Chapter 13. How brains seem to work

Brains: How They Seem to Work, ISBN: 9780132139991
To go back to the beginning, the foundation for thinking about the nervous system when I was a first-year medical student in 1960 was primarily knowledge about nerve cells and neural signaling based on work that Hodgkin, Huxley, Kuffler, Katz, and their collaborators had carried out during the preceding decades. This seminal body of research had led to a good, if incomplete, understanding of how action potentials carry information in the nervous system and how chemical synaptic transmission conveys this information from one neuron to another. Although a wealth of important detail about these processes has been added since, this basic understanding of neural function remains much the same as when I first learned it. The other body of knowledge for my generation was the welter of facts about the anatomy of the brain and the rest of the nervous system that grew concurrently. But in contrast to the clarity of what was known about neural signaling, the functional significance of brain structures and their complex interconnections was—and still is—poorly understood. The problem, both in the 1960s and today, is how brains use neural signaling to accomplish all the remarkable functions we take for granted.

It seemed obvious when I was starting out that the goal of neuroscience in the coming decades would be to put meat on the bones of the generally vague notions about what sensory, motor, and other brain systems are actually doing. As it turned out, the most influential pioneers in carrying this program forward were David Hubel and Torsten Wiesel, whose work in the visual system was inspired by Kuffler's study of the retina in the early 1950s, and by Vernon Mountcastle's description of columns in the somatic sensory cortex (see Chapter 7). By the 1970s, recording neural signals from the brains of
anesthetized animals while activating a system of interest with increasingly complex stimuli had become a staple of work on sensory and motor systems. In addition to Hubel and Wiesel’s work on vision, each of these efforts—mainly in the somatic sensory and auditory systems—soon had its own history and major players. Other researchers used the same general approach to explore the motor system, building on Sherrington’s work earlier in the twentieth century, to ask how the motor regions of the brain are organized to produce actions. Beginning in the 1980s, understanding brain systems was further advanced by the introduction of techniques for simultaneously recording from dozens of electrodes implanted in different brain regions, and by increasingly powerful noninvasive brain imaging techniques that reveal brain activity in human subjects carrying out cognitive and other tasks. The latter approach, in particular, provided a much stronger link between the functional architecture of the brain and the inchoate anatomy that as students we were taught by rote.

Each of these ways of investigating the brain has helped close the gap between the reasonably complete understanding of neural signaling and the much less perfect grasp of how the brain uses these basic mechanisms to achieve its biological goals. But if one were to ask today what all this information indicates about how brains accomplish what they do, most neuroscientists would probably agree that the answer is not all that much clearer than it was in 1960. Why, then, despite the enormous increase in information about the properties of neurons in different regions, their connectivity, their transmitter pharmacology, and the behavioral situations in which they become active, do we remain relatively ignorant about perception, cognition, consciousness, and other brain functions that interest us most?

Part of the reason is the absence of some guiding principle or principles that would help to understand the neural underpinnings of perceptual, behavioral, and cognitive phenomenology in a more general way. One doesn’t have to be steeped in history to recognize that other domains of biological science have typically advanced under the banner of some overarching framework. Obvious examples are the cell theory as a principle for understanding the structure and function of organs that emerged in the seventeenth century, the theory of evolution by natural selection as a principle for understanding species diversity that emerged in the nineteenth century, and the
theory of genes as a principle for understanding inherited traits that emerged in the late nineteenth and early twentieth centuries. However, in the case of the brain, no overarching idea has arisen that serves in this way. Given the complexity of the brain and the rest of the nervous system, this absence is perhaps to be expected. But if the history of science is any guide, sooner or later a principle of this sort is bound to emerge.

