# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>7</td>
</tr>
<tr>
<td>Abstract</td>
<td>9</td>
</tr>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>The Neocortex as a Uniform Cognitive Architecture</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Computation and the Brain</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Cognitive Architecture</td>
<td>15</td>
</tr>
<tr>
<td>2.3 The Importance of the Cortex in Cognition</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Anatomical Uniformity of the Cortex</td>
<td>20</td>
</tr>
<tr>
<td>2.5 The Fundamental Hypothesis</td>
<td>29</td>
</tr>
<tr>
<td>2.6 Some Objections to the Hypothesis</td>
<td>31</td>
</tr>
<tr>
<td>2.6.1 Modular Anatomy and Localization of Function</td>
<td>31</td>
</tr>
<tr>
<td>2.6.2 Diversity of Functional Properties of Cortical Neurons</td>
<td>32</td>
</tr>
<tr>
<td>2.6.3 From Uniform Neural Circuitry to Uniform Cognitive Architecture</td>
<td>34</td>
</tr>
<tr>
<td>2.7 Summary</td>
<td>36</td>
</tr>
<tr>
<td>A General Theory of the Neocortex</td>
<td>38</td>
</tr>
<tr>
<td>3.1 The Model Cortical Neuron</td>
<td>38</td>
</tr>
<tr>
<td>3.2 Population Coding</td>
<td>42</td>
</tr>
<tr>
<td>3.3 The Local Circuitry</td>
<td>44</td>
</tr>
<tr>
<td>3.4 Feedforward Activation in Layers 2–4</td>
<td>47</td>
</tr>
<tr>
<td>3.5 Motor Production in Layer 5</td>
<td>51</td>
</tr>
<tr>
<td>3.6 Feedback Activation From Layer 6</td>
<td>55</td>
</tr>
<tr>
<td>3.7 Lateral Activation in Layers 2–4</td>
<td>60</td>
</tr>
<tr>
<td>3.8 Plasticity</td>
<td>61</td>
</tr>
<tr>
<td>3.8.1 Properties of Cortical Plasticity</td>
<td>61</td>
</tr>
<tr>
<td>3.8.2 Weight Modification Rules</td>
<td>63</td>
</tr>
<tr>
<td>3.8.3 Molecular Mechanisms</td>
<td>67</td>
</tr>
<tr>
<td>3.9 Large-Scale Organization of the Cortex</td>
<td>69</td>
</tr>
<tr>
<td>3.10 The Focusing Mechanism</td>
<td>72</td>
</tr>
<tr>
<td>3.10.1 Properties of the Mechanism</td>
<td>73</td>
</tr>
<tr>
<td>3.10.2 Anatomical Substrate</td>
<td>78</td>
</tr>
<tr>
<td>3.11 Summary</td>
<td>80</td>
</tr>
<tr>
<td>Early Visual Processing</td>
<td>82</td>
</tr>
<tr>
<td>4.1 Review of the Subcortical Visual System</td>
<td>83</td>
</tr>
<tr>
<td>4.1.1 Luminance Processing in Retina and LGN</td>
<td>85</td>
</tr>
<tr>
<td>4.1.2 Wavelength Processing in Retina and LGN</td>
<td>85</td>
</tr>
<tr>
<td>4.2 Overview of Early Processing</td>
<td>87</td>
</tr>
<tr>
<td>4.3 Luminance</td>
<td>88</td>
</tr>
<tr>
<td>4.3.1 Simple and Complex Receptive Fields</td>
<td>89</td>
</tr>
<tr>
<td>4.3.2 Laminar and Columnar Organization</td>
<td>93</td>
</tr>
<tr>
<td>4.3.3 The End-Stopping Property</td>
<td>94</td>
</tr>
<tr>
<td>4.3.4 Plasticity of Receptive Field Properties</td>
<td>94</td>
</tr>
<tr>
<td>4.4 A Model of Simple and Complex Receptive Fields</td>
<td>95</td>
</tr>
<tr>
<td>4.4.1 Description of the Model</td>
<td>95</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The human brain, medial section</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>The human brain, frontal section</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Brodmann’s areas</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Layers of the neocortex</td>
<td>20</td>
</tr>
<tr>
<td>2.5</td>
<td>Cortical neurons</td>
<td>22</td>
</tr>
<tr>
<td>2.6</td>
<td>Cortical neurons</td>
<td>23</td>
</tr>
<tr>
<td>2.7</td>
<td>The thalamus</td>
<td>24</td>
</tr>
<tr>
<td>2.8</td>
<td>Feedforward and feedback connections</td>
<td>26</td>
</tr>
<tr>
<td>2.9</td>
<td>Areas and connections of the visual cortex</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td>Equivalent circuit</td>
<td>39</td>
</tr>
<tr>
<td>3.2</td>
<td>Firing rate of model neuron</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>The local circuitry</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Effect of the local circuitry</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>The canonical circuit (1)</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>The canonical circuit (2)</td>
<td>52</td>
</tr>
<tr>
<td>3.7</td>
<td>Population coding in the motor cortex</td>
<td>54</td>
</tr>
<tr>
<td>3.8</td>
<td>The canonical circuit (3)</td>
<td>55</td>
</tr>
<tr>
<td>3.9</td>
<td>Contextual enhancement</td>
<td>57</td>
</tr>
<tr>
<td>3.10</td>
<td>Connections from layer 6 to layer 4</td>
<td>58</td>
</tr>
<tr>
<td>3.11</td>
<td>The canonical circuit (4)</td>
<td>59</td>
</tr>
<tr>
<td>3.12</td>
<td>The $\phi$ function in the BCM rule</td>
<td>63</td>
</tr>
<tr>
<td>3.13</td>
<td>Graph of LTP and LTD</td>
<td>65</td>
</tr>
<tr>
<td>3.14</td>
<td>Firing rate trace</td>
<td>66</td>
</tr>
<tr>
<td>3.15</td>
<td>Body map in somatosensory cortex</td>
<td>70</td>
</tr>
<tr>
<td>3.16</td>
<td>The focusing mechanism</td>
<td>75</td>
</tr>
<tr>
<td>3.17</td>
<td>Interference in active storage</td>
<td>76</td>
</tr>
<tr>
<td>4.1</td>
<td>The visual pathway</td>
<td>83</td>
</tr>
<tr>
<td>4.2</td>
<td>Retinal ganglion cell receptive fields</td>
<td>84</td>
</tr>
<tr>
<td>4.3</td>
<td>Spectral sensitivity of rods and cones</td>
<td>86</td>
</tr>
<tr>
<td>4.4</td>
<td>Wavelength selective receptive fields</td>
<td>87</td>
</tr>
<tr>
<td>4.5</td>
<td>Simple receptive fields</td>
<td>88</td>
</tr>
<tr>
<td>4.6</td>
<td>Hubel-Wiesel model for simple cells</td>
<td>89</td>
</tr>
<tr>
<td>4.7</td>
<td>Complex receptive field model</td>
<td>91</td>
</tr>
<tr>
<td>4.8</td>
<td>Structure of the model</td>
<td>96</td>
</tr>
<tr>
<td>4.9</td>
<td>Image and convolution</td>
<td>97</td>
</tr>
<tr>
<td>4.10</td>
<td>Image and convolution</td>
<td>98</td>
</tr>
<tr>
<td>4.11</td>
<td>Initial connectivity of a middle layer unit</td>
<td>100</td>
</tr>
<tr>
<td>4.12</td>
<td>Initial connectivity of an upper layer unit</td>
<td>101</td>
</tr>
<tr>
<td>4.13</td>
<td>Trained weights of three middle layer units</td>
<td>105</td>
</tr>
<tr>
<td>4.14</td>
<td>Receptive fields from an artificial image</td>
<td>106</td>
</tr>
<tr>
<td>4.15</td>
<td>Trained weights of an upper layer unit</td>
<td>107</td>
</tr>
<tr>
<td>4.16</td>
<td>Subfields of an upper layer unit</td>
<td>108</td>
</tr>
<tr>
<td>4.17</td>
<td>Receptive fields of a single orientation</td>
<td>109</td>
</tr>
</tbody>
</table>
Acknowledgments

