# BRAINS, MACHINES, AND MATHEMATICS

MICHAEL A. ARBIB

# Brains, Machines, and Mathematics

**Michael A. Arbib**

*University of New South Wales*
*Sydney, Australia* and
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

**McGraw-Hill Book Company**
*New York San Francisco Toronto London*

**Brains, Machines, and Mathematics**



*Library of Congress Catalog Card Number* 63-21473

02170

# Preface

This book forms an *introduction* to the common ground of brains, machines, and mathematics, where mathematics is used to exploit analogies between the working of brains and the control-computation-communication aspects of machines. It is designed for a reader who has heard of such currently fashionable topics as cybernetics, information theory, and Gödel's theorem and wants to gain from one source more of an understanding of them than is afforded by popularizations. Here the reader will find not only *what* certain results are, but also *why*. The number of pages has been deliberately kept small so that a first reading is feasible in an evening or two. Yet a lot of ground is covered, and the reader who wants to go further should find himself reasonably well prepared to tackle the technical literature. Full use of the book does require a moderate mathematical background—a year of college calculus (or the equivalent "mathematical maturity"). However, much of the book should be intelligible to the reader who chooses to skip the mathematical proofs, and no previous study of biology or computers is required at all.

Before reviewing the contents, I should say a few words as to the present status of neurophysiology and the nature of our model making.

The use of microelectrodes, electron microscopes, and radioactive tracers has yielded a huge increase in neurophysiological

knowledge in the past few decades. Even a multivolume work such as the "Handbook of Neurophysiology" cannot fully cover all the facts. Many neurophysiological theories, once widely held, are being questioned as improved techniques reveal finer structures and more sophisticated chemicoelectrical cellular mechanisms. This means that our presentation of mathematical models in this book will have to be based on a grossly simplified view of the brain and the central nervous system. The reader may well begin to wonder what value or interest the study of such systems can have.

There is a variety of properties—memory, computation, learning, purposiveness, reliability despite component malfunction—which it might seem difficult to attribute to "mere mechanisms." However, herein lies one important reason for our study: By making mathematical models, we have proved that there do exist purely electrochemical mechanisms which have the above properties. In other words, we have helped to "banish the ghost from the machine." We may not *yet* have modeled *the* mechanisms that the brain employs, but we have at least modeled *possible* mechanisms, and that in itself is a great stride forward.

There is another reason for such a study, and it goes much deeper. Many of the most spectacular advances in *physical* science have come from the wedding of the mathematicodeductive method and the experimental method. The mathematics of the last 300 years has grown largely out of the needs of physics—applied mathematics directly, and pure mathematics indirectly by a process of abstraction from applied mathematics (often for purely esthetic reasons far removed from any practical considerations). In these pages we coerce what is essentially still the mathematics of the physicist to help our slowly dawning comprehension of the brain and its electromechanical analogs. It is probable that the dim beginnings of *biological* mathematics here discernible will one day happily bloom into new and exciting systems of pure mathematics. Here, however, we *apply* mathematics to derive far-reaching conclusions from clearly stated premisses. We can test the adequacy of a model of the brain by expressing it in mathematical form and using our mathematical tools to prove general theorems.

In the light of any discrepancies we find between these theorems and experiments, we may return to our premisses and reformulate them, thus gaining a deeper understanding of the workings of the brain. Further, such theories can guide us in building more useful and sophisticated machines.

The beauty of this mathematicodeductive method is that it allows us to *prove* general properties of our models and thus affords a powerful adjunct to model making in the wire and test-tube sense.

Biological systems are so much more complicated than the usual systems of physics that we cannot expect to achieve a fully satisfactory *biological* mathematics for many years to come. However, the quest is a very real and important one. This book strives to introduce the reader to its early stages. He will, I hope, find that the results so far obtained are of interest. Certainly they represent only a very minute fraction of what remains to be found —but the start of a quest is nonetheless exciting for being the start. I do not believe that the application of mathematics will solve all our physiological and psychological problems. What I do believe, though, is that the mathematicodeductive method must take an important place beside the experiments and clinical studies of the neurophysiologist and the psychologist in our drive to understand brains, just as it has already helped the electrical engineer to build the electronic computers which, though many, many degrees of magnitude less sophisticated than biological organisms, still represent our closest man-made analog to brains.