Only a few ideas about the overall operation of the brain have been seriously considered in recent decades, and all have relied heavily on studies of vision and the visual brain. The one that has come up most often in earlier chapters is that the brain extracts features from sensory stimuli and combines them in representations of the world in higher-order cortical areas that we then perceive and use to inform behavior and cognition. This perspective takes the brain to be directly related to the real world by logical operations carried out by neurons, neural circuits, and brain systems. The explicit or implicit advocacy of this essentially rationalist interpretation of brains has focused on both the biology of neural processing and the logical instructions (computational algorithms) that might be doing the job. A more stochastic twist on this perspective is that because stimuli are inevitably noisy and their sources are uncertain, the brain must to a degree operate probabilistically, generating perceptions and behaviors predicated on inferences (best guesses) about the most likely state of the world. These frameworks for thinking about how brains work express in different ways the seemingly obvious idea that sensory information is processed by online neural computations that represent the “real world” somewhere in the brain as a basis for all the higher-order functions that the brain performs and the behaviors that follow.

I fully accepted this general concept of brain function gleaned from my professors and colleagues in the 1960s. This framework taps into the subjective sense we all have about our relationship to the world, and for the next 20 years or so, I had no particular reason to doubt its propriety. Indeed, it seemed clear that exploring the properties of the nerve cells and neural circuits from the bottom up would eventually reveal the underlying logic of vision and the other senses, and indicate how they were related to perception and, ultimately, action. This was the implicit message of Hubel and Wiesel’s effort to understand vision in terms of an anatomical and functional hierarchy in
which simple cells feed onto complex cells, complex cells feed onto hypercomplex cells, and so on up to the higher reaches of the extrastriate cortex. (As noted earlier, Hubel and Wiesel were wary of anything that smacked of theoretical speculation and wisely stuck close to their results, leaving others to argue about what they might mean for the function of the brain more generally.) Nearly everyone believed that the activity of neurons with specific receptive field properties would, at some level of the visual system, represent the combined image features of a stimulus, thereby accounting for what we see. This idea seems so straightforward that it was, and is, hard to imagine an alternative.

Nevertheless, this conception of how brains work has not been substantiated despite an effort that now spans 50 years. When a path in science is pursued for this long without the emergence of a deeper understanding of the issue being addressed, doubts are usually warranted. Since the 1980s, it has become increasingly apparent that neuronal responses to even the simplest visual stimuli are difficult to rationalize in terms of a hierarchy that begins with the detection of image features at the retinal level and ends with feature representation in the cerebral cortex. Another problem has been the growing realization that the activity of visual neurons is not just a result of whatever is going on within the small region of visual space that defines the “classical” receptive field of a visual neuron; stimuli that fall outside this region also affect the activity of neurons in ways that have been difficult to understand. This less direct source of activation and its complexity is not surprising, given that the vast majority of the inputs to neurons in the various stations of the visual system come not from the retina, but from other neurons at the same or higher levels (such as feedback to the thalamus from the primary visual cortex, and feedback to the primary visual cortex from extrastriate cortical regions). Yet another concern is the observation that different stimuli can elicit the same general pattern of visual cortical activity. If the evidence in previous chapters that the stimuli can be perceived in radically different ways as a function of context is added to this mix, one is left with deep uncertainty about how the activity of sensory neurons produces perceptions, and how perception is then related to action.

One way to deal with these concerns was to take a frankly computational approach to the brain, a view that was gaining popularity during the same time that visual physiologists were grappling with these
these problems. The idea was that if brains operate as computers do, then understanding brain functions might advance rapidly by riding on the back of computational theory. The potential of computers and computational theory had gotten off to a rough start with the grandiose and ultimately failed plans of Charles Babbage in the nineteenth century (Figure 13.1), but the enormous progress of computation during and after World War II encouraged its application to understanding brain functions, vision in particular. The most influential exponent of visual processing as computations that could be understood in logical terms was David Marr. A brilliant theorist, Marr summarized his ideas in a book published in 1982 that many regard as having given birth to the field of computer vision. Computer (or “machine”) vision is the attractive notion that one should be able to build and program a device that sees in much the same way we do, based on rational engineering principles. For Marr (and many others since), vision is a computational task that, similar to the framework underlying the physiological approach that was then in ascendancy, leads to a representation of stimulus features in the brain. Marr (who died of cancer at age 35 just before his book was published) integrated into this perspective much of what was then known about visual perception, physiology, and anatomy. He was motivated by his conviction that vision scientists were describing the physiological and anatomical underpinnings of vision but were missing the boat by not getting straight to the logical heart of the matter.