I especially want to thank my supervisor, Rudi Lutz, for his invaluable suggestions and criticisms as well as practical assistance throughout this project. I also thank him for giving me licence to pursue this rather unusual thesis, which bears almost no relation to my original research proposal! I thank Andy Clark for reading an earlier version of Chapter 2. Chapter 2 also benefited a great deal from the comments of two anonymous referees of the journal *Mind & Language*. Finally I thank my family for being there.

This research was supported for two years by a research studentship (ref. 88304970) from the Science and Engineering Research Council.
I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree.
ABSTRACT

The cerebral neocortex is a part of the brain that plays a major role in all cognitive activity. However, the structure of the cortex is remarkably uniform: the neural circuitry is roughly the same in all cortical areas. This thesis investigates the proposal that the cortex is a *uniform cognitive architecture*, providing a system of general-purpose computational mechanisms that are applied in diverse psychological domains. Evidence for cortical uniformity is reviewed, and assumptions and implications of the proposal are discussed. Objections to the proposal are considered, including modular organization, diversity of neuronal physiology, and the relation between computation and neural circuitry. A general theory of the cortex is then introduced. Among the general-purpose mechanisms are: a basic feedforward flow of information within layers 2–4 and between cortical areas, supplemented by feedback and lateral interactions; production of motor output in layer 5; representation of information by population coding; a form of gain control and enhancement of pattern-selectivity performed by localized excitatory and inhibitory circuitry; a form of synaptic plasticity causing neurons to learn patterns that occur frequently in the input; and a selective attention mechanism. This theory is then used to explain a broad range of empirical data in visual perception, including: early processing of luminance, motion, spatial frequency, binocular disparity, texture, lightness and colour; general aspects of shape perception and object recognition; and special topics in shape perception, including the Gestalt grouping effects, illusory contours, multistability and hysteresis, stereopsis, the figure-ground phenomenon, and characteristics of selective attention as studied by visual search. A simulation demonstrates how neurons can acquire simple and complex receptive fields, and organize themselves into a smooth map of orientation selectivity, from exposure to natural visual input.
One of the principal goals of cognitive science is to obtain a general scientific understanding of human cognition and behaviour—that is, of thinking, perceiving, acting, imagining, learning, remembering, attending, deciding, believing, and so on. But is it possible to develop a unified theory of the mind—a single theory that explains human cognition and behaviour in a wide variety of tasks by means of a common theoretical framework? Or would the ultimate theory of the mind necessarily be like a patchwork of miniature theories, more-or-less independent of each other, each concerned with one particular aspect of cognition? Interest in the possibility of a unified theory has increased since Allen Newell advocated the idea in his 1987 William James Lectures. To quote Newell (1990, p. 1):

Psychology has arrived at the possibility of unified theories of cognition— Theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition.

I do not say they are here. But they are within reach and we should strive to obtain them.

This thesis presents an argument that this is indeed a feasible goal, subject to certain limitations, and outlines a theory which, I claim, could ultimately be developed into a unified theory of cognition. The approach is quite different from that which underlies the candidate unified theory (the Soar architecture) developed by Newell and his colleagues (Laird, Newell & Rosenbloom, 1987). It is based on a remarkable observation from experimental neuroscience: a substantial degree of uniformity in the anatomy and physiology of the cerebral neocortex, a part of the brain that plays a major role in all cognitive activity. This observation has received a fair amount of attention from neuroscientists (e.g., Creutzfeldt, 1977; Mountcastle, 1978; Szentágothai, 1978; Rockel, Hiorns & Powell, 1980; Powell, 1981; Martin, 1988; Shepherd, 1988; Douglas, Martin & Whitteridge, 1989), and some theorists have made explicit proposals for common, general-purpose mechanisms that might be embodied in the cortex (e.g., Marr, 1970; Grossberg, 1980; Eccles, 1984; Barlow, 1985; Carpenter & Grossberg, 1987; Edelman, 1987; Mel & Koch, 1990; Poggio, 1990; Singer, 1990; Mumford, 1991, 1992; James & Hoang, 1993). In this thesis I follow this line of research by (1) presenting an argument that it is possible to base a unified theory of cognition on a general theoretical model of the neocortex, (2) supporting this case by outlining a general theory of the neocortex, and (3) attempting to account for a substantial body of empirical data from the psychology and neurophysiology of visual perception in terms of this theory. I emphasize that the thesis does not present a unified theory of cognition per se; it is merely a foundation for such a theory.