We can now review the scope of this book:

We will first take a very quick look at neurophysiology, and from this we will formulate our crude first model of the brain as a network of components called McCulloch-Pitts neurons. We will see that anything an electronic computer can do can be done by such a network. We shall study the relation of these networks with finite automata and Turing machines; review work on the visual system of the frog as an example of complicated brain structure; and study the Perceptron (a machine that "learns"). We shall then review neurological evidence for neuron malfunction. This review

will stress the need to understand how to design networks which function reliably despite component malfunction. After a glance at the early von Neumann approach, we shall study Shannon's communication theory. We shall then be able to consider the Cowan-Winograd solution to the problem of reliable design. Then we turn to the study of Norbert Wiener's cybernetics—the study of control and communication in the animal and the machine. We shall examine the fundamental concept of feedback and the resultant insights gained into the functioning of the nervous system. We then take Greene's scheme of resonant frequencies in neural nets as an antidote to a too-ready identification of real brains with McCulloch-Pitts neural nets. After a discussion of homeostasis and prosthesis, we shall turn to Gestalt and the recognition of universals—how we perceive auditory and visual forms. The final chapter will be devoted to Gödel's incompleteness theorem. We shall give a historical outline of the trends in mathematical thought which led up to Gödel's work, prove the theorem, discuss its dramatic philosophical consequences for the foundations of mathematics, and finally look at its role in the brain-machine controversy.

This book is a revision of the lecture notes of a course delivered in June–August of 1962 at the University of New South Wales in Sydney, Australia. I want to thank John Blatt for inviting me to the Visiting Lectureship; Derek Broadbent for inviting me to broadcast the lectures; and Joyce Kean for her superb job of typing up the original lecture notes.

I have spent the last two years with the Research Laboratory of Electronics and the Department of Mathematics at the Massachusetts Institute of Technology on a research assistantship (supported by the U.S. Armed Forces and National Institute of Health). I owe so many debts of gratitude to the people there that I cannot fully do justice to them. However, I do particularly want to thank Warren McCulloch for his continual help and encouragement. It was George W. Zopf who first urged publication of the lectures. Bill Kilmer gave the original lecture notes a helpful and critical reading. For years I have nurtured the desire to claim at a

point such as this, "Any mistakes which remain are thus solely his responsibility." However, this would be a sorry expression of a very genuine gratitude, and so I follow convention and admit that any errors which remain are my responsibility.

Finally, I should like to thank all those authors whose work I have quoted and their publishers for so graciously granting me permission to use their material.

*Michael A. Arbib*

# Contents

# 1

# Neural Nets, Finite Automata, and Turing Machines

## 1.1 Introductory Neurophysiology

I want to start by giving a very sketchy account of neurophysiology—merely sufficient as a basis for our first mathematical model. We may regard the nervous system of man as a three-stage system as shown in Fig. 1.1.†

*Our fundamental hypothesis in setting up our model is that all the functioning of the nervous system relevant to our study is mediated solely by the passage of electrical impulses by cells we call neurons.* Actually, the human brain contains more *glial* cells than it contains *neurons*. Until recently, it was neurophysiological orthodoxy to believe that these glial cells served only to support and nourish the neurons—functions irrelevant to our study. However, the last 15 years have seen a growing number support the view that the

† The purpose of the arrows drawn from right to left will be made clear in the discussion of feedback in Sec. 4.1.

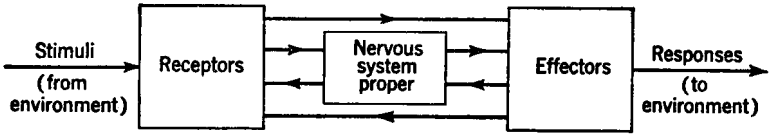


**Figure 1.1** The nervous system of man considered as a three-stage system.

glial cells actually carry out functions, such as memory, which are of interest here. Throughout this book, we shall ignore such posited glial functions. We shall also ignore such modes of neural interaction as continuously variable potentials and transmission of hormones. In setting up our *possible* mechanisms, neural impulses will fully suffice—*future developments will, of course, require the ascription of far greater importance to the other neural functions and perhaps to the glia.*

In the light of our fundamental hypothesis, then, we shall simply view the nervous system proper as a vast network of neurons, arranged in elaborate structures with extremely complex interconnections. This network receives inputs from a vast number of receptors: ıe rods and cones of the eyes, the pain, touch, hot, and cold receptors of the skin, the stretch receptors of the muscles, etc., all converting stimuli from the body or the external world into patterns of electrical impulses which convey information into the network. These interact with the enormously complicated patterns already traveling through the neurons (there are estimated to be $10^{10}$ neurons in the neural net which is the human brain!) and result in the emission of impulses which control the effectors, such as our muscles and glands, to give our responses. Thus we have our three-stage system: receptors, neural net, and effectors.

