosophy is losing ground to empirically constrained theorizing and inventive experimentation.

If the aforementioned changes have emerged from discoveries in the various neurosciences—including neuroanatomy, neurophysiology, neuropharmacology, and cognitive science—wherefore *philosophy*? What is *neurophilosophy*, and what is *its* role? Part of the answer is that the nature of the mind (including the nature of memory and learning, consciousness, and free will) have traditionally been subjects within the purview of philosophy. Philosophers, by tradition,

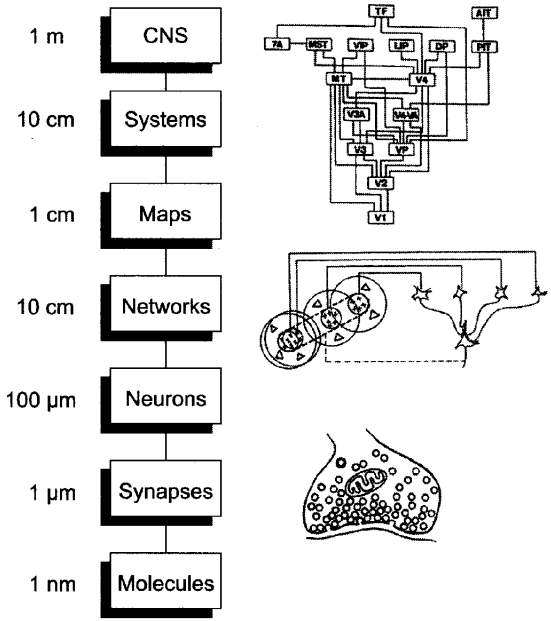

Figure 1.1 Organized structures are found at many spatial scales in nervous systems. Functional levels may be even more fine-grained. Thus dendrites are a smaller computational unit than neurons, and networks may come in many sizes, including local networks and long-range networks. Networks may also be classed according to distinct dynamical properties. Icons on the right depict distinct areas in the visual system (top), a network (middle), and a synapse (bottom). (Based on Churchland and Sejnowski 1988.)

have wrestled with these topics, and the work continues. Neurophilosophy arises out of the recognition that at long last, the brain sciences and their adjunct technology are sufficiently advanced that real progress can be made in understanding the mind-brain. More brashly, it predicts that philosophy of mind conducted with no understanding of neurons and the brain is likely to be sterile. Neurophilosophy, as a result, focuses on problems at the intersection of a greening neuroscience and a graying philosophy.

Another part, perhaps the better part, of the answer is that philosophy, traditionally and currently, is quintessentially the place for synthesizing results and integrating theories across disciplinary domains. It is panoramic in its scope and all-encompassing in its embrace. It unabashedly bites off much more than it can chew. *Any* hypothesis, be it ever so revered or ever so scorned, is considered fair game for criticism. Philosophy deems it acceptable to kick the

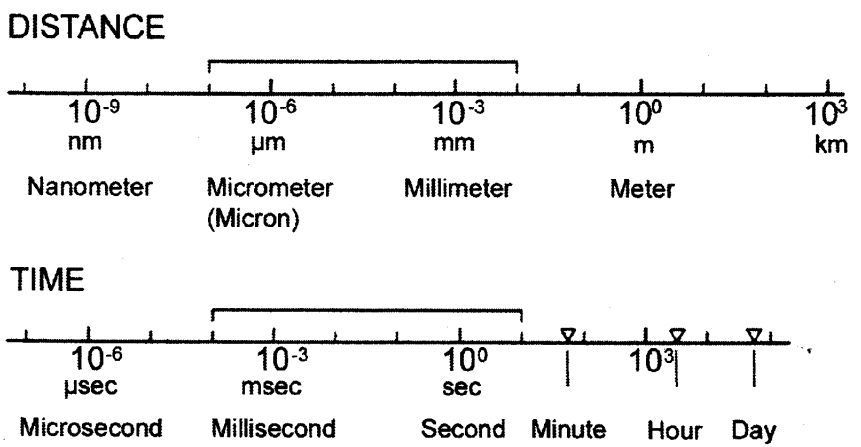

Figure 1.2 Logarithmic scales for spatial and temporal magnitudes. Brackets indicate the scales especially relevant to synaptic processing. (Based on Shepherd 1979.)

tires of every governing paradigm, examine every sacred cow, and peer behind the curtains of every magic show.

Under *this* description, we are all philosophers from time to time. Certainly, scientists have their philosophical hours, when they push back from the bench and stew on the broad questions, or when they beat on the conventional wisdom and strike a blow for originality. Such philosophical hours prepare the ground for the germination of new ideas and new experimental techniques.

Politely, we can consider philosophy the theoretical companion to experimental science; less politely, we can consider it merely woolgathering and freelancing. Certainly, some philosophy *is* just horsing around. Yet that is no bad thing, especially when a science is in its nascent stages. Neuroscience *is* a nascent science, and theoretical innovation is needed in every subfield of that broad über-field. Most theoretical ideas are bound to be losers, of course, but unless we are courageous enough to nurture lots and lots of new ideas, the rightful winners will never see the light of day.

This description highlights the positive side of philosophizing, but as with anything else, there is a seamier side. This is the side revealed when one is lulled into taking one's untested theoretical fancies as fact, or equating theory *beautiful* with theory *true*, or rejecting unorthodox ideas as heresy *because* they are unorthodox, or supposing that some chummy circle has the corner on clever ideas. If this applies to philosophy, it applies just as well to science, government, finance, and war.

This book is about neurophilosophy. It aims to take stock of various philosophical problems concerning the nature of the mind, given the recent bonanza of developments in neuroscience and cognitive science. In finding a path through the thicket of relevant neuroscientific studies and discoveries, I found material assembling itself into two classical categories: metaphysics and epistemology. Ethics gets a brief look in my discussion of free will and responsibility, but is mainly undiscussed on this occasion. Religion is the subject of the closing chapter, and has both a metaphysical and an epistemological dimension.

Before plunging on, we shall limber up with a few brief historical points and a short discussion on *reductionism*, a pivotal concept whose clarity is no luxury as we begin to assay the integration of hitherto separated domains.[1]

2 Natural Philosophy

Greek thought in the period 600 B.C. to 200 A.D. was the fountainhead for Western philosophy generally, as well as for modern science. In those days, *philosophy* literally meant "love of wisdom," and for the ancient Greeks, philosophy targeted a vast range of questions, such as, What is the nature of change such that water can freeze or wood burn? What is the nature of the moon and stars, and where did Earth come from? Are there fundamental particles of which all objects are composed? How do living things reproduce? In addition, of course, they raised questions about themselves—about what it is to be human, to think and perceive, to reason and feel, to plan and decide, to live a good life, to organize a harmonious and productive political state.

Theories about the *natural* world were considered part of *natural philosophy*. By contrast, theories of ethics and politics and practical life were part of *moral* philosophy. To a first approximation, this classification separates questions about *how things are* from questions about *what we should do*. Though distinct, these two domains share concepts and theories. In particular, sometimes questions about the mind will have one foot in each of these areas.

When did philosophy come to be considered a separate discipline? By the end of the nineteenth century, advances in some domains of natural philosophy had developed so extensively that separate subfields—physics, chemistry, astronomy and biology—branched off as distinct sciences. With progress and specialization, the expression "natural *science*" gained currency, while the more old-fashioned term, "natural *philosophy*" faded from use, now being essentially archaic. Nonetheless, this broad title can still be found on science buildings and

doorways in older universities such as Cambridge in England and St. Andrews in Scotland. Until the middle of this century, St. Andrews's degrees in physics were officially degrees in Natural Philosophy. The title Ph.D. (*Philosophae Doctor*, or "teacher of philosophy") is awarded not only to philosophers, but to scientists of all sorts. It is a vestige of the older classification, which embraces all of science as a part of natural philosophy.

If the stars, the heart, and the basic constituents of matter became understood well enough to justify a separate science, what about the mind? Ancient thinkers, such as the physician Hippocrates (460–377 B.C.), were convinced that thoughts, feelings, and perceptions were activities of the brain. He believed that events such as sudden paralysis or creeping dementia had their originating causes in brain damage. And this implied, in his view, that normal movement and normal speech had *their* originating causes in the *well-tempered* brain. On the other hand, philosophers favoring a nonnatural framework—Plato (427–347 B.C.), and especially later Christian thinkers such as St. Thomas Aquinas (1225–1274) and St. Augustine (354–430)—believed the soul to be distinct from the body and divine in origin. Plato, in perhaps the first systematic theorizing on the soul, hypothesized it to have a sensible part (which determines perceptions), an emotional part (by virtue of which we feel honor, fear, and courage), and a rational part. This last was considered unique to humans and allowed us to reason, think, and figure things out. Theologically minded philosophers concluded that the mind (or, one might say, *the soul*) was a subject for study by means other than those available to natural science. If supernaturalism was true of the soul, then the nature of the soul could not be revealed by natural science, though perhaps other methods—such as meditation, introspection, and reason—might be useful.

Descartes (1595–1650) articulated the modern version and systematic defense of the idea that the mind is a *nonphysical* thing. This dual-substance view is known as *dualism*. Reason and judgment, in Descartes's view, are functions inhering in the mental, *immaterial* mind. He surmised that the mind and the body connect at *only* two points: sensory input and output to the muscles. Apart from these two functions, Cartesian dualism assumes that the mind's operations in thought, language, memory retrieval, reflection and conscious awareness proceed *independently* of the brain. When clinical studies on brain-damaged patients showed clear dependencies between brains and *all* these ostensibly brain-independent functions, classical dualism had to be reconfigured to allow that brain-soul interactions were not limited to sensory and motor

functions. Achieving this correction without rendering the soul explanatorily redundant has been the bane of post-Cartesian dualism.