The structure of the thesis is as follows. Chapter 2 presents the general argument that the neocortex can be considered a uniform cognitive architecture, at least approximately. I first discuss a fundamental assumption of the approach, that the brain is a computational system. I then examine the concept of cognitive architecture, and distinguish a uniform architecture from a heterogeneous one. I point out the importance of the cortex in human cognition, and review the anatomical evidence for structural uniformity of the cortex. I then present the fundamental hypothesis, that the cortex is an approximately uniform cognitive architecture and that a unified theory of cognition could be built on a general model of the cortex. Finally, some objections to this proposal are discussed. I consider
modular structure and localization of function in the cortex, the manifest non-uniformity of the functional properties of cortical neurons, and the question of whether uniformity of neural mechanisms implies uniformity at the computational level.

Chapter 3 introduces a general theory of the cortex. The theory is based on arguments presented in Chapter 2 that the cortex can be considered a perceptual-motor processor that is largely dedicated to pattern recognition. Among the general-purpose computational mechanisms that I propose to be operating in all cortical areas are: a basic feedforward flow of information within layers 2–4 and between cortical areas, supplemented by relatively weak feedback and lateral interaction; the production of motor output in layer 5; representation of information by population coding; a form of gain control and enhancement of pattern-selectivity performed by localized excitatory and inhibitory circuitry; a form of synaptic plasticity that causes neurons to learn patterns that occur frequently in the input; and a selective attention mechanism. These mechanisms collectively account for most of the known components of the cortical circuitry.

Chapter 4 applies this theory to low-level visual processing. Following a review of the subcortical visual system, I discuss in some detail the ‘simple’ and ‘complex’ orientation-sensitive cells discovered in the visual cortex by Hubel and Wiesel (1962, 1968). I then present a computer simulation that demonstrates how neurons can acquire simple and complex receptive fields, and organize themselves into a smooth map of orientation selectivity, from exposure to natural visual input. The remainder of the chapter describes in less detail how the theory might account for the early processing of other types of visual information, including motion, spatial frequency, binocular disparity, texture, lightness and colour.

Chapter 5 applies the general theory to higher-level visual processing, in particular shape perception and object recognition. Special emphasis is given to the visual responses of neurons in the inferior temporal cortex, such as those that respond to faces. This is followed by a discussion of a number of special topics in shape perception, including: the Gestalt grouping effects, illusory contours and contour interpolation, the perception of three-dimensional shape in pictures, multistability of ambiguous pictures like the Necker cube, hysteresis effects in grouping, stereoscopic depth perception, the figure-ground phenomenon, and the characteristics of selective attention as studied by visual search.

The final chapter summarizes the main assumptions of the theory, and the thesis concludes with the obligatory speculations about consciousness.
Chapter 2
The Neocortex as a Uniform Cognitive Architecture

This chapter presents a case that the cerebral neocortex can be considered a uniform cognitive architecture, in a sense to be defined below, and that a unified theory of cognition could be built on a general theory of how the cortex works. The chapter is a revised version of a previously published article (Ebdon, 1993).

2.1 Computation and the Brain

It is a basic assumption of this thesis that the brain is a kind of computer. It is important to explain what this means, and more importantly what it does not mean.

In spite of the astonishing advances in speed, storage capacity and miniaturization that have taken place over the last fifty years or so, nearly all of the electronic computers in use today still conform to the basic design now called the von Neumann architecture. These computers differ from brains in a great many ways. Transistors have only a few connections; neurons have thousands of connections. Digital components have only two states, and switch from one state to the other at discrete intervals; neurons are essentially analogue devices that change state continuously. Every artificial computer of a particular make has an identical wiring diagram; each individual brain has its own unique connectivity, within the broad pattern common to the species. The von Neumann computer has a single ‘master’ processor (the central processing unit) that communicates with a number of ‘slave’ processors through channels of just a few tens of wires (the buses); brains have hundreds of processors (nuclei and cortical areas) that communicate through thousands of channels (tracts) each containing tens of thousands of wires (axons). (It is this latter difference, incidentally, that causes von Neumann computers to be ‘serial’ and brains ‘parallel’.) Artificial computers were designed by engineers; brains arose by natural selection—a fact that has important consequences, as we shall see.

This list could easily be extended but the point has been made: any similarity between the neural circuitry of the brain and the digital circuitry of artificial computers is at best superficial and at worst downright misleading. Many critics have inferred from this that the comparison between brains and computers is merely the latest chapter in a long history of trying to explain the brain in terms of the latest wonder of technology; to us the brain is a computer, to our grandparents it was a telephone exchange, to Descartes it was a hydraulic system of pipes and valves, etc. But there is more to it than this.

To claim that the brain is a computer is to claim that it is possible to regard certain physical states of the brain as representations. Representations are structures that have a semantic interpretation (they have content or meaning) and computations are processes that transform representations in a meaningful, content-preserving way. Certain aspects of the anatomy and physiology of the brain (in particular, the neurons and their electrical and chemical interactions) are said to realize or implement the representations and computations. For example, the computational approach to understanding the visual system is to assume that the brain is computing representations that specify the locations, shapes, sizes, colours, movements, etc. of the objects the animal is looking at—information that is merely implicit in the pattern of light falling on the two retinas.
This perspective has three advantages. First, because representations have meaning, they can explain how the brain performs the kinds of activities that we call cognitive—perceiving, thinking, learning, remembering, attending, intending, and the like. At the same time, because computations are implemented in physical stuff, they are consistent with the established laws of physics, and there is no problem in understanding how they can interact with the body and the rest of the world. Thus, the computational perspective appears to provide a solution to at least one aspect of the ancient problem of the relation between mind and brain; it explains how a purely physical system can do meaningful things. Second, the computational perspective does not commit us to the view that advanced cognitive activity can only possibly be performed by biological brains. It seems intuitively plausible that, in principle, a thinking machine could be constructed from non-biological (say, electronic or optical) components, and that such a machine would be just as capable of perceiving, imagining, learning, and the like as brains are. The computational perspective is consistent with this intuition, because the computations implemented by the human brain could also be implemented in a quite different physical substrate. The third justification is less philosophical and more practical: computer science provides a very useful vocabulary with which to talk about the brain. It is very natural to regard a tract of nerve fibres as carrying information from one place to another, or to talk of a particular brain structure as representing a certain kind of information; so natural, in fact, that even critics of computationalism tend to use such descriptions whenever their philosophical guard is down. But computational terminology is not merely convenient, it is also theoretically neutral, at least in comparison with psychological terms like attention, memory, image, and the like. For example, referring to a particular set of synaptic connections as a memory begs the question of how that anatomical structure relates to the hugely complex phenomenon that we normally mean by that word. Calling the structure a representation avoids such conceptual entanglements.

