1 Introduction

1.1 COMPUTATIONAL MODELS AS TOOLS

At some point in our childhood, many of us played with model planes made of balsa wood or cardboard. Such models often have flat wings and a twisted rubber band connected to a small propeller; when the plane is launched into the air, the tension on the rubber band is released, driving the propeller to spin, and the plane soars through the air for a few minutes of flight. A future scientist playing with such a toy could learn many general principles of aviation; for example, in both the toy plane and a Boeing 747, stored energy is converted to rotary motion, which provides the forward speed to create lift and keep the plane in the air.

Aerodynamic engineers use other types of airplane models. In the early days of aviation, new planes were developed by using wooden models of airplane shapes, which were placed in wind tunnels to test how the air flowed across the wings and body. Nowadays, much of the design and testing is done with computer models rather than wooden miniatures in wind tunnels. Nevertheless, these computer-generated models accomplish the same task: They extract and simplify the essence of the plane's shape and predict how this shape will interact with wind flow.

Unlike the toy airplane, the engineer's aerodynamic model has no source of propulsion and cannot fly on its own. This does not mean that the toy airplane is a better model of a real airplane. Rather, each model focuses on a different aspect of a real airplane, capturing some properties of airplane flight. *The value of these models is intrinsically tied to the needs of the user; each captures a different design principle of real planes.*

A model is a simplified version of some complex object or phenomenon. The model may be physical (like the engineer's wind tunnel) or virtual (like the computer simulation). In either case, it is intended to capture some of the properties of the object being modeled while disregarding others that, for the time being, are thought to be nonessential for the task at hand. Models are especially useful for testing the predictive and explanatory value of

abstract theories. Thus, in the above examples, theories of propulsion and lift can be tested with the toy plane, while theories of aerodynamic flow and turbulence can be tested with the engineer's wind tunnel model or the computer simulation of that wind tunnel.

Of course, these are not the only models that could be used to test principles of aviation; many different models could be constructed to test the same ideas. The superficial convergence of a model and the world does not prove that the model is correct, only that it is plausible.

We believe that models should be evaluated primarily in accord with how useful they are for discovering and expressing important regularities and principles in the world. Like a hammer, a model is a tool that is useful for some tasks. However, no single tool in a carpenter's kit is the most correct; similarly no single model of the brain, or of a specific brain region, is the most correct. Rather, different models work together to answer different questions.

In evaluating a model's usefulness, it is important to keep in mind that the utility of a model depends not only on how faithful it is to the real object, but also on how many irrelevant details it eliminates. For example, neither the rubber-band toy nor the aerodynamic model incorporates passenger seating or cockpit radar, even though both features are critical to a real airplane. These additions would not improve the toy plane's ability to fly, nor would they add to the engineers' study of wind resistance. Adding such details would be a waste of time and resources and would distract the user from the core properties being studied.

The ideals of simplicity and utility also apply to brain models. Some basic aspects of brain function are best understood by looking at simple models that embody one or two general principles without attempting to capture all the boggling complexity of the entire brain. By eliminating all details except the essential properties being studied, these models allow researchers to investigate one or two features at a time. *By simplifying and isolating core principles of brain design, models help us to understand which aspects of brain anatomy, circuitry, and neural function are responsible for particular types of behaviors.* In this way, models are especially important tools for building conceptual bridges between neuroscience and psychological studies of behavior.

The brain models presented in this book are all simulated within computers, as are the aerodynamic models used by modern airplane designers. Chapters 3 through 5 will explain in more detail how such computer simulations of neural network models are created and applied.

Most of the research described in this book proceeds as follows. A body of behavioral and neurobiological data is defined, fundamental principles and

regularities are identified, and then a model is developed and implemented as a computer simulation of the relevant brain circuits and their putative functions. Often, these brain models include several components, each of which corresponds to a functionally different region of the brain. For example, there might be one model component that corresponds to the cerebral cortex, one for the subcortical areas of the brain, and so forth. By observing how these components interact in the model, we may learn something about how the corresponding brain regions interact to process information in the normal brain.