What about dualism appealed to Descartes? First, he was particularly impressed by the human capacity for reasoning and language, and the degree to which language use seems to be governed by reasons rather than causes. More exactly, he confessed that he was completely unable to imagine how a mechanical device could be designed so as to reason and use language appropriately and creatively.

What sort of mechanical devices were available to propel Descartes's imagination? Only clockwork machines, pumps, and fountains. Though some of these were remarkably clever, even the most elaborate clockwork devices of the seventeenth century were just *mechanical.* Well beyond the seventeenth-century imagination are modern computers that can guide the path of a cruise missile or regulate the activities of a spacecraft on Mars. In an obvious way, Descartes's imagination was limited by the science and technology he knew about. Had he been able to contemplate the achievements of computers, had he had even an inkling of electronics, his imagination might have taken wing. On the other hand, the core of Descartes's argument was revived in the 1970s by Chomsky[2] and Fodor[3] to defend their conviction that nothing we will ever understand about the brain will help us very much to understand the nature of language production and use.

The second reason dualism appealed is closely connected to the first. Descartes was convinced that exercise of free will was inconsistent with causality. He was also sure that humans did indeed have free will, and that physical events were all caused. So even if the body was a just a mechanical device, the mind could not be. Minds, he believed, must enjoy *uncaused* choice. We can undertake an action for a reason, but the relations between reasons and choices are not causal. Animals, by contrast, he believed to be mere automata, without the capacity for reason or for free choice. In its core, if not in its details, this argument too is alive and well even now, and it will be readdressed in greater detail in chapter 5 in the context of the general topic of free will.

Third, Descartes was impressed by the fact that one seems to know one's own conscious experiences simply by *having* them and *attending* to them. By contrast, to know about *your* experiences, I must draw inferences from your behavior. Whereas I know I have a pain simply by having it, I must draw an inference to know that my body has a wound. I cannot be wrong that *I* am conscious, but I can be wrong that *you* are conscious. I can even be wrong that

you exist, since "you" might be nothing but *my* hallucination. According to Descartes's argument, differences in *how* we know imply that the thing that has knowledge—the mind—is fundamentally different from the body. The mind, he concluded, is essentially immaterial and can exist after the disintegration of the body. Like the other two arguments for dualism, this argument has remained powerful over the centuries. It has been touched up, put in modern dress, and in general reworked to look as good as new, but Descartes's insights regarding knowledge of mental states constitute the core of virtually all recent work on the nonreducibility of consciousness.[4] Because it continues to be persuasive, this argument will be readdressed and analyzed in detail when we discuss self-knowledge and consciousness. (See especially chapter 3, but also chapters 4 and 6.)

How, in Descartes's view, is the body able causally to affect the mind so that I feel pain when touching a hot stove? How can the mind affect the body so that when I decide to scratch my head, my body does what I intend it should do? Although Descartes envisioned interaction as limited to sensory input and motor output, notice that the business of interaction—*any* interaction—turns out to be a vexing problem for dualism, no matter how restricted or rich the interactions are believed to be. The interaction problem was, moreover, recognized as trouble right from the beginning. How could there be any causal interaction *at all*, was the question posed by other philosophers, including his contemporary, Princess Elizabeth of Holland, who put her objection bluntly in a letter of 10/20 June 1643: "And I admit that it would be easier for me to concede matter and extension to the soul than to concede the capacity to move a body and be moved by it to an immaterial thing" (*Oeuvres de Descartes*, ed. C. Adam and P. Tannery, vol. III, p. 685). As Princess Elizabeth realized, the mind, as a mental substance, allegedly has *no* physical properties; the brain, as a physical substance, allegedly has *no* mental properties. Slightly updated, her question for Descartes is this: how can the two radically different substances interact? The mind allegedly has no extension, no mass, no force fields—*no physical properties at all.* It does not even have spatial boundaries or locations. How could a nonphysical thing *cause a change* in a physical thing, and vice versa? What could be the causal basis for an interaction? Somewhat later, Leibniz (1646–1716) described the problem as intractable:[5] "When I began to meditate about the union of soul and body, I felt as if I were thrown again into the open sea. For I could not find any way of explaining how the body makes anything happen in the soul, or vice versa, or how one substance can communicate with another created substance. Descartes had given up the game at this

point, as far as we can determine from his writings" (from *A New System of Nature*, translated by R. Ariew and Daniel Garber, p. 142).

Descartes almost certainly did recognize that mind-body interaction was a devastating difficulty, and indeed it has remained a stone in the shoe of dualism ever since. (For additional discussion, see chapter 2.)

The difficulty of giving a positive account provoked some philosophers, Leibniz being the first, to assert that events in a nonphysical mind are simply separate phenomena running in parallel to events in the brain. The mind causes nothing in the brain, and the brain causes nothing in the mind. Known as psychophysical parallelism, the idea was that the parallel occurrence of mental and brain events gives the illusion of causal interaction, though in fact no such causation ever actually occurs. What keeps the two streams in register? Some parallelists, such as Malebranche, thought this was a job God regularly and tirelessly performs for every conscious subject every waking hour. Leibniz, who preferred the idea that God kicked off the two streams and then let them alone, disparaged "occasionalists" such as Malebranche: "[Descartes's] disciples ... judged that we sense the qualities of bodies because God causes thoughts to arise in the soul on the occasion of motions of matter, and that when our soul, in turn, wishes to move the body, it is God who moves the body for it" (p. 143).

Descartes's best attempt to explain the interaction between mind and body was the suggestion that some unobserved but very, *very* fine material—*material*—in the pineal gland of the brain brokered the interaction between nonphysical mind and physical brain. His critics, such as Leibniz, were not fooled.

Perhaps Descartes was not fooled either. Some historians argue that Descartes's defense of a fundamental difference between mind and body was actually motivated by political rather than intellectual considerations.[7] Descartes was unquestionably a brilliant scientist and mathematician. This is, after all, the Descartes of the Cartesian coordinate system, a stunning mathematical innovation for which he is rightly given credit. He also understood well the bitter opposition of the Church to developments in science, and had left France to live in Holland to avoid political trouble. It is possible that he feared that developments in astronomy, physics, and biology would be cut off at the knees unless the Church was reassured that the "soul" was its unassailable proprietary domain. Such a division of subject matter might permit science at least to have the body as *its* domain. Whether this interpretation does justice to the truth remains controversial.

Certainly some of Descartes's arguments, both for the existence of God as well as for the mind/body split, are sufficiently flawed to suggest that they are

ostentatiously flawed. On this hypothesis, the genius Descartes knew the logic full well and planted the flaws as clues for the discerning reader. And certainly Descartes had good reason to fear the Church's power to thwart scientific inquiry and to punish the scientist. Burning, torturing, and exiling those who inquired beyond official Church doctrine was not uncommon. Galileo, for example, was "shown the instruments of torture" to force him to retract his claim that Earth revolved around the Sun, a claim based on observation and reasoning. Recant he did, rather than submitting to the rack and iron maiden, but even so, he spent the rest of his life under house arrest by Church authorities. By vigorously postulating the mind/body division, perhaps contrary to his own best scientific judgment, Descartes may have done us all a huge, if temporary, favor in permitting the rest of science to go forward.

And go forward it did. By the end of the nineteenth century, physics, chemistry, astronomy, geology, and physiology were established, advanced scientific disciplines. The science of nervous systems, however, was a much slower affair. Though some brilliant *anatomical* work had been done on nervous systems, particularly by Camillo Golgi (1843–1926) and Santiago Ramón y Cajal (1852–1934), even at the end of the nineteenth century, little was known about the brain's functional organization, and almost nothing was understood concerning how neurons worked. That neurons signaled one another was a likely hypothesis, but how and to what purpose was a riddle.

Why did progress in neuroscience lag so far behind progress in astronomy or physics or chemistry? Why is the blossoming of neuroscience really a late-twentieth-century phenomenon? This question is especially poignant since, as noted, Hippocrates some four hundred years B.C. had realized that the brain was the organ of thought, emotion, perception, and choice.

The crux of the problem is that brains are exceedingly difficult to study. Imagine Hippocrates observing a dying gladiator with a sword wound to the head. The warrior had lost fluent speech following his injury, but remained conscious up to the end. At autopsy, what theoretical resources did Hippocrates possess to make sense of something so complex as the relation between the loss of fluent speech and a wound in the pinkish tissue found under the skull? Remember, in 400 B.C. nothing was understood about the nature of the cells that make up the body, let alone of the special nature of cells that make up the brain. That *cells* are the basic building blocks of the body was not really appreciated until the seventeenth century, and neurons were not seen until 1837, when Purkyně, using a microscope, first saw cell bodies in a section of brain tissue (figure 1.3).[8] Techniques for isolating neurons—brain cells—to

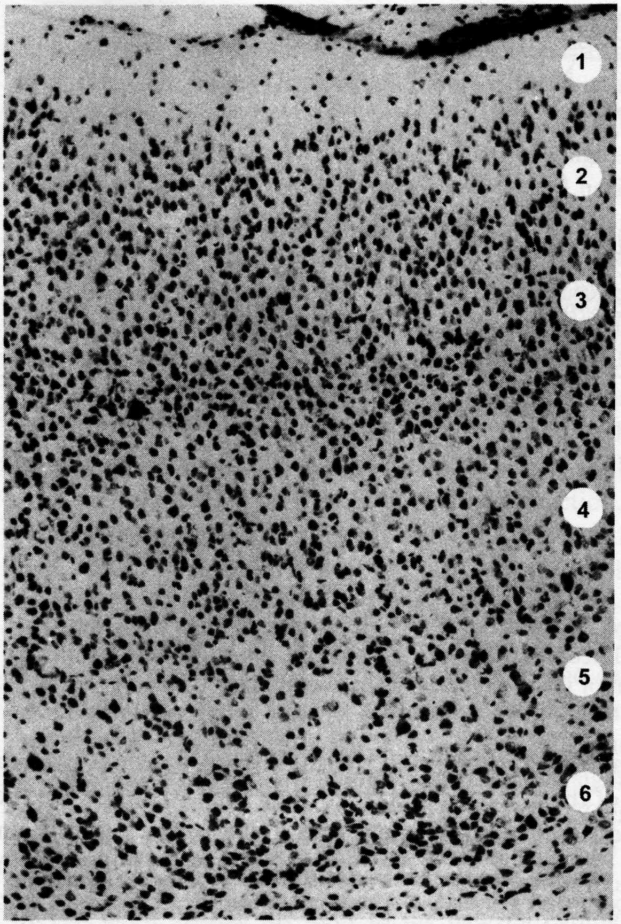

Figure 1.3 A cross-section through the mink visual cortex, with cresyl violet used to stain all cell bodies. Cortical layers are numbered at the right. (Courtesy of S. McConnell and S. LeVay.)