To address this deficiency, Marr proposed that the information in retinal images is subject to algorithmic processing at three levels, which he referred to as the construction of the primal sketch, the $2\frac{1}{2}$-D sketch, and the 3-D model representation. The purpose of the primal sketch is to make information about the elemental geometries and intensities of the retinal image explicit (to represent basic image features), the purpose of the $2\frac{1}{2}$-D sketch is to make explicit the orientation and approximate depth of surfaces in a viewer-centered frame of reference, and the purpose of the 3-D model is to represent shapes and their spatial organization in an object-centered frame of reference, representing what observers see. For Marr, this overall process of internal representation was a “formal system for making explicit certain entities or types of information, together with a specification of how the system does this.” His attempt to rationalize vision in a
Figure 13.1 Babbage’s “Difference Engine,” a mechanical computer designed in 1821. Although this and Babbage’s later designs were never built to completion during his lifetime (his machinist eventually quit and the British government tired of supporting his increasingly expensive project), Babbage’s machines are widely considered the first examples of serial-processing computers, the computers on which most contemporary machine-vision strategies are implemented. Because its mechanisms are visible, Babbage’s machines (some of which are on display at the British Science Museum in London) give a more vivid impression of what logical computation entails than looking at an integrated circuit. (Courtesy of Eric Foxley)

comprehensive computational theory was highly ambitious and widely admired. Although much of this initial enthusiasm has faded, determining how the features of the retinal image are detected, processed, and represented according to a series of algorithms remains a central theme in vision science, particularly in the ongoing attempts of computer scientists, biomedical engineers, and the military to create artificial visual systems that could help patients or enable robotic devices to navigate more efficiently in real-world environments.
Wherever Marr might have taken these ideas had he lived, it seems unlikely that the brain operates like the serial-processing computers we are all familiar with today—a machine that uses a series of specified, logical steps to solve a problem (a so-called universal Turing machine). Of course, algorithmic computations based on physical information acquired by photometers, laser range scanners, or other devices that directly measure aspects of the world can solve some of the problems that confront biological vision. And the “visually guided” behavior of robotic vehicles today is impressive. But these and other automata are “seeing” in a fundamentally different way than we do. The limitation of machine vision in this form is its inability to meet the challenge that has evidently driven the evolution of biological vision: The unknowability of the physical world by any direct operation on images (the inverse problem). Machines such as photometers and laser range finders accurately determine some physical property of the world (such as luminance or distance) by direct measurement. But as should be apparent from previous chapters, this is not an option for biological vision, nor for machine vision if it is ever to attain the sort of visual competence that we enjoy. Only by evolving circuitry that reflects the outcome of trial-and-error experience with all the variables that affect the successful behavior is a machine likely to generate “perceptions” and “visually guided” behavior that works well in real-world circumstances.

Given this caveat, it is important to recognize that computers can solve complex problems in another way, an alternative that gives cause for some optimism about the future of machine vision and ultimately understanding what the complex connectivity of the brain is accomplishing. In a paper published in 1943, MIT psychologist Warren McCulloch and logician Walter Pitts pointed out that instead of depending on a series of predetermined steps that dictate each sequential operation of a computer in logical terms, problems can also be solved by devices that comprise a network of units (neurons, in their biologically inspired terminology) whose interconnections change progressively according to feedback arising from the success (or failure) of the network dealing with the problem at hand (Figure 13.2). The key attribute of such systems—which quickly came to be called artificial neural networks (or just neural nets)—is the ability to solve a problem without previous knowledge of the answer, the steps needed to reach it, or the designer’s conception of how the problem might be...
solved in rational terms. In effect, neural nets reach solutions by trial and error, gradually generating more useful responses by retaining the connectivity that led to improved behavior. As a result, the architecture of the trained network—analogous to evolved brain circuitry—is entirely a result of the network’s experience.