The computational perspective is established orthodoxy in cognitive science (see, e.g., Pylyshyn, 1984; Churchland & Sejnowski, 1992, Chap. 3). But the approach is not so broad that it is beyond criticism. It is incompatible with dualism: computations are physical processes, so if some aspect of cognition (such as consciousness) depends on non-physical processes, whatever that may mean, computationalism is false. Some philosophers have denied that representations can really be said to be meaningful (e.g., Searle, 1980). Some neuroscientists, psychologists, and even artificial intelligence researchers argue on various grounds that the processing of representations is an inappropriate way to describe what the brain is doing (e.g., Gibson, 1966; Edelman, 1987; Brooks, 1991). A sociologist might claim that computational terminology is hardly neutral but carries some hefty ideological baggage of its own. Interesting though these debates are (well, the first three anyway) I do not want to address them here.

Even within the computational framework there is a broad diversity of views. For example, some argue that the brain’s representations are largely autonomous with regard to their implementation (e.g., Marr, 1982; Fodor & Pylyshyn, 1988; Newell, 1990). According to this view, the representations and computations used in the brain can be understood in their own terms quite independently of the underlying neural machinery, in just the same way that a program written in Lisp can be understood without reference to the computer’s electronic circuits. If this is the case, the anatomy and physiology of the brain place very few constraints on the nature of the computations used in human cognition, except for some very general limitations on such things as storage capacity and processing speed. At the other end of this spectrum are advocates of computational neuroscience (e.g., Churchland, 1986; Churchland, Koch & Sejnowski, 1990; Churchland & Sejnowski, 1992), who argue that computation in the brain is tightly constrained by the neural machinery that realizes it, so that neuroscience has a large contribution to make towards understanding cognition. Naturally, the present thesis endorses the latter perspective. Specifically, the theory introduced in Chapter 3 takes the form of a simplified and idealized model of the neural circuitry of the cerebral neocortex. An argument justifying this approach is presented at the end of this chapter. For now, I will simply point out that to claim that the brain is a computer is not to imply that the representations and computations used in the brain have much in common with Lisp programs, any more than it implies that neurons have much in common with logic gates.
2.2 Cognitive Architecture

The use of the term *architecture* that originated in computer science (e.g., Tanenbaum, 1984) has become an important concept in the theoretical foundations of cognitive psychology (Anderson, 1983; Pylyshyn, 1984; Newell, Rosenbloom & Laird, 1989; Newell, 1990). There is no generally agreed-upon definition of cognitive architecture in the literature, but I will use the term here to refer to *all the structures and mechanisms of the cognitive system that are genetically specified and not modifiable by learning*. It therefore excludes those aspects of the system that are acquired, or at least can be altered, by experience—memories, knowledge, skills, habits, and the like. The cognitive architecture is assumed to determine the gross division of the cognitive system into processing modules and communication pathways, and to determine the basic *computational mechanisms* provided by each module. These mechanisms include the forms of representation that are available, and the elementary, unmodifiable processing operations that are directly implemented in the neural substrate. The architecture largely or completely specifies a great many things that are of interest to cognitive psychologists, such as: the number of different kinds of memory available (e.g., short-term versus long-term, procedural versus declarative, episodic versus semantic); basic characteristics of these memories (e.g., their organization, capacity, method of accessing information, causes of retrieval failure); the ways in which knowledge can be represented (e.g., symbol structures, images, skills) and the efficiency with which these different representations can be processed (as measured, e.g., by speed or error rates); and the characteristics of arousal and selective attention. It is natural to suppose, therefore, that a general theory of cognition will be based on a theory of the cognitive architecture (Newell, 1990; Newell, Rosenbloom & Laird, 1989).

Human cognition and behaviour in any situation is determined by three factors: the cognitive architecture, sensory inputs from the body and the environment, and the information represented in the architecture. This third factor includes the person’s thoughts and the stock of knowledge and skills that the person has accumulated up to that time; it also includes the person’s goals and desires, insofar as these are the result of learning and not built into the architecture. The psychologist who is interested in human behaviour in a given task and environment needs to know how the general capacities provided by the architecture get applied in that situation. There is no reason to suppose that this will be a trivial research task in general, even given a complete understanding of the architecture. Psychology is therefore an open-ended science: there is an endless variety of tasks and environments in which humans can operate and in which psychologists might be interested. A theory of the architecture is therefore necessary, but not sufficient, for a complete understanding of the mind.

In conventional digital computers, the distinction between the architecture and its information content is reasonably clear; it is roughly, although not exactly, the distinction between hardware and software. There are a number of problems, however, in applying the architecture concept to the brain. First, just as humans differ from one another in height, colour, build, and so on, it seems reasonable to suppose that there is some variation between individuals in cognitive architecture. However, these variations are probably very small relative to the similarities, and so it should be possible to formulate a simplified model of the architecture that is applicable to all normal humans. Second, it is generally believed that memories, skills, and other results of learning are stored in the brain by modifying the strengths of synaptic connections between neurons; that is, by changing the brain’s ‘hardware’. Therefore—and it is important to emphasize this—a theory of the architecture will not simply be a theory of the brain’s physical structure, but will cover only those aspects of its structure that are not modifiable by learning. This will probably include, for example, the number of neurons and their gross organization into nuclei and cortical areas, but not all aspects of the detailed synaptic connections (although the theory should certainly include a description of the laws by which those connections are modified.) The third and most serious difficulty is that the architecture is not fixed and immutable: the brain’s structure can be modified in ways that we would not wish to call learning. We can easily exclude alterations that are clearly exceptional, like those caused by disease and injury, but it is difficult to ignore the effects of maturation. The problem is that the human cognitive architecture is not fixed at birth, but continues to mature through the early years of life. The effects of learning and maturation are notoriously hard to disentangle.

I do not think that these difficulties are sufficiently severe to invalidate the use of the
architecture concept in the case of the brain. Clearly, a great deal of the brain’s organization is fixed and genetically specified. As for the effects of maturation, we can largely avoid the problem by concentrating on the adult brain, in which learning is presumably the only normal cause of significant changes in physical structure (except, perhaps, for a small amount of neuron shrinkage and death). Alternatively, following Newell (1990), we can accept that the architecture itself can undergo very slow change over a relatively long time scale.