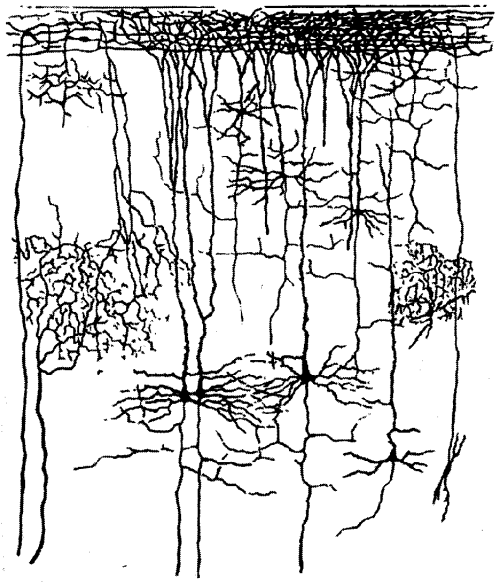

Figure 1.4 A drawing of Golgi-stained neurons in the rat cortex. About a dozen pyramidal neurons are stained, a tiny fraction of the neurons packed into the section. The height of the section depicted is about 1 mm. (Based on Eccles 1953.)

reveal their long tails and bushy arbors were not available until the second half of the nineteenth century, when stains that filled the cell were invented by Deiters (carmine stain) and then Golgi (silver nitrate stain) (figure 1.4). Neurons are *very* small, and unlike a muscle cell, each neuron has long branches—its axon and dendrites. There are about a 10^5 neurons per cubic *millimeter* of cortical tissue, for example, and about 10^9 synapses. (A handy rule of thumb is about 1 synapse/μm^3.) Techniques for isolating living neurons to explore their function did not appear until well into the twentieth century.[9]

By contrast, Copernicus (1473–1543), Galileo (1564–1642), and Newton (1643–1727) were able to make profound discoveries in astronomy without highly sophisticated technology. Through a clever reinterpretation of traditional astronomical measurements, Copernicus was able to figure out that Earth was not the center of the universe, thus challenging geocentrism. With a low-tech telescope, Galileo was able to see for the first time the moons of Jupiter and the craters of our own moon, thus undermining the conventional wisdom concerning the absolute perfection of the Heavens and the uniqueness of Earth.

Figuring out how neurons do what they do requires *very* high-level technology. And *that*, needless to say, depends on an immense scientific infrastructure: cell biology, advanced physics, twentieth-century chemistry, and post-1953 molecular biology. It requires sophisticated modern notions like molecule and protein, and modern tools like the light microscope and the electron microscope, and the latter was not invented until the 1950s. Many of the basic ideas can be grasped quite easily now, but discovering those ideas required reaching up from the platform of highly developed science.

To have a prayer of understanding nervous system, it is essential to understand how neurons work, and that was a great challenge technically. The most important *conceptual* tool for making early progress on nervous systems was the theory of electricity. What makes brain cells special is their capacity to signal one another by causing fast microchanges in each others' *electrical* states. Movement of ions, such as Na^+, across the cell membrane is the key factor in neuronal signaling, and hence in neuronal function. Living as we do in an electrical world, it is sobering to recall that as late as 1800, electricity was typically considered deeply mysterious and quite possibly occult. Only after discoveries by Ampere (1775–1836) and Faraday (1791–1867) at the dawn of the nineteenth century was electricity clearly understood to be a *physical* phenomenon, behaving according to well-defined laws and capable of being harnessed for practical purposes. As for neuronal membranes and ions and their role in signaling, understanding these took much longer (figures 1.5 and 1.6).

Once basic progress was made on how neurons signal, it could be asked *what* they signal; that is, what do the signals mean. This question too has been extremely hard to address, though the progress in the 1960s correlating the response of a visual-system neuron to a specific stimulus type, such as a moving spot of light, opened the door to the neurophysiological investigation of sensory and motor systems,[10] and to the discovery of specialized, mapped areas.

Beginning in the 1950s, progress had been made in addressing learning and memory at the systems level, and by the late 1970s, intriguing data on neuronal changes mediating system plasticity permitted the physiology of learning and memory to really take off. Meanwhile the role of specific neurochemicals in signaling and modulating neuronal function was beginning to be unraveled, and associated with large-scale effects such as changes from being awake to being asleep, to memory performance, to pain regulation, and to pathological conditions such as Parkinson's disease and obsessive-compulsive disorder. By the 1980s, attention functions came within the ambit of neuroscience, and changes at the neuronal level could be correlated with shifts in attention.

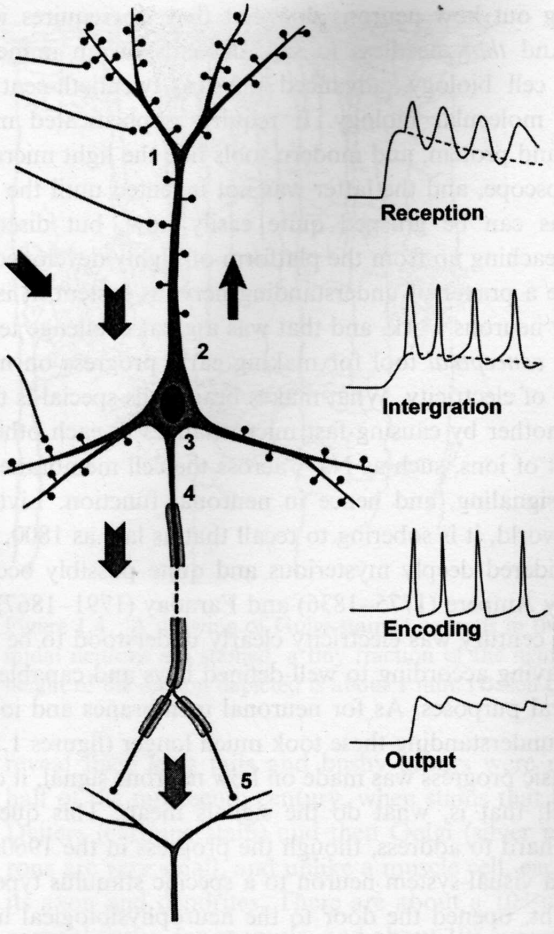

Figure 1.5 Neurons have four main structural regions and five main electrophysiological functions. The dendrites (2) have little spines (1) projecting from them, which are the major sites of in-coming signals from other neurons. The soma (3) contains the cell nucleus and other organelles involved in cell respiration and polypeptide production. Integration of signals takes place along the dendrites and soma. If signal integration results in a sufficiently strong depolarization across the cell membrane, a spike will be generated on the membrane where the axon emerges and will be propagated down the axon (4). Spikes may also be propagated back along dendritic membrane. When a spike reaches the axon terminal, neurotransmitter may be released into the synaptic cleft (5). The transmitter molecules diffuse across the cleft and some bind to receptor sites on the receiving neuron. (Adapted from Zigmond et al. 1999.)