![Diagram of a simple neural network](image)

**Figure 13.2** Diagram of a simple neural network comprising an input layer, an output layer, and a hidden layer. The common denominator of this and more complex artificial neural networks is a richly interconnected system of nodes, or neurons. The strengths of the initially random connections between the nodes are progressively changed according to the relative success of trial-and-error responses, which is then fed back to the network during training. The result is that the connectivity of the system is gradually changed as the network deals ever more effectively whatever problem it has been given. (After Purves, Brannon, et al., 2008)

The answer to the question of whether the brain operates like a computer therefore depends on what kind of a computer one has in mind. Unlike the rationalist idea of computation based on feature extraction and representation according to a series of logical steps, the conception of the brain as a computer is plausible considered as a neural network whose connectivity changes over evolutionary and individual time according to feedback from trial and error. This perspective
can be made even more biologically attractive by letting a computational equivalent of natural selection operate on an evolving population of neural networks. By changing the weights of the connections between nodes according to the relative success of autonomous networks competing with each other in a virtual environment and reproducing differentially based on some criteria of “fitness,” it should be possible to explore how the structure and function of real brains evolved.

By the mid-1990s, mathematically inclined psychologists joined the fray with yet another perspective about how vision, and brains more generally, might work. Their approach was based on a statistical methodology generally referred to as Bayesian decision theory. Thomas Bayes was an English mathematician and Presbyterian minister who published a paper in 1763 showing formally how conditional probabilities lead to valid inferences. Although Bayes’s motivation for studying this issue is unclear, his theorem has been widely applied to statistical problems whose solution depends on an assessment of hypotheses that are only more or less likely to be true because they depend on two or more probabilities. For example, the theorem has been used in medicine to evaluate the likelihood of a diagnosis or a clinical outcome given a set of tests or other data that together contribute to the overall probability of having the disease or responding to therapy. Bayes’s theorem is usually written in this form:

\[
P(H|E) = \frac{P(H) \times P(E|H)}{P(E)}
\]

H is a hypothesis to be tested, E is the evidence pertinent to its validity, and P is the probability. The first term on the right side of Bayes’s equation, P(H), is called the prior probability distribution, or simply “the prior,” and is a statistical measure of confidence in the hypothesis, without any present evidence about its truth or falsity. With respect to vision, the prior describes the probabilities of different physical states of the world that might have produced a retinal image. Therefore, the needed priors are the frequency of occurrence in the world of surface reflectance values, illuminants, distances, object sizes, and so on. The second term, P(E|H), is called the likelihood function. If hypothesis H was true, this term indicates the probability that the evidence E would have been available to support it. Given a
particular state of the physical world (a particular combination of illumination, surface properties, object sizes, etc.), the likelihood function describes the probability of that state having generated the retinal projection in question. The product of the prior and the likelihood function, divided by the normalization constant \( P(E) \), gives the posterior probability distribution, \( P(H|E) \). The posterior distribution defines the probability of hypothesis \( H \) being true, given the evidence \( E \), and would therefore indicate the relative probability of a given retinal image having been generated by some set of possible physical realities. The most likely physical reality in the posterior probability distribution is taken to be what an observer sees.

Despite the difficulty of these more abstract ideas, Bayes’s theorem correctly (and usefully) spells out the logical relationship among the factors that underlie any empirical approach to vision. Nonetheless, the way it has been used in vision science has tended to obscure instead of clarify the way brains seem to work. For most people using Bayesian decision theory, the visual system is taken to compute statistical inferences about possible physical states of the world based on prior knowledge acquired through experience with image features. The evidence in the preceding chapters, however, argues that the accumulated knowledge stored in brain circuitry is based on feedback from behavioral success, not image features. As a result, what we see is not the state of the world that most likely produced the image. On the contrary, the qualities we see define a subjective universe that comprises the various perceptual spaces described earlier. These perceptions successfully guide behavior not because they represent likely states of the world but because of the trial-and-error manner in which the relevant perceptual spaces and the underlying neural circuitry have been generated.