We are not going to formulate models of the receptors or effectors here, but we do want a model of the neural net. To do this, we shall first model the neuron. The neurons of our nervous system come in many forms, but we shall restrict our study to neurons like that of Fig. 1.2.

The *neuron* is a cell and so has a nucleus, which is contained in the *soma* or *body* of the cell. One may think of the *dendrites* as a very fine filamentary bush, each fiber being thinner than the axon, and of the *axon* itself as a long, thin cylinder carrying

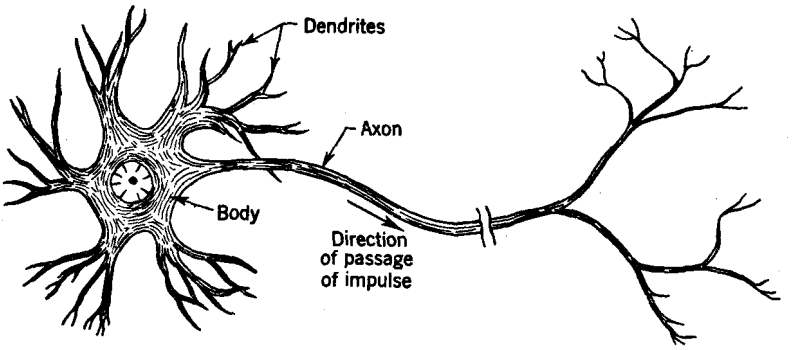


**Figure 1.2** Schematic drawing of a neuron.

impulses from the soma to other cells. The axon splits into a fine arborization, each branch of which finally terminates in a little *endbulb* almost touching the dendrites of a neuron. Such a place of near contact is called a *synapse.* Impulses reaching a synapse set up graded electrical signals in the dendrites† of the neuron on which the synapse impinges, the interneuronal transmission being sometimes electrical and sometimes by diffusion of chemicals. A particular neuron will only fire an electrical impulse along its axon if sufficient impulses reach the endbulbs impinging on its dendrites in a short period of time, called the *period of latent summation.* Actually, these impulses may either help or hinder the firing of an impulse and are correspondingly called *excitatory* or *inhibitory.* The condition for the firing of a neuron is then that the excitation should exceed the inhibition by a critical amount called the *threshold of the neuron.* If we assign a suitable positive weight to

† A synapse may also occur on other axons. This "interaction of afferents" is discussed in Sec. 3.1.

each excitatory synapse and a negative weight to each inhibitory synapse, we can say that

> *a neuron fires only if the total weight of the synapses which receive impulses in the period of latent summation exceeds the threshold* (1.1.1)

This picture of simple linear summation is, again, a gross simplification. Further, the threshold is a time-varying parameter—however, this time variance has rarely been considered in formal neuron modeling and plays no part in the models we shall consider here. The reader who is close to despair at this ever-widening departure from reality is advised to re-read the introduction for encouragement!

There is a small time delay between a period of latent summation and the passage of the corresponding axonal impulse to its endbulbs, so that the arrival of impulses on the dendrites of a neuron determines the firing of its axon at a slightly later time.

After an impulse has traveled along an axon, there is a time called the *refractory period* during which the axon is incapable of transmitting an impulse. Hence, during a length of time equal to one refractory period, at most one impulse may be fired along the axon. If we now choose as our unit of time the refractory period of the neuron, we may specify the firing behavior of our neuron by specifying for each of the first, second, third, etc., time intervals whether or not the neuron fired. We are thus led to the simplifying assumption that our neuron (already far removed from reality) may only fire at times $t = 1, 2, 3, 4, \ldots$ units of time after some suitable origin. We next make the *gross* assumption that we may use the same *discrete time scale* for all the neurons of our net. That is, we assume the firing behavior of our net is completely specified by the firing pattern of the individual neurons at the discrete times $t = 1, 2, 3, \ldots$. In line with this, we assume that the axonal firing of a neuron is determined by the firing pattern of inputs at its synapses one moment of our discrete time scale earlier.

## 1.2 The McCulloch-Pitts Model

The highly simplified neurophysiological considerations of the last section lead to the McCulloch-Pitts model of the neuron:

***Definition 1.2.1*** A *module* (or *formal neuron*) is an element with, say, $m$ inputs $x_1, \ldots, x_m$ ($m \geq 1$) and one output $d$. It is characterized by $m + 1$ numbers, its threshold $\theta$, and the weights $w_1, \ldots, w_m$, where $w_i$ is associated with $x_i$. The module operates on a discrete time scale $t = 1, 2, 3, 4, \ldots$, the firing of its output at time $n + 1$ being determined by the firing of its inputs at time $n$ according to the following rule (cf. Statement 1.1.1): The module fires an impulse along its axon at time $n + 1$ if and only if the total weight of the inputs stimulated at time $n$ exceeds the threshold of the neuron.