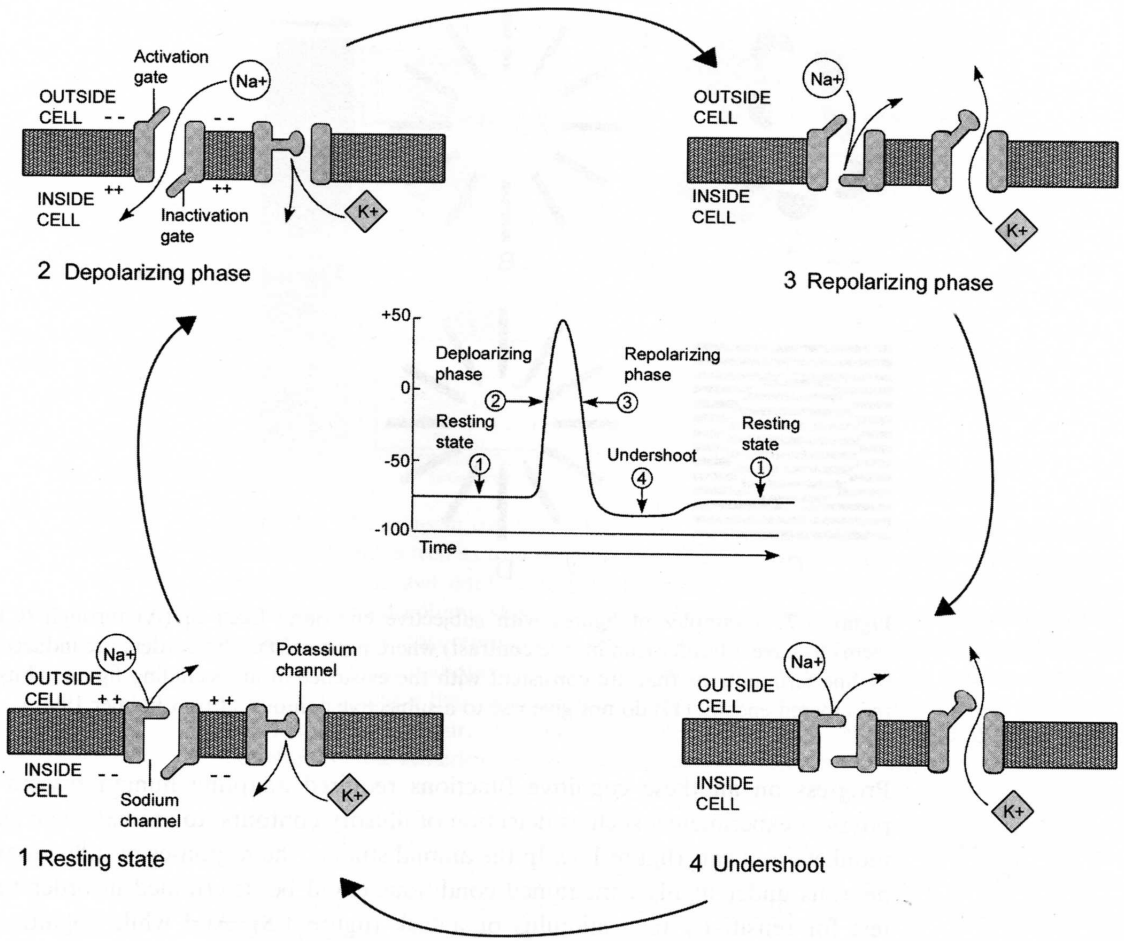

Figure 1.6 In the neuron's resting state (1), both the sodium (Na^+) and potassium (K^+) channels are closed, and the outside of the cell membrane is positively charged with respect to the inside. Hence there is a voltage drop across the membrane. If the membrane is depolarized (2), sodium ions enter the cell until the cell's polarity is reversed; that is, the inside of the cell is positively charged with respect to the outside. In the repolarization phase (3), the potassium channel then opens to allow eflux of potassium ions, the sodium gate closes, and sodium ions are actively pumped out of the cell. All of these activities help bring the membrane back to its resting potential. Because the potassium gate does not close as soon as the resting potential is reached (4), the voltage drop across the membrane briefly drops a little below the resting voltage. Equilibrium is reached once the resting potential is restored. (Based on Campbell 1996.)

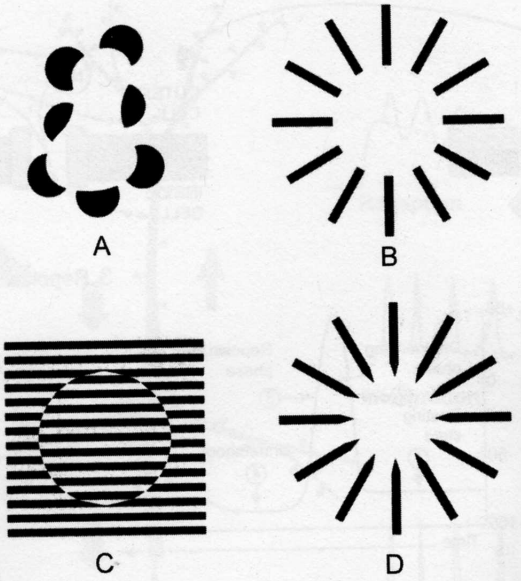

Figure 1.7 Examples of figures with subjective contours. Each of (A) through (C) seems to have a border (luminance contrast) where none exists. The borders are induced by line terminations that are consistent with the existence of an occluding figure. Thus the tapered ends in (D) do not give rise to a subjective contour. (From Palmer 1999.)

Progress on all these cognitive functions required adapting human psychophysical experiments, such as detection of illusory contours, to animals such as monkeys and cats (figure 1.7). In the animal studies, the responses of individual neurons under highly constrained conditions could be determined in order to test for sensitivity to a stimulus or a task (figure 1.8). And while cognitive functions at the network and neuronal level were being explored, details continued pour in to update the story of the ultrastructure of neurons—their synapses, dendrites, and gene expression within the nucleus—and how cognitive function was related to various ultrastructural operations.

Nevertheless, many fundamental questions about how the nervous system works remain wide open. In particular, bridging the gap between activity in individual neurons and activity in networks of neurons has been difficult. Macrolevel operations depend on the orchestrated activity of many neurons in a network, and presumably individual neurons make somewhat different contributions in order for the network to achieve a specific output, such as recog-

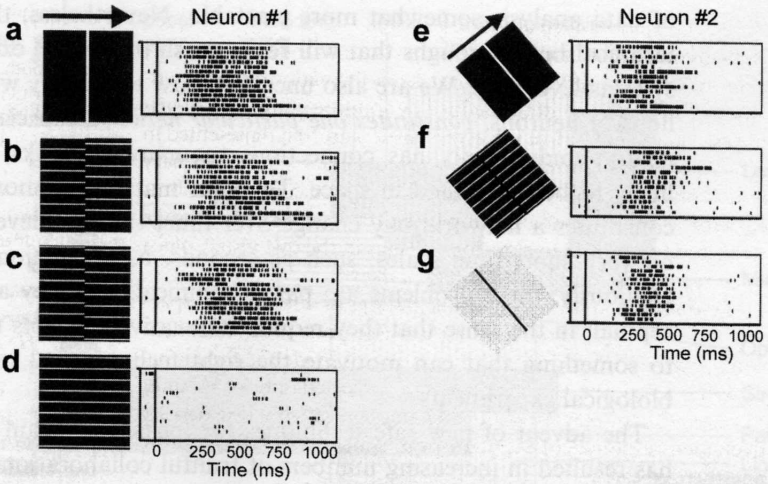

Figure 1.8 Neurons in owl visual forebrain areas (visual Wulst) respond to subjective contours about as well as to a real contour. The four contours (a) to (d) were randomly presented to the owl until each was viewed 15 times. The left column illustrates the stimuli; the right column shows the corresponding dot-raster displays for several presentations. Black dots represent the occurrences of spikes. Arrows indicate the direction of motion of the contours (motion onset at 0 ms). Notice that the neuron responds poorly in (d), where there is no subjective contour, but responds as well to (b) and (c) as to (a), the real contour. (Reprinted with permission from Nieder and Wagner 1999. Copyright by the American Association for the Advancement of Science.)

nition of visual motion or a command to move the eyes to a specific location. Moreover, understanding the dynamics of patterns of activity in neural networks and across many networks is undoubtedly essential to understanding how integration and coherence are achieved in brains. For example, there appear to be "competitions" between networks as the brain settles on a decision whether to fight or flee, and if to flee, whether to run in *this* direction or *that*, and so on. We are just beginning to feel our way toward concepts that might be helpful in thinking about the problems of coherencing.[11]

Until very recently, neuronal responses could be probed only one neuron at a time, but if we cannot access many neurons in a network, we have trouble figuring out how any given neuron contributes to various network functions, and hence we have trouble understanding exactly how networks operate. Significant technical progress has been made in recording simultaneously from more than one neuron, and the advent of powerful computers has made the problems

of data analysis somewhat more tractable. Nevertheless, the search is on for technical breakthroughs that will really mesh microlevel experimentation with systems-level data. We are also uncertain how to identify what, among the billions of neurons, *constitutes one particular network*, especially since any given neuron undoubtedly has connections to many networks, and networks are likely to be distributed in space. To make matters yet more interesting, what constitutes a network may change over time, through development, and even on very short time scales, such as seconds, as a function of task demands. Obviously, these problems are partly technical, but they are also partly conceptual, in the sense that they require innovative concepts to edge them closer to something that can motivate the right technological invention for neurobiological experiments.

The advent of new safe techniques for measuring brain activity in humans has resulted in increasing numbers of fruitful collaborations between cognitive scientists and neuroscientists. When the results of techniques such as functional magnetic resonance imaging (fMRI)[12] and positron emission tomography (PET)[13] converge with results from basic neurobiology, we move closer to an integrated mind-brain science (figure 1.9). These techniques can show something about the changes in regional levels of activity over time, and if set up carefully, the changes can track changes in cognitive functions. It is important to understand that none of the imaging techniques measure neuronal activity directly. They track changes in blood flow (hemodynamics). Because the evidence suggests that localized increases in blood flow are a measure of local increases in neuronal activity (more active neurons need more oxygen and more glucose), they are believed to be an indirect indication of changes in levels of activity in the local neuronal population. Note also that the recorded changes are insensitive to what individual neurons in a region are doing. The best spatial resolution of PET is about 5 mm, and in fMRI it is about 2 mm, though these resolutions may improve. Since one mm^3 of cortex contains about 100,000 neurons, the spatial resolution of these techniques does not get us very close to single-neuron activity.[14]

If the images from scanning techniques reflect changes across time, one conceptual problem concerns how to interpret the changes, and that means figuring out what should count as the baseline activity in any given test. Suppose that a subject is awake and alert, and is given a task, for example, visually imaging moving his hand. How do we characterize the state before he is to begin the task? We ask the subject to just rest. But his brain does not rest. His brain will be doing *lots* of things, including making eye movements, monitoring