Although the human brain is enormously complex, the evidence based on what we see and why implies that the brain and the rest of our nervous systems are doing just one basic thing: linking sensory information (perceptions or the unconscious equivalent) to successful behavior by means of synaptic connectivity that has been entirely determined by trial and error. The reason for this blunt assertion is that the inverse problem in vision and other sensory modalities doesn’t allow much choice.
The biological mechanisms that create neural circuitry on this basis are straightforward, at least in principle. Whenever the neural connections inherited by an individual link perception and action a little more successfully, the slightly greater reproductive fitness of the individual tends to increase the prevalence of that brain circuitry in the population. Over eons, the neuronal associations underlying successful behavior in response to stimuli will therefore increase and unsuccessful associations will decrease, leading to the brain circuitry humans and other animals have today. All that is needed to implement this strategy is plenty of time and a way of providing feedback about the relative success of behavior. The history of life on Earth has provided ample time (approximately 3.5 billion years), and natural selection is a wonderfully powerful mechanism for assigning credit to behavior. Indeed, our perceptions and actions have become so effective in dealing with our circumstances world that it is difficult to imagine that what we see, hear, or otherwise perceive is only an operationally useful surrogate for the world, not physical reality.

The critical importance of this inherited neural infrastructure for survival is self-evident. Nonetheless, the neural connectivity we are born with is continually refined by individual experience (see Chapters 2–7). The reasons for this are also clear. From a developmental perspective, neural circuits must adjust to the changing body, as well as reaping the obvious benefits that come from encoding additional information about the specific objects, conditions, and contingencies encountered in any particular life. Although the mechanisms of synaptic plasticity that enable lifetime learning are quite different from the mechanisms of inheritance, both serve to encode experience in neural circuitry. The common denominator is the association of input from the external and internal environment with empirically successful behavioral output. From this perspective, the complexity of brain function and structure boils down to using, making, and modifying neuronal connections. The best evidence so far that brains really do work in this wholly empirical way is that so many peculiarities of subjective experience can be explained in this way. For each of the basic visual qualities—lightness, brightness, color, geometrical form, and motion—a variety of otherwise puzzling phenomena can be predicted by the frequency of occurrence of the stimulus characteristics derived from databases that serve as proxies for human
experience. The universal discrepancies between perception and physical measurement (the more flagrant examples of which are wrongly categorized as “illusions”) are neither anomalies nor the result of the limitations in neural processing circuitry. They are the signatures of this strategy of brain operation. Whereas the end result gives us a compelling sense that our brains must be detecting features by analyzing stimuli and producing perceptions that represent the world, terms such as feature detection, feature analysis, and feature representation are not appropriate descriptors of what the machinery of the brain is doing, the problem this mode of operation is solving, or what we actually perceive.

Accounting in these terms for everything we see or otherwise perceive, however, is going to be much harder than suggested by the ability to predict the relatively simple visual responses used so far to support the empirical case. Even in the most basic circumstances, what we see with respect to any particular visual quality is, for good reason, affected by the sensory information pertinent to other qualities (see Chapter 9). To complicate matters further, such interactions extend across different sensory modalities for the same reasons. One of many examples of how what we hear affects what we see is illustrated in Figure 13.3. This further evidence implies that percepts and behaviors ultimately depend on neural processing that is going on at the same time in many brain systems, taking into account the influence of all the neural activity elicited by the stimuli acting on us at any given time. The advantages accruing from such interactions help explain why the connectivity of the brain is so complex. Any brain that did not take into account the full range of information available to it to determine behavior would not be doing its job. If one adds the influences of memory, motivation, and emotional state to sensory input, what we perceive and do at any given moment is determined by the processing going on in much, if not all, of the brain.