Once we are confident that a model captures observed learning and memory behaviors and reflects the anatomy of an intact brain, we can then ask what happens when one or more model brain regions are removed or damaged. We would hope that the remaining parts of the model behave like a human or animal with analogous brain damage. If the behaviors of the brain-damaged model match the behaviors of animals or people with similar damage, this is evidence that the model is on the right track. This is the approach taken by many of the models presented in this book.

The usefulness of the models as tools for furthering research comes from novel predictions that the models make. For example, the model might predict that a particular form of brain damage will alter learning and memory in a particular way. These predictions are especially useful if the predictions are surprising or somehow unexpected given past behaviors or data. If the predictions are correct, this strengthens one's confidence in the model; if the predictions are incorrect, this leads to revisions in the model.

However, even a model of relatively simple behaviors can quickly become so complex that it seems an advanced degree in mathematics is required just to understand it. When theories and models are comprehensible only to other modelers, they lose their ability to function as effective tools for guiding empirical research. Rather, *it should be possible for most psychologists and neuroscientists to understand the intuitive ideas behind a computational model without getting mired in the details.*

In this book we have tried to summarize—at a conceptual level—neural network modeling of hippocampal function with little or no reference to the underlying math.

1.2 GOALS AND STRUCTURE OF THIS BOOK

The goal of the first five chapters—constituting part I—is to level the playing field so that the rest of the book is accessible to anyone in the behavioral and neural sciences, including clinical practitioners such as neuropsychologists, psychiatrists, and neurologists.

Some of the material in part I is likely to be too elementary for many readers. For example, we expect that many neuroscientists and their graduate students will be able to skip chapter 2 ("The Hippocampus in Learning and Memory") because that material is covered in most neuroscience graduate programs (and some psychology programs). In contrast, chapters 3 through 5 cover material that is likely to be too rudimentary for computer scientists, engineers, and others with a strong background in the formal basis of neural network models. These readers may wish to skip from chapter 2 to chapter 6, the beginning of part II.

As a caveat, we note that the tutorial material in part I does not conform to the standard organization and scope found in most textbooks. Rather, we have given the material our own spin, emphasizing the themes and issues that we believe are most essential to appreciating the models and research presented in the second half of the book.

For example, our coverage of the hippocampus and memory in chapter 2 focuses not on the more traditionally recognized hippocampal-dependent behaviors—such as the recall of past episodes or explicit facts—but rather on simpler behaviors, especially classical Pavlovian conditioning, that have formed the basis for a great deal of computational modeling.

To better understand the methods of information-processing theories of brain function, chapters 3 through 5 provide an introduction to the fundamentals of neural network modeling. Again, this tutorial is nonstandard in that it emphasizes the historical roots of neural network theories within psychological theories of learning and the relevance of these issues to modern studies of the hippocampus and learning behaviors.

Chapter 3 serves as an introduction to simple neural network models. It focuses on learning rules used for the formation of associations and the relevance of these rules to understanding the neural circuits necessary for classical conditioning. For continuity with the rest of the book, these networks will be illustrated through their application to classical conditioning. In this chapter, we introduce an early forerunner of modern neural network theories: the Rescorla-Wagner model of classical conditioning, which is in many ways a "model" model. It has stood for nearly thirty years as an example of how it is possible to take a set of complex behaviors, pare away all but the essence, and express the underlying mechanism as an intuitively tractable idea. Moreover, it now appears that the Rescorla-Wagner model may be more than just a psychological description of learning; it may also capture important properties of the kinds of learning that occur outside the hippocampus, especially in the cerebellum.

Chapter 4 introduces a fundamental problem that is common to the fields of psychology, neuroscience, and neural networks: How are events in the

outside world transformed into their neural representations, that is, the corresponding physical changes and processes within the brain? Researchers from each discipline have grappled for years with the problem of representation, and each group has added important novel insights to our understanding of this problem.

Chapter 5 considers how animals and people learn from just the mere exposure to stimuli and what kinds of neural networks can capture this type of learning. It introduces a class of neural network architectures, called autoassociators, which have often been used to describe the functional significance of the very specialized circuitry found in the hippocampus. Interestingly, it is exactly this kind of learning from mere exposure that seems to be especially sensitive to hippocampal-region damage in animals. In addition, an autoassociator is capable of storing arbitrary memories and later retrieving them when given partial cues—exactly the sort of memory ability that is lost in amnesic patients with hippocampal-region damage.