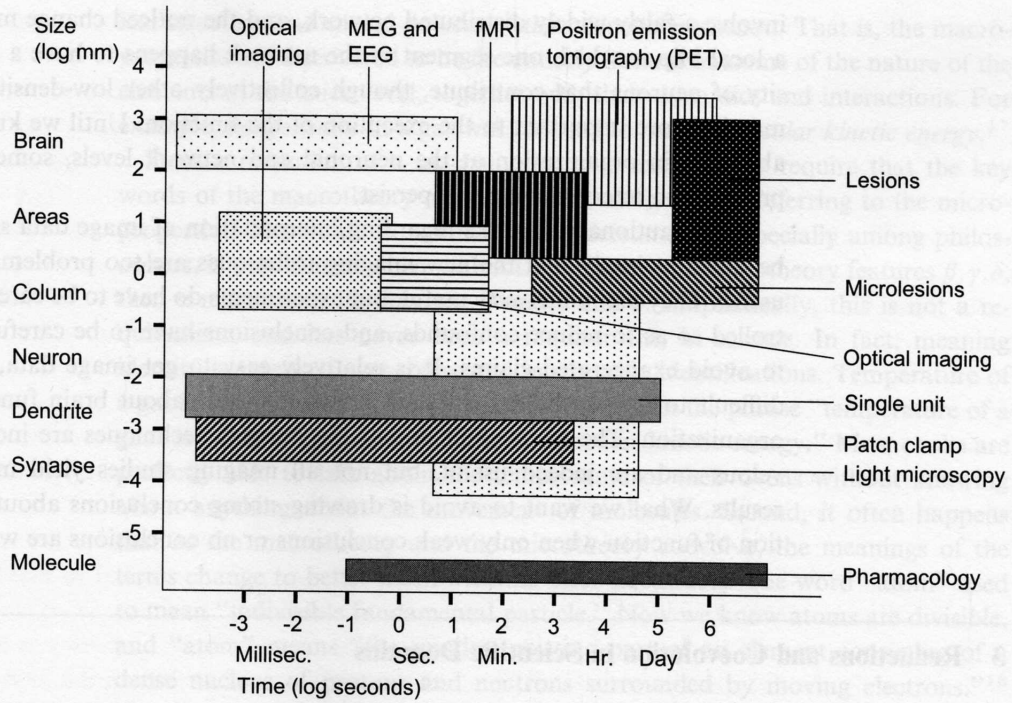

Figure 1.9 Comparison of the temporal and spatial resolutions of various brain-mapping techniques. MEG indicates magnetoencephalography; ERP, evoked response potential; EROS, event-related optical signal; MRI, magnetic resonance imaging; fMRI, functional MRI; PET, positron emission tomography; and 2-DG, 2-deoxyglucose. (Adapted from Churchland and Sejnowski 1988.)

glucose levels, perhaps thinking about missing breakfast, feeling an itch in his scalp, maintaining posture, and so forth. The subject cannot command the cessation of all cognitive functions, and certainly not all brain functions.

The problem of the baseline was recognized right from the beginning, and various strategies for reducing confounds have been developed, especially by Michael Posner and his colleagues.[15] These involve subtracting the level of activity in the "rest" condition from the level in the task condition, to reveal the difference made, presumably, by the task. There are other problems in getting meaningful interpretations of image data. For example, if a region shows increased activity during a cognitive task, does that mean it is *specialized* for that task? At most, it probably shows that the region has some role in executing the task, but this is a much weaker conclusion. Performance of the task may

involve a fairly widely distributed network, and the noticed change may reflect a local blip in which one segment of the network happens to have a high density of neurons that contribute, though collectively other low-density regions may be more important to the execution of the function. Until we know more about brain organization at the neuronal and network levels, some of these problems in interpretation will persist.

These cautionary remarks regarding interpretation of image data should *not* be taken to imply that the new imaging techniques are too problematic to be useful. They are in fact *very* useful, but experiments do have to be carefully controlled so as to reduce confounds, and conclusions have to be carefully stated to avoid exaggerated claims. It is relatively easy to get image data, but very difficult to know whether the data reveal anything about brain function and organization. The main point is that the imaging techniques are indeed marvelous and are indeed useful, but not all imaging studies yield meaningful results. What we want to avoid is drawing strong conclusions about localization of function when only weak conclusions or no conclusions are warranted.

3 Reductions and Coevolution in Scientific Domains

The possibility that mental phenomena might be understood in a neuroscientific framework is associated with *reductive explanation* in science generally. An example where one phenomenon is successfully reduced to another is the reduction of heat to molecular kinetic energy. In this case, the prereductive science was dealing with two sets of phenomena (i.e., heat and energy of motion), and had a good deal of observational knowledge about each. It was not initially obvious that heat had anything at all to do with motion, which seemed a wholly separate and unrelated phenomenon. As it turned out, however, they have quite a lot to do with each other, initial appearances notwithstanding.

An understanding of mental phenomena—such as memory, pains, dreaming, and reasoning—in terms of neurobiological phenomena is a *candidate* case of reduction, inasmuch as it looks reasonable to expect that they are brain functions. Because the word "reduction" can be used in wildly different ways, ranging from an honorific to a term of abuse, I now outline what I do and do not mean by "reduction."[16]

The baseline characterization of scientific reduction is tied to real examples in the history of science. Most simply, a reduction has been achieved when the causal powers of the macrophenomenon are explained as a function of the phys-

ical structure and causal powers of the microphenomenon. That is, the macroproperties are discovered to be the entirely natural outcome of the nature of the elements at the microlevel, together with their dynamics and interactions. For example, *temperature* in a gas was reduced to *mean molecular kinetic energy.*[17]

Does a reduction of a macrotheory to a microtheory require that the key words of the macrotheory *mean* the same as the words referring to the microproperties? Not at all. A common misunderstanding, especially among philosophers, is that if macrotheory about α is reduced to microtheory features β, γ, δ, then α must *mean the same* as β and γ and δ. Emphatically, this is not a requirement, and has never been a requirement, in science. In fact, meaning identity is rarely, if ever, preserved in scientific identifications. Temperature of a gas is *in fact* mean molecular kinetic energy, but the phrase "temperature of a gas" is not *synonymous* with "mean molecular kinetic energy." Most cooks are perfectly able to talk about the temperature of their ovens without knowing about anything about the movement of molecules. Second, it often happens that as the macrotheory and the microtheory coevolve, the meanings of the terms change to better mesh with the discovered facts. The word "atom" used to mean "indivisible fundamental particle." Now we know atoms are divisible, and "atom" means "the smallest existing part of an element consisting of a dense nucleus of protons and neutrons surrounded by moving electrons."[18] Usually, the meaning change is first adopted within the relevant scientific community and propagates more widely thereafter.

What does the history of science reveal about reductive explanations that might be helpful in understanding what a reduction of psychology to neuroscience will entail? A nagging question about the connection between cognition and the brain is this: can we ever get beyond mere *correlations* to actual identification and hence reduction? If so, how? Let us try to address this question by briefly discussing three cases. The first concerns the discovery that the identification of temperature of a gas with the mean kinetic energy of its constituent molecules permits thermal phenomena such as conduction, the relation of temperature and pressure, and the expansion of heated things to get a coherent, unified *explanation.* Correlations give you reasons for testing to evaluate the explanatory payoff from identification, but without *explanatory* dividends, correlations remain mere correlations. In the case of thermal phenomena, the first explanatory success with gases allowed the extension of the same explanatory framework to embrace liquids and solids, and eventually plasmas and even empty space. As a theory, statistical mechanics was far more successful than the caloric theory, the accepted theory of heat in the nineteenth century. Let us

look at little more closely at how people came to realize that temperature was actually molecular motion.

It is very natural to think of heat as a kind of stuff that moves from hot things to cold things. As natural philosophers investigated the nature of changes in temperature, they gave the name "caloric" to the stuff that presumably made hot things hot. Caloric was thought to be a genuine fluid—a fundamental stuff of the universe, along with atoms, and existing in the spaces between atoms. When Dalton (1766–1844) proposed his atomic theory, his sketches of tiny atoms showed them as surrounded by tiny atmospheres of caloric fluid. Within this framework, a hot cannonball was understood to have more caloric than a cold cannonball; snow has less caloric than steam.

Given that caloric is a kind of fluid, this entails that a thing should weigh more when hot than when cold. Weighing a cannon ball before and after heating tested this theory. The results showed that no matter how hot the cannon ball became, its weight remained the same. Faced with a possible refutation of a very plausible theory (what *else* could heat be?), some scientists were tempted by the hypothesis that caloric fluid was *very* special in that it had no mass.

Heat created through friction was also a puzzle, because there was no evident fluid *source* of caloric. The conventional wisdom settled on the idea that rubbing released the caloric fluid that was normally sequestered in the spaces between atoms. Rubbing jostled the atoms, and the jostling allowed the caloric to escape. To test the solution to the friction puzzle, Count Rumford Benjamin Thompson (1753–1814) traveled from England to a factory in Bavaria that bored holes in iron cannons. The boring, of course, continuously produced a huge amount of heat through friction, and the cannons under construction were constantly cooled by water. Rumford reasoned that if caloric fluid was released by friction during boring, then the caloric should eventually run out. No additional heat should be produced by further boring or rubbing. Needless to say, he observed that heat never ceased to be produced as the holes down the cannon shaft were continuously bored. At no point did the caloric fluid in the iron show the slightest sign of depletion.