This understanding of perception and sensory systems generally implies yet another difficult conclusion about how brains work: Perceptions, and the behaviors they lead to, are entirely reflexive. Although the concept of a reflex is not precisely defined in neuroscience, the term is typically used to refer to involuntary sensory-motor behaviors that are assumed to occur with a minimum of the higher-order cortical processing thought to underlie voluntary
actions. The usual example presented in textbooks is one of the spinal reflexes that Sherrington studied in the early twentieth century (Figure 13.4), or an autonomic reflex such as salivation in response to food that Ivan Pavlov used at about the same time to study conditioned learning. For Sherrington and Pavlov, a reflex meant an automatic response that depends on a relatively simple neural pathway with little or no cognitive input. However, Sherrington was well aware that the idea of a simple reflex isolated from the rest of the nervous system is, in his words, a “convenient...fiction.” Sherrington went on to emphasize that “all parts of the nervous system are connected together and no part of it is ever capable of reaction without affecting and being affected by other parts”. Clinicians know very well that spinal reflexes depend on cortical processing and that they no longer operate normally when the descending cortical input to spinal neurons is compromised by disease or injury.

Despite Sherrington’s caveat, reflexes are usually regarded as behaviors that are immune from the influence of cortical processing, particularly from “top-down” cognitive influences. But the reality is that brain systems and circuits are invariably interconnected to take full advantage of the range of information the nervous system provides at any given moment. From this vantage, the conventional distinction between involuntary (reflexive) and voluntary neural processing makes little sense. Any perception or behavior is equally a product of all the neural processing occurring in the nervous system.
at that time. Depending on the circumstances, some parts of the nervous system will be more—or less—active than others. But it would be a mistake to think that processing activity in any part of the brain is irrelevant to processing in another region, or to perceptual and behavioral outcome. A subjective sense of what our brains are doing has given pride of place to what we assume to be “voluntary” behaviors or thought processes, but no neurobiological evidence supports this bias.

The reflexive nature of perception and behavior points to a still deeper problem at the center of a debate that has persisted since the Stoic and Epicurean philosophers were at each other’s throats over this issue in ancient Greece: Does the way brains work allow us to act freely, or is what we think and do fully determined? The evidence discussed here and in previous chapters inevitably bears on this vexing question. Based on empirically determined sensory circuitry, the brain generates behavior that works surprisingly well in a world that our senses can’t tell us about directly. This way of producing perceptual content and behavior is very different from our subjective impression of what is going on. Everyday experience suggests that
our brains (a stand-in for the “me” or the “I” that we take for granted) analyze stimulus features and represent these features, and that we then make decisions about how to act based on representations of the world “as it really is.” But if brain activity and its perceptual and behavioral consequences simply reflect neural associations created by accumulated phylogenetic and ontogenetic experience, then a wholly empirical understanding of brain function comes down squarely on the side of determinism. In this framework, terms such as inferences, decisions, and choices are apt descriptors of our subjective sense of the way we relate to the world, but not of any underlying brain function.

To sum up how brains seem to work, the circuitry of nervous systems such as ours has evolved to contend with one fundamental challenge: How to generate useful perceptions and behaviors in response to a world that is unknowable directly by means of sensory stimuli. The strategy that has emerged to deal with this problem is governed by history, not logical principles or algorithms. Based on feedback from the empirical consequences of behavior, accumulated information about operational success is realized over evolutionary time in inherited neural circuitry whose organization is then modified to a limited extent by individual experience. Accordingly, our perceptions never correspond to physical reality despite the fact that they provide successful operational guides to behavior. The evidence that supports these conclusions is the ability to predict many otherwise puzzling perceptual phenomena using databases that serve as proxies for aspects of accumulated human experience.

These ideas about what we perceive, what we do as a result, and ultimately what we are in consequence may be anathema to neuroscientists committed to a rationalist perspective of brain function, or to anyone who has difficulty disengaging from the subjective sense we have of the world and our relation to it. But if the evidence continues to support a wholly empirical interpretation of how brains work, we will simply need to pursue brain function and structure in these terms, perhaps gaining in the process a clearer and more useful conception of ourselves and our place in nature.
This page intentionally left blank