This completes part I.

Part II of the book gets into the details of modeling hippocampal function in learning. Chapters 6 through 10 share a common format and organization. Each introduces a different behavioral or neurobiological phenomenon, reviews one or more computational models of these phenomena, relates these models to qualitative (noncomputational) theories of learning and the brain, and then closes with a discussion of the implications of these models for understanding human memory and its clinical disorders.

Chapter 6 builds on the discussions in chapters 4 and 5 of representation and mere-exposure learning and describes how these issues have motivated two different models of the interaction between the hippocampus and cortex during associative learning. These hippocampal models are compared to several qualitative noncomputational theories of the interaction between the hippocampus and cortex. The chapter shows how modeling of animal conditioning has led to new insights into why brain-damaged amnesic patients can sometimes learn associations faster than normal control subjects.

Chapter 7 focuses on the role of the hippocampus and medial temporal lobes in the processing of background stimuli such as the constant sounds, noises, and odors that are present in an experimental laboratory. Indeed, several early influential theories of the hippocampus argued that its chief function was in processing this kind of contextual information.

To more fully understand what the hippocampal region is doing, it is necessary to have some understanding of its inputs—and hence what the sensory cortices are doing to information from the eyes, ears, and other sense organs before they pass this information on to the hippocampus. Chapter 8

shows how certain types of network models can be related to cortical architecture and physiology. It presents a specific model that combines a cortical module with a hippocampal-region module and explores how these brain systems might interact. Finally, chapter 8 presents an example of how research into cortical representation has led to a real-world application to help children who are language-learning-impaired.

Chapter 9 continues the discussion of cortical representation by focusing on one particular region of cortex: the entorhinal cortex. The entorhinal cortex is physically contiguous to the hippocampus and is considered part of the hippocampal region as that term is used in this book. This chapter reviews three different computational models of the entorhinal cortex and its interaction with other brain regions. It then discusses the implications of theories of entorhinal (and hippocampal) function for understanding and diagnosing the earliest stages of Alzheimer's disease, which is characterized by cell degeneration and physical shrinkage in the entorhinal cortex and hippocampus.

An emerging theme from these studies and models is that the hippocampal region does not operate in isolation; rather, to understand the hippocampus, one must understand how it interacts and cooperates with the functioning of other brain regions. Accordingly, chapter 10 considers additional brain regions that provide chemical messengers that alter the functioning of hippocampal-region neurons. This chapter first provides a brief review of neurotransmission and neuromodulation, with particular attention to acetylcholine and how it affects memory. Next, the chapter discusses computational models that suggest that acetylcholine released from the medial septum into hippocampus is integral in mediating hippocampal function and a model that addresses the effects of changes in acetylcholine levels on learning and memory.

The final chapter, chapter 11, reviews several key themes that recur throughout the book. They are:

1. Hippocampal function can best be understood in terms of how the hippocampus interacts and cooperates with the functioning of other brain systems.
2. Partial versus complete lesions may differ in more than just degree.
3. Disrupting a brain system has different effects than removing it.
4. Studies of the simplest forms of animal learning may bootstrap us toward understanding more complex aspects of learning and memory in humans.
5. The best theories and models exemplify three principles: Keep it simple, keep it useful, and keep it testable.

These five themes represent the core message of this book. In the ten chapters that follow, we will elaborate on these themes as they are exemplified in a variety of specific research programs. Through this story we hope to communicate to the broader scientific community how and why computational models are advancing our understanding of the neural bases of learning and memory.

Many questions about hippocampal function in learning still remain unanswered. Some of these open questions are empirical, and we will suggest, at several places throughout the book, what we think are some of the more pressing empirical issues that need to be resolved by further experimentation. Other open questions are of a more theoretical nature, and we will suggest several new modeling directions for future efforts. Although we have aimed this book primarily at non-modelers, we hope that we may excite a few of our readers to go on to become modelers themselves, or to incorporate computational modeling into their own research programs through collaboration with modelers.