Either there was an *infinite* amount of this allegedly massless fluid in the iron, or something was fundamentally wrong with the whole idea of caloric. Rumford realized that the first option was not seriously believable. Were it true, even one's hands would have to contain an infinite amount of caloric, since you can keep rubbing them without decline in heat production. Rumford concluded that not only was caloric fluid not a *fundamental* kind of stuff, it was not a stuff

of *any* kind. Heat required a different sort of explanation altogether. Heat, he proposed, just *is* micromechanical motion.[19]

Notice that a really determined calorist could persist in the face of Rumford's experiments, preferring to try to develop the option that every object really does contain an infinite amount of (massless) caloric fluid. And undoubtedly some believers did persist well after Rumford's presentation. The possibility of such persistence shows only that refutations of empirical theories are not as straightforward as refutations of mathematical conjectures. The caloric-fluid theory of heat was eventually rejected because its fit with other parts of science slowly became worse rather than better, *and* because, in the explanatory realm, it was vastly outclassed in explanatory and predictive power by the theory that heat is a matter of molecular motion. The fit of the newer theory with other parts of science, moreover, became *better* rather than worse. These developments also led to the distinction between heat (energy transfer as a result of difference in temperature) and temperature (movement of molecules).

The explanation of the nature of light can be seen as another successful example of scientific reduction. In this instance, visible light turned out to be electromagnetic radiation (EMR), as did radiant heat, x-rays, ultraviolet rays, radio waves, and so forth (see plate 1). Note also that in these examples, as in most others, further questions always remain to be answered, even after the reductive writing is on the wall. Hence, there is a sense in which the reduction is always incomplete. If the *core* mysteries are solved, however, that is usually sufficient for scientists to consider an explanation—and hence a reduction—to be well established and worthy of acceptance as the basis for further work.

Reductions can be very messy, in the sense that the mapping of properties from micro to macro can be *one-many* or even *many-many*, rather than the ideal *one-one*. While the case of light reducing to EMR is relatively clean, the case of phenotypic traits and genes is far less clean. Genes, as we now know, may not be single stretches of DNA, but may involve many distinct segments of DNA. The regulatory superstructure of noncoding DNA means that identification of a stretch of coding DNA as a "gene for ..." is a walloping simplification. Additionally, a given DNA segment may participate in different macroproperties as a function of such things as stage of development and extracellular milieu. Despite this complexity, molecular biologists typically see their explanatory framework as essentially reductive in character. This is mainly because a causal route from base-pair sequences in DNA to macrotraits, such as head/body segmentation, can be traced. The details, albeit messy,

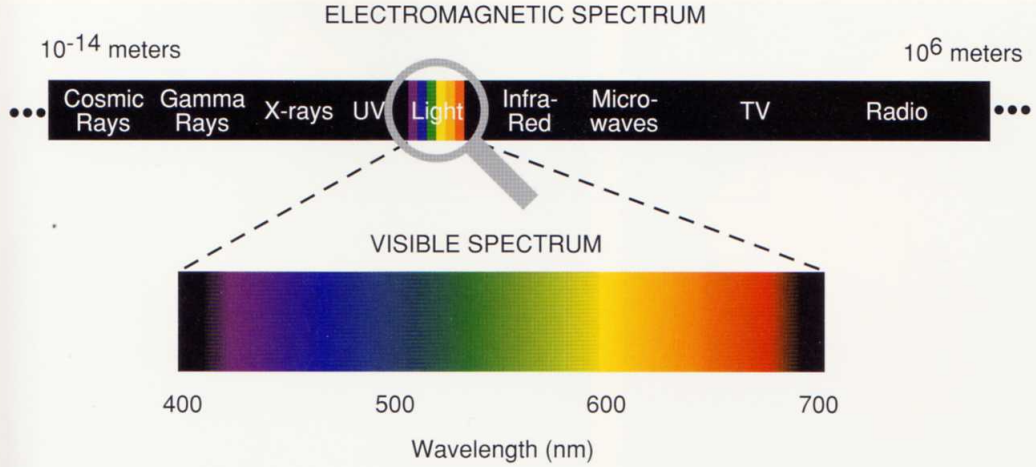

Plate 1 The electromagnetic spectrum. Radiant energy is characterized by its wavelength, which varies continuously from very small to very large. Visible light occupies the limited range from 400 to 700 nanometers (10^{-9} meters). It is the only form of electromagnetic radiation that people sense directly. (From Palmer 1999.)

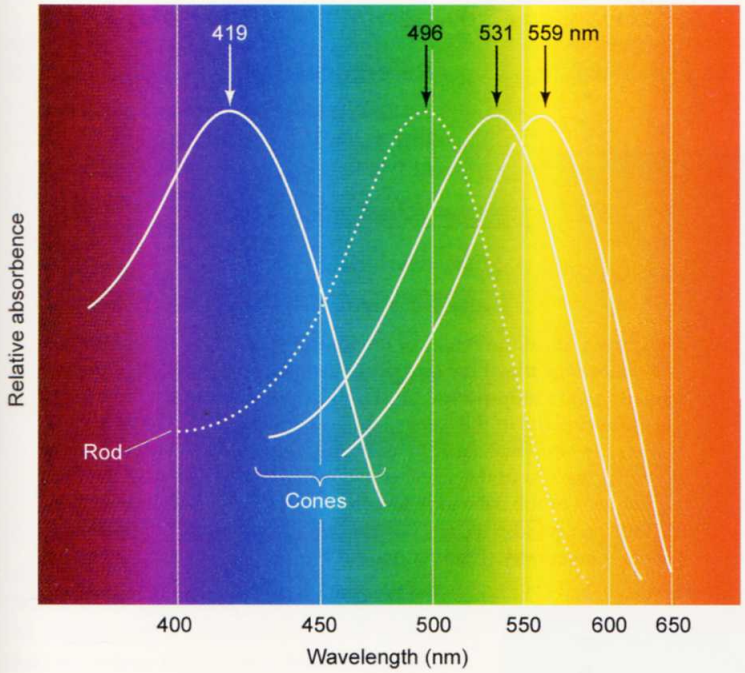

Plate 2 The absorption spectra of the four photopigments in the normal human retina. There are three types of cones, distinguished by three types of photopigments sensitive to light at distinct wavelengths. The sensitivity curve for rhodopsin, the photopigment in the rods, is also shown.

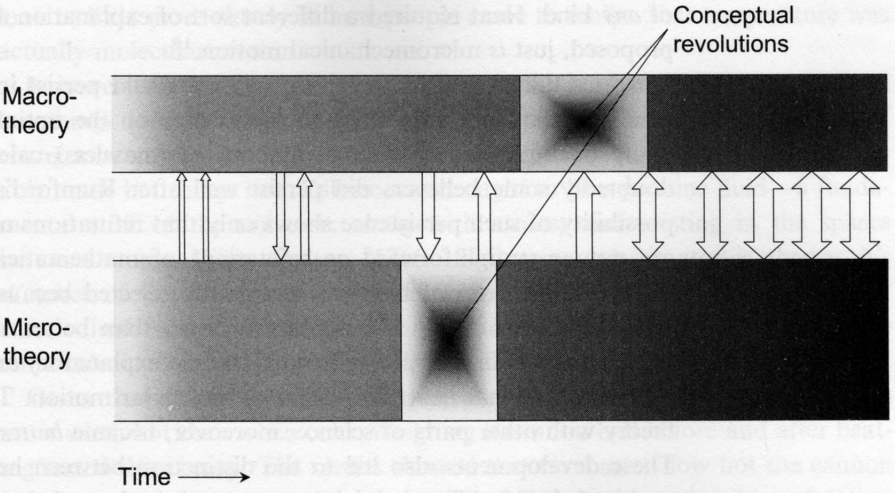

Figure 1.10 Macrolevel theories and microlevel theories coevolve through time. Initially, the connection between the macro- and microlevels may be tenuous and only suggestive, but their interactions may increase as experiments reveal correlations between macro- and microphenomena. As the experimental and theoretical interactions increase, the theories become increasingly interdigitated. The central concepts classifying macrophenomena and microphenomena are inevitably revised, and when the conceptual revision is very dramatic, this may be described in terms of a scientific revolution. Such revolutions are crudely indicated by a tunnel in the darkening pattern.

can be expected to fill out, at least in general terms, as experimental results come in.

This brings us to a second major point. Reductive explanations typically emerge in the later stages of a long and complicated courtship between higher-level and lower-level scientific domains. Earlier phases involve the *coevolution* of the scientific subfields, where each provides inspiration and experimental provocation for the cohort subfield, and where the results of each suggest modifications, revisions, and constraints for the other (figure 1.10). As theories coevolve, they gradually knit themselves into one another, as points of reductive contact are established and elaborated. Initially, contact between a high-level science and a lower-level science may be based merely on suggestive correlations in the occurrences of phenomena. Some such suggestive connections may prove to be genuine; some may turn out to be coincidental.

Reductive links begin to be forged when mechanisms at one level begin to explain and predict phenomena at another level. Not until there exist reason-

ably well-developed theories on both levels do the reductive explanations emerge. If you don't know beans about the macrolevel phenomenon of heat, you will not get very far trying to explain it in terms of some deeper and invisible property of matter. Sometimes the coevolution involves major revisions to the basic ideas defining the sciences, and the history of science reveals a wide spectrum of revisionary modifications. Caloric fluid, as we saw, got the boot as thermodynamics and statistical mechanics knit themselves together. Galileo and Newton rewrote the book on *momentum* and threw out the medieval conception of "impetus." Michael Faraday demonstrated, contrary to received opinion, that electricity is fundamentally the same phenomenon, whether it is produced by a battery, an electromagnetic generator, an electric eel, two hot metals brought into contact, or a hand rubbing against cat fur. In reality, the varieties of electrical phenomena are at bottom just one thing: electricity.