Cognition is conventionally divided into several different faculties, such as visual perception, auditory perception, memory, imagination, language understanding, language generation, reasoning, and the like. The classification is hierarchical: visual perception, for example, can be decomposed into such faculties as shape perception, colour perception, motion perception, and so on. Of course, the correct classification of cognition into natural-kind faculties is a theoretical issue on which no consensus exists, and it is possible that our folk-psychological classification is wide of the mark (Churchland, 1986), but it is clear that some such division must exist. Two broad types of cognitive architecture can be distinguished with respect to such a classification. A uniform architecture is one in which very similar computational mechanisms are used in all faculties. A heterogeneous architecture is one in which a fundamentally different system of mechanisms is used in each faculty. It is important to distinguish heterogeneity from modularity: a modular architecture is one that consists of a number of separate processors, called modules, each of which is dedicated to one faculty. A heterogeneous architecture will almost certainly be modular, because it makes sense to isolate the specialized mechanisms in separate modules. However, a modular architecture is not necessarily heterogeneous. A modular architecture is also uniform if each of the separate processors uses the same internal mechanisms; that is, different modules are responsible for different faculties, but they all work in basically the same way internally, using the same kinds of representation and the same set of basic operations.

Figure 2.1: Section through the midplane of the human brain showing the medial surface of the right cerebral hemisphere. The cerebral cortex is stippled. (Adapted from R. Nieuwenhuys, J. Voogd and C. van Huijzen, 1978, The Human Central Nervous System: A Synopsis and Atlas, Berlin: Springer, Fig. 12. Copyright 1978 Springer-Verlag.)
Although the terminology has changed over the years, the issue of whether the human cognitive architecture is uniform or heterogeneous has been debated throughout the history of psychology (see Fodor, 1983). In our own time, the heterogeneous view dominates cognitive science. Chomsky has argued forcefully that our capacity for language is subserved by specialized mechanisms, unique to our species (e.g., Chomsky, 1980). Marr (1982) viewed the visual system as a modular, heterogeneous architecture, with specialized mechanisms for deriving shape from different sources (stereopsis, motion, surface texture, and so on). This is a perspective that is still almost universally accepted in the field of visual perception. Tulving (1983) has divided the faculty of long-term memory into procedural, episodic and semantic sub-systems, each having its own operating characteristics. These are merely the most explicit statements of a general perspective that is widely accepted.

A few theorists have argued for a limited amount of uniformity in the cognitive architecture. Anderson (1983) and Newell (1990) have accepted that ‘peripheral’ faculties like perception and motor control have their own specialized mechanisms but that ‘central’ faculties like problem-solving and semantic memory are subserved by a uniform architecture. Even Fodor (1983) has some sympathy with this perspective, although, paying homage to Chomsky, he excludes language from central cognition.

In this thesis I argue that the human cognitive architecture as a whole is indeed heterogeneous, but that a very substantial part of it is approximately uniform. Specifically, the uniform part of the architecture is that realized in the cerebral neocortex, whereas the specialized systems are those realized in the other parts of the brain. I do not claim that the uniformity of the cortical architecture is perfect—there is a significant amount of specialization from one faculty to the next—but the degree of uniformity is sufficient for a general account of the architecture to be possible. A general theory of the computational mechanisms provided by the neocortex could form the basis of a unified theory of a large part of human cognition.
2.3 The Importance of the Cortex in Cognition

The human cerebral cortex consists of two convoluted sheets of grey matter, each about 1200 cm² in area (Haug, 1987) and a few millimetres in thickness, one on the surface of each cerebral hemisphere. The cortex is illustrated in relation to the rest of the brain in Figures 2.1 and 2.2. When viewed in cross-section, the cortex is seen to have a laminar structure, which is caused by variations in the sizes and densities of cells and fibres at different depths below the cortical surface. Two broad types of cortex are generally distinguished: three-layered allocortex, which includes olfactory cortex (a specialized processor for the sense of smell) and the hippocampus; and six-layered neocortex. Neocortex is the most phylogenetically-recent type, being unique to mammals, and has expanded dramatically over the course of primate evolution. In humans, the neocortex dwarfs the rest of the brain and is estimated to contain over $10^{10}$ neurons (Haug, 1987). In the remainder of the thesis I will adopt the usual convention of using the simple term cortex to refer exclusively to the neocortex, except in contexts where confusion might arise.

On the basis of differences in thickness, laminar structure, functional properties of neurons, and other criteria, the cortical sheet can be parcellated into many distinct fields or areas. Figure 2.3 shows one widely-used scheme of parcellation of the human cortex, due to Brodmann (1909). Brodmann’s map is based on rather crude anatomical criteria; more detailed studies will certainly show many of his areas to consist of several smaller areas, as they have in the monkey. (Amazingly little is known about human neuroanatomy. Available information on the human cortex seems entirely consistent with data from other primates; the parcellation of the cortex into areas is likely to be the single most important species-difference.) Physiological studies of animals and neuropsychological investigations of humans demonstrate conclusively that different areas are dedicated to different cognitive faculties. With regard to the intrinsic structure of the cortex, however, there is remarkably little variation from one area to another. That is, there appears to be a basic pattern of neuronal organization which is common to all cortical areas. In the following section I review the evidence for this claim as it currently stands. Before doing this, however, it is worthwhile briefly reviewing the vital importance of the cortex in human cognition.

**Perception.** Large regions of the cortex are devoted to perception. Vision has the lion’s share: it has been estimated that 55% of the macaque monkey’s cortex is directly involved in visual perception (Felleman & Van Essen, 1991). In humans the fraction is presumably smaller, but visual cortex still makes up the whole of the occipital lobe as well as large regions of the parietal and temporal lobes. Audition, somesthesis and taste also have their cortical territories. (Smell, as noted above, has its own specialized mechanisms in the allocortex.) Much more is known about cortical processing in perception, especially visual perception, than in any other domain of psychology. In particular, the primary visual area (Brodmann’s area 17, now usually called V1), which is the first stage in cortical visual processing, has been the focus of most experimental work on the cortex since the pioneering studies of Hubel and Wiesel (1962, 1968). From the perspective of the present thesis the concentration on V1 is somewhat unfortunate, because there are several indications (see below) that this area is particularly highly specialized, and is therefore a poor advertisement for the hypothesis of basic cortical uniformity.