If we introduce the symbolism

$m(t) = 0$ for "$m$ does not fire at time $t$"
$m(t) = 1$ for "$m$ does fire at time $t$"

(where $m$ may be an axonal output or a synaptic input of a neuron), we see that the above rule may be expressed symbolically as

$$d(n + 1) = 1 \quad \text{if and only if } \Sigma w_i x_i(n) \geq \theta$$

Note that a positive weight $w_i > 0$ corresponds to an excitatory synapse (i.e., module input) whereas a negative weight $w_i < 0$ means that $x_i$ is an inhibitory input.

In terms of this very simple model of a neuron, we may immediately define our first model of a neural net:

***Definition 1.2.2*** A *modular net* is a collection of modules, each with the same time scale, interconnected by splitting the output of any module into a number of lines and connecting some or all of these to the inputs of other modules. An output may thus lead to any number of inputs, but an input may only come from at most one output.

The *input lines* of a net are those inputs $l_0, l_1, \ldots, l_{m-1}$ of modules of the net which are not connected to modular outputs. The *output lines* of a net are those lines $p_0, p_1, \ldots, p_{r-1}$ from modular outputs which are not connected to modular inputs.

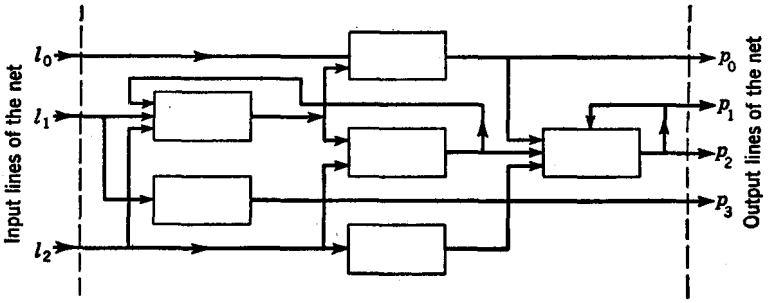


**Figure 1.3** A simple modular net.

In the example of Fig. 1.3, there are three input lines and four output lines—note that the input lines may split and that the output lines need not come from distinct modules.

We have now set up a model of the brain. To the reader who thinks of a model as an actual collection of wires and transistors, my use of the word "model" here may seem somewhat strange. Therefore, let me stress that throughout this book, the word is used in the mathematical sense. The *engineer* feels he has modeled a system when he has actually constructed an apparatus which he can hope will behave similarly to the original system. The *mathematician*, on the other hand, feels that he has modeled a system when he has "captured" some properties of the system in precise mathematical definitions and axioms in such a form that he can deduce further properties of this "formal" (i.e., mathematical) *model;* thus, hopefully, *explaining* known properties of the original system and *predicting* new properties. The concept of a "modular net" has a precise mathematical definition (and we shall prove theorems about it in subsequent sections), and it is in this mathematical sense that we consider it to be a model of

the brain. Before we study it, let us stress that *we have only obtained it at the cost of drastic simplifications:*

**a.** We have assumed complete synchronization of all the neurons.
**b.** We have fixed the threshold and weights of each neuron for all time.
**c.** We have ignored the effects of hormones and chemicals (e.g., alcohol) in changing the behavior of the brain.
**d.** We have ignored all interaction between neurons (e.g., due to the electrical field associated with their impulses) save that taking place at the synapses.
**e.** We have ignored the glial cells.

The list can be extended, and it must be realized that *our first model is only a starting point for our study and not an end in itself.* However, our simplifications have not rendered our model completely powerless, and a modular network can indeed store information and carry out computations. We shall demonstrate this in Sec. 1.4 by "blueprinting" a digital computer as a modular network. First, however, we shall use Sec. 1.3 to introduce the concept of "finite automaton" and relate it to that of "modular network"; while in Sec. 1.7 we shall consider a trivial model of perception by giving a mathematical characterization of the dichotomies which may be made of its input sequences by a finite automaton.

## 1.3 Finite Automata and Modular Nets

In this section, we introduce the concept of finite automaton in such a way as to make it clear that every modular net is a finite automaton. Our objective will then be to show that, conversely, the input-output behavior of a finite automaton can always be carried out by a suitably constructed modular net. Since it is much easier, in general, to design a finite automaton for a given task than to design the corresponding modular net, the result clarifies for us what tasks our modular nets are capable of performing (cf. Sec. 1.4).