Reductive achievements sometimes fall short of the complete reduction of one theory to another because the available mathematics are insufficient to the task. Thus quantum mechanics has succeeded in explaining the macroproperties of the elements, such as the conductivity of copper or the melting point of lead, but not why a specific protein folds up precisely as it does. Whether more is forthcoming depends on developments in mathematics. In the case of quantum mechanisms, the mathematical limitations entail not that the macroproperties of complex molecules (e.g., serotonin) are emergent in some spooky sense, but only that we cannot now fully explain them.

It may come as a surprise that the great majority of philosophers working now are not reductionists, and are not remotely tempted by the hypothesis that understanding the brain is essential to understanding the mind. Such philosophers typically also see the details of neuroscience as *irrelevant* to understanding the nature of the mind.[20] The reason for their skepticism about the role of neuroscience is not rooted in substance dualism. Rather, the key idea is that the mind is analogous to software running on a computer. Like Adobe Photoshop, the cognitive program can be run on computers with very different hardware configurations. Consequently, although mind software can be run on the brain, it can also run on a device made of silicon chips or Jupiter goo. Hence, the argument goes, there is nothing much we can learn about cognition per se from looking at the brain.

Known as functionalism, this view asserts that the nature of a given type of cognitive operation is wholly a matter of the role it plays in the cognitive economy of the person.[21] Thus the draw operation of Adobe Photoshop is what it is solely and completely in virtue of its *role* in Adobe Photoshop. Its

nature, so to speak, is exhausted by the description of its interactions when Photoshop is running. Obviously, therefore, understanding the draw operation in Photoshop will not be helped by understanding the capacitors and transistors and circuits of one's computer.[22] Likewise, understanding what it is for a person to want a banana or believe that cows can fly will not be helped by understanding neurons, circuits, or anything else about how the brain works.[23]

Considerations of this sort motivated Jerry Fodor to emphasize the importance of experimental psychology, but also to firmly reject the relevance of neuroscience. He defends a thesis he calls the *autonomy of psychology*. This is a methodological claim. Its label embodies his conviction that psychology, as a science, is independent in its concepts and generalizations, of the concepts and generalizations of neuroscience. Briefly, the crux of the claim is that cognition cannot be explained in neurobiological terms and will not be usefully explored by neuroscientific techniques. The claim supports investigating cognition using *behavioral* measures, such as reaction times, and developing theories by constructing models that reflect the cognitive organization supposedly revealed by behavioral and introspective experiments. Neuroscientific data allegedly have a bearing only on how the cognitive program can be implemented in a particular physical arrangement, but have very little bearing on the actual nature of the cognitive functions. Neuroscience, from this perspective, may be of clinical interest, but it has no major significance for cognitive science.

There are many well-known criticisms of the autonomy-of-psychology thesis.[24] One powerful objection, repeatedly raised but never answered by those who live by the software analogy, is that the conceptual distinction between hardware and software does not correspond to any real distinction in nervous systems.[25] There are many levels of brain organization, ranging from protein channels in membranes, to neurons, microcircuits, macrocircuits, subsystems, and systems (see again figures 1.1 and 1.2). At many brain levels there are operations fairly describable as computations, and *none* of these levels can be singled out as *the* hardware level. For example, computations are performed by parts of dendrites, as well as by whole neurons, as well as by networks of neurons. Learning and memory, for example, involve computational operations at many levels of structural organization.[26] (This will be discussed in more detail in chapter 8.) The fact is, in nervous systems there are no levels of brain organization identifiable as *the* software level or *the* hardware level. Consequently, the linchpin analogy (mind/brain = software/hardware) is about as accurate as saying that the mind is like a fire or the mind is like a rich tapestry. In a poetic context, the metaphors are perhaps charming enough, but they are far too

unconnected to the real phenomena do very much to advance the scientific project of understanding cognition.

Another major concern is as practical as wearing boots in the snow. There is no point in turning your back on a vast range of data that might very well narrow your search space. To do so is perversely counterproductive. Keeping psychology pure from the taint of neuroscience seems strangely puritanical. Why not take advantage of every strategy, every technique, every well-controlled and well-run experiment? Why turn up your nose at some data when it might be useful?

Fodor, however, takes the software/hardware analogy to license assurance that neuroscientific data will not be useful. As noted, the analogy stipulates that neuroscientific data pertain to *implementation* rather than software. Unfortunately, and rather obviously, this response is untenable, because the analogy is untenable. By insisting that experimental psychology cut itself off from potentially useful neurobiological data, theory dualism is steering resolutely into the past instead of into the future. In a curious way, brain-averse functionalism is methodologically close to Cartesianism. In place of Descartes's *nonphysical mental substance*, functionalism substituted "*software*." Otherwise, things are much the same: no interest in or search for mechanisms of cognitive functions, no credence given to the possibility that we might learn fundamental facts about the mind by understanding how the brain works.

Notwithstanding the strictures of functionalism, the fact is that neuroscience and cognitive science are coevolving, like it or not. This coevolution is motivated not by ideology, but by the scientific and explanatory rewards derived from the interactions. Increasingly, this trend means that data from neuroscience are having an impact on how we frame questions about the mind and how we rethink how best to characterize psychological phenomena themselves. Examples of these developments will be seen in later chapters, and they will make us wonder whether some folk-psychological "verities" are as much in need of revision as were the "verities" of geocentrism. Exactly how the cognitive sciences and the neurosciences will knit into one another and how coevolution will change *both* is not easily predicted.

Though we can expect in a general way that mental phenomena will reduce to neurobiological phenomena, in the qualified sense of "reduction" used here, that achievement is certainly not yet in hand and could well be thwarted by the reality of the brain. For all we can be sure of now, a loose, if revealing, integration of domains may be the best we can achieve. Detailed explanatory mechanisms may elude us, and we might have to settle for general explanatory

principles that give us a story about mechanisms. Then again, maybe not. Science often surprises us with progress we thought impossible.[27]

There are still some very general worries about reduction to be addressed and allayed in advance of further progress, and I shall turn to three of those now.[28]

3.1 If We Get an Explanatory Reduction of Mental Life in Terms of Brain Activity, Should We Expect Our Mental Life to Go Away?

This worry is based on misinformation concerning what reductions in science do and do not entail. The short answer to the question, therefore, is "No." Pains will not cease to be real just because we understand the neurobiology of pain. That is, a reductive explanation of a macrophenomenon in terms of the dynamics of its microstructural features does not mean that the macrophenomenon is not real or is scientifically disreputable or is somehow explanatorily unworthy or redundant. Even after we achieved an explanation of light in terms of EMR, the classical theory of optics continues to be useful, even in discovering new things. Nobody thinks that light is not real, as result of Maxwell's explanatory equations. Rather, we think that we understand more about the real nature of light than we did before 1873. Light is real, no doubt about it. But we now see visible light as but one segment of a wider spectrum that includes x-rays, ultraviolet light, and radio waves (plate 1). We can now explain a whole lot at the macrolevel that we were unable to explain before, such as why light can be polarized and why light is refracted by a lens.

Sometimes, however, hitherto respectable properties and substances do turn out to be unreal. The caloric theory of heat, as we mentioned, did not survive the rigors of science, and caloric fluid thus turned out not to be real. As neuroscience proceeds, the fate of our current conception of consciousness, for example, will depend on the facts of the matter and the long-term integrity of current macrolevel concepts.[29]

3.2 Should We Expect a One-Step Integration of the Behavioral Domain with the Neuronal Domain?

Nervous systems appear to have many levels of organization, ranging in spatial scale from molecules such as serotonin, to dendritic spines, neurons, small networks, large networks, areas, and systems. Although it remains to be empirically determined what exactly are the functionally significant levels, it is

unlikely that explanations of macroeffects such as perceiving motion will be explained directly in terms of the lowest microlevel. More likely, high-level network effects will be the outcome of interacting subnetworks; subnetwork effects the outcome of participating neurons and their interconnections; neuron effects the outcome of protein channels, neuromodulators, and neurotransmitters; and so forth. One misconception about the integrationist strategy sees it as seeking a *direct* explanatory bridge between the highest level and lowest levels. This idea of "explanation in a single bound" does stretch credulity, and neuroscientists are not remotely tempted by it. My approach predicts that integrative *explanations* will proceed stepwise from highest to lowest, and that the *research* should proceed at all levels simultaneously.[30]

3.3 How Can You Have Any Self-Esteem If You Think You Are Just a Piece of Meat?

The first part of the answer is that brains are not *just* pieces of meat. The human brain is what makes humans capable of painting the Sistine Chapel, designing airplanes and transistors, skating, reading, and playing Chopin. It is a truly astonishing and magnificent kind of "wonder-tissue," as the philosopher Dennett jokingly puts it.[31] Whatever self-esteem justly derives from our accomplishments does so *because* of the brain, not in spite of it.

Second, if we thought of ourselves as glorious creatures before we knew that the brain is responsible, why not continue to feel so after the discovery? Why does the knowledge not make us more interesting and remarkable, rather than less so? We can be thrilled by the spectacle of a volcano erupting or a calf being born or a bone healing before we understand what volcanoes are and how reproduction and healing work. Being the creatures we are, however, commonly we are even more thrilled in the embrace of the knowledge about volcanoes and birth and bones. Understanding why we sleep and dream or how we distinguish so many smells makes us so much more glorious, rather than less so. At the same time, understanding why someone is demented or gripped by a hand-washing compulsion or tormented by a phantom arm after amputation helps replace superstition with sympathy and panic with calm reason.