**Motor control.** The cortex is also heavily involved in the production of movement, although a large part of the motor system is subcortical—for example, the basal ganglia, cerebellum, and various nuclei in the brain stem (Figs. 2.1, 2.2). An area called the primary motor cortex in the precentral gyrus of the frontal lobe (Brodmann’s area 4) has traditionally been considered the main cortical centre for motor control, since it is particularly easy to elicit bodily movements by electrical stimulation in this area. However, it is now realized that practically the whole of the cortex projects fibres into the subcortical motor system. Even the primary visual area, for example, has a substantial output to the superior colliculus, one of the functions of which is controlling the direction of gaze (Wurtz & Albano, 1980). Hence, it is probably too simplistic to regard only certain areas as motor areas; all parts of the cortex are involved in motor control (Diamond, 1979).

**Memory.** The cortex is also believed to be the major site of memory storage. It is well-known that an allocortical structure, the hippocampus (Fig. 2.2), is crucially involved in the establishment of at least some kinds of long-term memory, since bilateral damage to the hippocampus causes profound
anterograde amnesia (reviewed by Squire, 1992). However, memories from before the time of damage often remain largely intact, implying that the hippocampus is not itself the site of storage. The memories are presumably held in neocortex, probably in the temporal lobe.

**Language.** It has been known since the early studies of aphasia by Broca and Wernicke that the cortex is crucially important for language. The major cortical regions that are important for language are Broca’s area (approximately Brodmann’s area 44) and Wernicke’s area (roughly posterior area 22). (Again, it is very likely that these are not single areas but complexes of several areas.) Language is processed mainly by the cortex of the left cerebral hemisphere in most individuals.

**Thought.** Finally, the cortex is undoubtedly the main substrate of such higher thought processes as imagery, reasoning, problem solving and planning. It is generally assumed that the prefrontal cortex, the anterior part of the frontal lobe, is the most important cortical region for these central faculties. The prefrontal region comprises nearly a third of the cortex in humans, a much larger fraction than in any other species (Fuster, 1989). Central thought processes are almost certainly not confined to prefrontal cortex, however; there is good evidence, for example, that visual imagery

---

**Figure 2.3: Brodmann’s division of the human cerebral cortex into areas.**

(A) Lateral view of the right hemisphere.  
(B) Section through the midplane. Because the human cortex is very convoluted the map gives a poor impression of the relative surface areas of the fields. (Adapted from R. Nieuwenhuyys, J. Voogd and C. van Huijzen, 1978, *The Human Central Nervous System: A Synopsis and Atlas*, Berlin: Springer, Fig. 5. Copyright 1978 Springer-Verlag.)
involves certain parts of the visual cortex (Farah, 1988). There is little empirical data on the physiological basis of these central faculties, in part because of the very limited capabilities of experimental animals. However, progress is being made with humans using evidence from non-invasive experimental techniques and from neuropsychology (e.g., Shallice, 1988).

In summary, the cortex plays a large and probably dominant role in most of the faculties of cognition. The observation of a substantial degree of uniformity in the neural circuitry of the cortex is therefore a surprising and significant one. In the following section the evidence for anatomical uniformity is considered in some detail.

2.4 Anatomical Uniformity of the Cortex

To claim that the anatomy of the cortex is uniform is not to say that it is simple or unstructured—a sort of porridge of cytoplasm. On the contrary, the cortex has a highly intricate internal structure, probably greater in complexity than any other part of the brain. However, the cortex does appear to be uniform in the sense that the neural circuitry beneath a small patch of the cortical surface is very similar across the whole cortical sheet.

Layered structure. As already mentioned, neocortex is defined by its highly laminated structure. Six layers are conventionally distinguished, numbered 1 to 6 from the cortical surface to the underlying white matter, although the boundaries between layers are often not distinct. Figure 2.4 gives an impression of the variations in cell size and density that define the cortical layers. There are

Figure 2.4: Layered structure of the neocortex as revealed by three different tissue stains. The Golgi stain reveals the structure of a small fraction of the neurons; the Nissl stain marks the somas of all the neurons; and the Weigert stain shows myelinated axons. (From S. W. Ranson and S. L. Clark, 1959, The Anatomy of the Nervous System: Its Development and Function, 10th Edition, Philadelphia: W. B. Saunders, Fig. 252. Copyright 1959 W. B. Saunders Company. Reproduced with permission of Elsevier Ltd. Originally from Brodmann, 1909.)
some differences between cortical areas in the detailed laminar structure. The primary visual cortex, V1, is particularly complex in this regard; in the monkey, layer 4 has four distinct sub-layers in this area (Lund, 1981). It should be noted that variations in laminar structure do not necessarily indicate differences in computational mechanisms, although they might do to some extent. For example, there is a correlation between the size of a neuron’s soma and the length of its axon, and so variations in neuron size between areas may be caused by differences in the targets of the outputs of those areas rather than differences in their intrinsic organization.

**Pyramidal neurons.** Although no two neurons are exactly alike, it is possible to classify cortical neurons into a number of reasonably well-defined categories. The main type of neuron in the cortex, making up at least three quarters of the total population, is the pyramidal neuron (reviewed by Feldman, 1984), so called because of the roughly pyramidal shape of its soma. Its major characteristics are a large apical dendrite which arises from the apex of the soma and ascends vertically through the cortex, generally reaching layer 1; a few short basal dendrites in the vicinity of the soma; and an axon which descends through the cortex into the underlying white matter, travelling to another cortical area or to another part of the nervous system. Examples are illustrated in Figure 2.5 (C–E). All pyramidal cell dendrites are densely covered with small protuberances called spines, on which input synapses are preferentially located. Before leaving the cortex, the axon generally gives off a few collateral branches which radiate out from the parent axon, for up to several millimetres, and make excitatory synaptic connections with neighbouring neurons. Although pyramidal neurons vary in size and morphology, they are found in all cortical areas of all mammalian species, and virtually all of the fibres that carry information out of the cortex are axons of pyramidal cells. This cell type is therefore the basic component of the cortical architecture.

**Nonpyramidal neurons.** The general class of nonpyramidal neurons can be subdivided into several categories (reviewed by Fairén, DeFelipe & Regidor, 1984), but they all share the characteristics that they lack apical dendrites and have axons that remain entirely within the cortex, forming synapses with neighbouring neurons no more than a millimetre away. Some examples are illustrated in Figures 2.5 and 2.6. Most types of nonpyramidal neuron have an inhibitory effect on the neurons they contact, since most of them release the inhibitory neurotransmitter γ-aminobutyric acid (GABA) (Ribak, 1978; Freund, Martin, Smith & Somogyi, 1983; Peters & Saint Marie, 1984). It is possible that not all types of nonpyramidal cell are present in all mammalian species, and those that are present seem to differ in the details of their morphology from one species to the next (Fairén et al., 1984). However, within any one species, those nonpyramidal neurons that are present seem to occur in all areas of the cortex, although they may be more common in some areas than in others. The major exception to this general rule is the spiny stellate neuron (Figs. 2.5B, 2.6F), which is found exclusively in layer 4 of certain areas (notably the primary sensory areas) of certain species (Lund, 1984). This particular cell class might therefore seem to be an objection to the hypothesis of basic anatomical uniformity in the cortex. However, it is possible that the spiny stellate neuron is actually a specialization of the pyramidal neuron (Lund, 1984). There are a number of ways in which spiny stellate cells are similar to pyramidal neurons but different from other nonpyramidal types. First, they have dendrites that are thickly covered with spines, whereas other nonpyramidal neurons have smooth or sparsely spinous dendrites. Second, the distribution of different types of synapse over the dendrites and somas of spiny stellate cells has a pattern similar to that of pyramidal cells, but different from that of other nonpyramidal cell types (LeVay, 1973; Saint Marie & Peters, 1985). Third, intermediate cell types exist: there are spiny stellate neurons in area V1 of primates with axons that leave the cortex, and there are a few pyramidal neurons with poorly developed apical dendrites and axons that remain entirely within the cortex (Lund, 1981). Finally, spiny stellate cells, like pyramidal cells, are thought to have an excitatory function.

**Neuron numbers.** Although the packing density of neurons differs quite substantially from one region of the cortex to another, the number of neurons beneath a unit area of the cortical surface is more uniform. In one study, Rockel, Hiorns and Powell (1980) counted the number of neurons beneath a surface area of 25 × 30 µm (after tissue shrinkage due to processing) from several different cortical areas of several different species, including man. In all cases, the number was about 110, except in area V1 of primates, where it was about 270. Subsequent studies (e.g., Beaulieu & Colonnier, 1989) have supported the general conclusion that there is approximate uniformity in the
Figure 2.5: Some components of the primary visual cortex of the cat. The thicker cables are dendrites, the thinner ones are axons and axon collaterals. (A) Two input fibres from the lateral geniculate nucleus, terminating mainly in layer 4. (B) Spiny stellate cells. (C–E) Pyramidal cells. (F) Smooth stellate cell. (G) Bipolar cell. (H) Basket cell. The basket cell’s axon is confined to layer 4 but appears more widespread because of the plane of the section. (From C. D. Gilbert, 1983, Microcircuitry of the visual cortex, *Annual Review of Neuroscience* 6: 217–247, Fig. 2. Copyright 1983 Annual Reviews Inc. Reproduced with permission.)
Figure 2.6: Nonpyramidal neurons in monkey sensorimotor cortex. (A) Arcade cell. (B) Double bouquet cell. (C) Small basket cell. (D) Chandelier cells. (E) Bipolar cell. (F) Spiny stellate cell. (G) Spiderweb cell. (H) Large basket cells. (From E. G. Jones, 1986, Connectivity of the primate sensory-motor cortex, in E. G. Jones and A. Peters (eds.), *Cerebral Cortex*, Vol. 5, pp. 113–183, New York: Plenum Press, Fig. 26. Copyright 1986 Plenum Press. Reproduced with permission of Springer-Verlag.)
number of neurons per unit surface area over many regions of the cortex, although there is probably greater variation than Rockel et al. found. It is also likely that the proportions of the various neuron types are fairly uniform. Hendry, Schwark, Jones and Yan (1987) studied the proportion of neurons that were immunoreactive for the neurotransmitter GABA (which is released by most of the nonpyramidal cell types) in several areas of monkey cortex and found that it was about 25%; again, area V1 was the main exception, where the proportion was only 20%. The exceptional nature of V1 may be caused by an unusually high number of small spiny stellate neurons in layer 4.

**Connections with the thalamus.** Most of the information that reaches the cortex is relayed through a subcortical structure, the thalamus (see Jones, 1985, for a comprehensive review). The thalamus consists of a number of distinct cell groups or nuclei, illustrated in Figure 2.7, and the fibres from each nucleus connect with particular areas of the cortex, summarized in Table 2.1. For example, visual information from the retina is relayed through the lateral geniculate nucleus to the visual areas of the cortex, principally to V1. Each thalamic fibre terminates as a localized cluster of collaterals (Fig. 2.5A), forming excitatory synapses with all types of neurons (White, 1986), mainly in the middle layers (layer 4 and lower layer 3) of the cortex (Lorente de Nó, 1949; Jones & Burton, 1976; Jones, 1985). A few small thalamic nuclei send fibres to layer 1 and to the deep layers of the cortex; these projections appear to be rather diffuse and widespread, and have been called ‘nonspecific’ to distinguish them from the more substantial inputs to the middle layers (Lorente de Nó, 1949; Jones, 1985; Herkenham, 1986). The thalamus also receives connections back from the cortex; there seems to be a general rule that each cortical area sends fibres back to the thalamic nuclei from which it receives inputs. Throughout the cortex, these connections arise from pyramidal cells in the deep layers, predominantly layer 6 (Gilbert & Kelly, 1975; Jones & Wise, 1977; Jones, 1981, 1984).

**Connections with the claustrum.** The claustrum is another subcortical structure (Fig. 2.2) with which most areas of the cortex are connected (Olson & Graybiel, 1980; Sherk, 1986). These connections appear to have a similar laminar pattern to the thalamic connections: claustral inputs arrive mainly in layer 4 and lower layer 3, with some inputs also to layer 1 and layer 6, and cortical outputs to the claustrum arise from the deep layers (LeVay & Sherk, 1981; Clascá, Avendaño, Román-Guindo, Llamas & Reinoso-Suárez, 1992). Unlike the thalamus, however, the claustrum does not receive any substantial inputs other than that which it receives from the cortex (Sherk, 1986).

**Other subcortical outputs.** All cortical areas send outputs to a variety of other subcortical structures. These outputs always originate from the deep layers, mainly from layer 5 (Gilbert & Kelly,
1975; Jones & Wise, 1977; Jones, 1981, 1984). Although there is a great deal of variation between areas in the targets of these projections, it is not unreasonable to regard them generally as motor outputs, in that the target structures are known to be heavily involved in motor control—for example, the basal ganglia, pons, superior colliculus and spinal cord.