Third, self-esteem, as we all know, depends on many complex factors, including things that happened or didn't happen during childhood and social recognition of a certain kind. None of this is altered one iota by realizing that one's feelings are caused by brain activity. When I step on a thorn, it still hurts in the same way, whether I know that the pain is really an activity in neurons

or not. When a teacher sincerely compliments a student's essay as insightful, well-researched, and clearly written, he esteems the student's accomplishment. In consequence, she is entitled to self-esteem, and it would be utterly irrelevant to add, "Too bad, though, this paper is just a product of your brain" as a deflationary remark.

4 Concluding Remarks

Three hypotheses underpin this book:

Hypothesis 1 Mental activity *is* brain activity. It is susceptible to scientific methods of investigation.

Hypothesis 2 Neuroscience needs cognitive science to know *what* phenomena need to be explained. To understand the scope of the capacity you want to explain—such as sleep, temperature discrimination, or skill learning—it is insufficient to simply rely on folk wisdom and introspection. Psychophysics, and experimental psychology generally, are necessary accurately to characterize the organism's behavioral repertoire and to discover the composition, scope, and limits of the various mental capacities.

Hypothesis 3 It is necessary to understand the brain, and to understand it at many levels of organization, in order to understand the nature of the mind.

Hypothesis 1 is a front-and-center topic of the entire book. It will be continually dissected, tested, and defended when we address the nature of the self, consciousness, free will, and knowledge. Ultimately, its soundness will be settled by what actually happens as the mind/brain sciences continue to make progress. Conceivably, it will turn out that thinking, feeling, and so on, are in fact carried out by nonphysical soul stuff. At this stage of science, however, the Cartesian outcome looks improbable. As noted earlier, hypothesis 3 is hotly contested by those psychologists and philosophers who favor the "mind as software" approach.[32] Hypothesis 2, on the other hand, though it may be embraced *in principle* by neuroscientists, is sometimes ignored in practice. For example, molecular-level neuroscientists may be apt to scoff at systems-level neuroscientists who are groping for ways to test psychophysical hypotheses in monkeys.

The more serious problem, however, is that brain-averse philosophers and psychologists tend to assume that those who believe hypothesis 3 (e.g., neuro-

scientists) are bound and determined to *dis*believe hypothesis 2.[33] *No such conclusion follows*, of course. The important point is that psychology and neuroscience are coevolving and will continue to do so. The fields are not mutually incompatible, but *mutually dependent*. Temporarily focusing on one level of organization is often a practical experimental expedient, but that is very different from making it a principle of research strategy.

One further observation concerns our ideas about ourselves, including our philosophical ideas. The main business of our brains is to help us adapt to changing circumstances, to predict food sources and dangers, to recognize mates and shelter, in general, to allow us to survive and reproduce. The human brain, as a rather fancy defense against variability and disaster, also generates stories—call them *theories*—to explain *why* things happen and thus help predict what *will* happen.

Some theories are better than others. The theory that bubonic plague is God's punishment is not as successful as the theory that it is a rat-borne bacterial infection. The first suggests prayer as a preventative, the second predicts that hand washing, rat killing, and water boiling will be more effective. As indeed they are. The theory that Zeus makes thunder by hurling luminous bolts is not as successful as the theory that lightning causes a sudden heating of adjacent air and therewith a sudden expansion. And so forth.

What about theories concerning ourselves—*our* natures? Our ideas about why people do certain things, and indeed why one does something oneself, are part of a wider network of story structures, with some cultural variability and some commonality. We explain and predict one another's behavior by relying on stories about attitudes, will power, beliefs, desires, superegos, egos, and selves. For example, we explain a certain basketball player's demands for attention in terms of his big ego; we may describe a backsliding smoker as lacking will power, an actor as moody or as obsessed with popularity or as having a narcissistic personality disorder, and so on. Freud (1856–1939) urged us to explain compulsive behavior in terms of superego dysfunction. But what, in neurobiological terms, *are* these states—will power, moods, personality, ego, and superego? Are some of these categories like the categories of now-defunct but hitherto "obvious" Aristotelian physics, categories such as "impetus" and "natural place"?

Given scientific progress in general, along with specific evidence about the brain and how it works, our shared conventional story structures may come to be modified where they prove less successful than experimentally tested theories. The details of theory modifications are essentially impossible to predict in advance. Already, however, we can see some story modification.

In the last fifty years, we have come to realize that epilepsy is best understood in neurobiological terms, not in terms of the divine touch. Hysterical paralysis is not a dysfunction of the uterus, but of the brain. In subjects' who are compulsive handwashers, possession by spirits or superego dysfunction explains and predicts far less than neuromodulator levels. The discovery that highly addictable subjects have a gene implicated in the quirks of their dopamine reward system begins to hint that we will want to reconsider what exactly having or lacking will power comes to. None of this is surprising, for what the history of science reveals is that *some* theory revision is typical and pretty much inevitable, no matter what the domain of inquiry—astronomy, physics, biology, or the nature of our minds. That the story structure giving shape to traditional philosophical inquiry may itself evolve, perhaps quite profoundly, accordingly presents an even deeper challenge to those who wish to isolate philosophy from science.

The overarching theme of this book is that if we allow discoveries in neuroscience and cognitive science to butt up against old philosophical problems, something very remarkable happens. We will see genuine progress where progress was deemed impossible; we will see intuitions surprised and dogmas routed. We will find ourselves making sense of mental phenomena in neurobiological terms, while unmasking some classical puzzles as preneuroscientific misconceptions. Neuroscience has only just begun to have an impact on philosophical problems. In the next decades, as neurobiological techniques are invented and theories of brain function elaborated, the paradigmatic forms for understanding mind-brain phenomena will shift, and shift again. These are still early days for neuroscience. Unlike physics or molecular biology, neuroscience does not yet have a firm grasp of the basic principles explaining its target phenomena. The real conceptual revolution will be upon us once those principles come into focus. How things will look *then* is anybody's guess.

Selected Readings

Basic Introductions

Allman, J. M. 1999. *Evolving Brains*. New York: Scientific American Library.

Bechtel, W., and G. Graham, eds. 1998. *A Companion to Cognitive Science*. Oxford: Blackwells.

Osherson, D., ed. 1990. *Invitation to Cognitive Science*. Vols. 1–3. Cambridge: MIT Press.

Palmer, S. E. 1999. *Vision Science: Photons to Phenomenology*. Cambridge: MIT Press.

Sekuler, R., and R. Blake. 1994. *Perception*. 3rd ed. New York: McGraw Hill.

Wilson, R. A., and F. Keil, eds. 1999. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge: MIT Press.

Zigmond, M. J., F. E. Bloom, S. C. Landis, J. L. Roberts, L. R. Squire. 1999. *Fundamental Neuroscience*. San Diego: Academic Press.

Additional Selected Readings

Bechtel, W., P. Mandik, J. Mundale, and R. S. Stufflebeam, eds. 2001. *Philosophy and the Neurosciences: A Reader*. Oxford: Oxford University Press.

Bechtel, W., and R. C. Richardson. 1993. *Discovering Complexity*. Princeton: Princeton University Press.

Churchland, P. M. 1988. *Matter and Consciousness*. 2nd ed. Cambridge: MIT Press.

Churchland, P. S. 1986. *Neurophilosophy: Towards a Unified Understanding of the Mind-Brain*. Cambridge: MIT Press.

Crick, F. 1994. *The Astonishing Hypothesis*. New York: Scribners.

Damasio, A. R. 1994. *Descartes' Error*. New York: Grossett/Putnam.

Kandel, E. R., J. H. Schwartz, T. M. Jessell, eds. 2000. *Principles of Neural Science*. 4th ed. New York: McGraw-Hill.

Moser, P. K., and J. D. Trout, eds. 1995. *Contemporary Materialism: A Reader*. London: Routledge.

History

Brazier, M. A. B. 1984. *A History of Neurophysiology in the 17th and 18th Centuries: From Concept to Experiment*. New York: Raven Press.

Finger, S. 1994. *Origins of Neuroscience: A History of Explorations into Brain Function*. New York: Oxford University Press.

Gross, C. G. 1999. *Brain, Vision, Memory: Tales in the History of Neuroscience*. Cambridge: MIT Press.

Young, R. M. 1970. *Mind, Brain, and Adaptation in the Nineteenth Century*. New York: Oxford University Press.

Journals with Review Articles

Annals of Neurology

Cognition

Current Issues in Biology

Nature Reviews: Neuroscience

Psychological Bulletin

Trends in Cognitive Sciences

Trends in Neurosciences

Websites

BioMedNet Magazine: http://news.bmn.com/magazine

Encyclopedia of Life Sciences: http://www.els.net

The MIT Encyclopedia of the Cognitive Sciences: http://cognet.mit.edu/MITECS

Neuroanatomy: http://thalamus.wustl.edu/course

Science: http://scienceonline.org

The Whole Brain Atlas: http://www.med.harvard.edu/AANLIB/home.html

Patricia Smith Churchland

Brain-Wise

Studies in Neurophilosophy

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong, and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Churchland, Patricia Smith.
Brain-Wise : studies in neurophilosophy / Patricia Smith Churchland.
p. cm.
Includes bibliographical references and index.
ISBN 0-262-03301-1 (hc : alk. paper) — ISBN 0-262-53200-X (pbk : alk. paper)
1. Neurosciences—Philosophy. 2. Cognitive science—Philosophy. I. Title: Studies in neurophilosophy. II. Title.
[DNLM: 1. Neuropsychology. 2. Knowledge. Metaphysics. 4. Neurology. 5. Philosophy. 6. Religion and Psychology. WL 103.5 C563b 2002]
RC343 .C486 2002
153′.01—dc21 2002066024

10 9 8 7 6 5 4 3 2 1