**Connections between cortical areas.** Every cortical area is connected with several (though by no means all) other cortical areas, both within the same cerebral hemisphere and in the opposite hemisphere; these are called corticocortical connections. One of the most significant discoveries in neuroanatomy in recent years is that the connections within each hemisphere fall into two main groups (Rockland & Pandya, 1979; Van Essen & Maunsell, 1983; Felleman & Van Essen, 1991). Connections of the feedforward type originate mostly from neurons in layers 2–4 of the source area, and terminate in layer 4 and lower layer 3 of the destination area—that is, the same pattern of termination as the specific thalamic inputs. Those of the feedback type originate mostly from neurons in the deep layers of the source area and terminate outside layer 4 (mainly in layers 1 and 6) of the destination area. This pattern is illustrated schematically in Figure 2.8. A small proportion of corticocortical connections do not appear to have any particular laminar distribution; these may be called lateral pathways. The evidence suggests that every feedforward projection is reciprocated by a feedback projection, and vice versa. Furthermore, it has been suggested that the areas within each hemisphere can be placed in a hierarchy, such that all the pathways going up the hierarchy are of the feedforward type and all those going down the hierarchy are feedback; the few lateral pathways connect areas at the same level in the hierarchy (Van Essen & Maunsell, 1983). The evidence for the validity of this hierarchical organization comes mainly from the connections between the visual areas of the macaque monkey (Van Essen, 1985; Zeki & Shipp, 1988; Felleman & Van Essen, 1991; Kaas & Krubitzer, 1991; Young, 1992) which are illustrated in Figure 2.9. There is accumulating evidence that the scheme is also applicable, at least approximately, to other species and other regions of the cortex (Galaburda & Pandya, 1983; Symonds & Rosenquist, 1984; Pandya & Yeterian, 1985; Barbas, 1986; Friedman, Murray, O’Neill & Mishkin, 1986; Burkhalter & Bernardo, 1989; Felleman & Van Essen, 1991; Rouiller, Simm, Villa, de Ribaupierre & de Ribaupierre, 1991; Coogan & Burkhalter, 1993). It remains to be seen whether all cortical areas can be placed into one big hierarchy.

<table>
<thead>
<tr>
<th>Nucleus</th>
<th>Subcortical input</th>
<th>Associated cortex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral geniculate</td>
<td>Retina</td>
<td>Primary visual area</td>
</tr>
<tr>
<td>Pulvinar</td>
<td>Superior colliculus</td>
<td>Visual cortex</td>
</tr>
<tr>
<td>Medial geniculate</td>
<td>Superior colliculus</td>
<td>Auditory cortex</td>
</tr>
<tr>
<td>Ventral posterior lateral,</td>
<td>Medial &amp; trigeminal lemnisci, spinothalamic tract</td>
<td>Somatosensory cortex</td>
</tr>
<tr>
<td>ventral posterior medial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral posterior</td>
<td>?</td>
<td>Somatosensory area 5</td>
</tr>
<tr>
<td>Ventral lateral (posterior)</td>
<td>Cerebellum, vestibular nucleus</td>
<td>Primary motor area</td>
</tr>
<tr>
<td>Ventral lateral (anterior)</td>
<td>Globus pallidus</td>
<td>Non-primary motor cortex</td>
</tr>
<tr>
<td>Ventral medial (principal)</td>
<td>Substantia nigra</td>
<td>Non-primary motor cortex</td>
</tr>
<tr>
<td>Ventral medial (basal)</td>
<td>Gustatory nucleus</td>
<td>Gustatory area</td>
</tr>
<tr>
<td>Ventral anterior</td>
<td>Globus pallidus, substantia nigra</td>
<td>Prefrontal cortex</td>
</tr>
<tr>
<td>Medial dorsal</td>
<td>Olfactory cortex, amygdala</td>
<td>Prefrontal cortex</td>
</tr>
<tr>
<td>Medial ventral</td>
<td>Fornix</td>
<td>Hippocampus</td>
</tr>
<tr>
<td>Lateral dorsal</td>
<td>Fornix</td>
<td>Cingulate cortex</td>
</tr>
<tr>
<td>Anterior complex</td>
<td>Mammillary nuclei</td>
<td>Cingulate and parahippocampal cortex</td>
</tr>
<tr>
<td>Posterior complex</td>
<td>Superior colliculus, spinothalamic</td>
<td>Insular cortex</td>
</tr>
</tbody>
</table>

Table 2.1: Summary of the main nuclei of the thalamus, their subcortical inputs and their associated areas of cortex. Omitted are the reticular nucleus, which connects with other thalamic nuclei, and the intralaminar nuclei, which send ‘nonspecific’ connections to the cortex. (Nomenclature and data mainly from Jones, 1985.)
Commissural connections. Nearly all cortical areas also have connections to and from areas in the opposite cerebral hemisphere, called commissural connections (Innocenti, 1986). Usually, each area is connected with its corresponding, symmetrical area, and with a few others. These fibres make up a massive tract joining the two hemispheres, the corpus callosum (Figs. 2.1, 2.2). It has been suggested that commissural connections can be classified as feedforward, feedback and lateral in the same way as the corticocortical connections within each hemisphere, with the difference that the commissural feedback fibres originate only from the upper layers of the cortex (Kennedy, Meissirel & Dehay, 1991).

Connections with the limbic system. Certain cortical areas have connections with a number of structures in the forebrain collectively known as the limbic system, in particular with the amygdala and the hippocampus—the latter has already been mentioned as an allocortical structure known to be important for memory. Actually, most connections between the neocortex and the hippocampus are indirect: certain areas connect with parahippocampal cortex, which in turn connects with the hippocampus (Van Hoesen, 1982; Pandya & Yeterian, 1985). The parahippocampal cortex has an internal structure that seems to be transitional between allocortex and neocortex. It is interesting that connections from the neocortex to the parahippocampal cortex and amygdala are of the feedforward pattern: they arise mostly from neurons in layer 3 (Insausti, Amaral & Cowan, 1987; Squire, Shimamura & Amaral, 1989; Amaral, Price, Pitkänen & Carmichael, 1992; Suzuki & Amaral, 1994). The reciprocal connections back to the neocortex are of the feedback type: the fibres terminate mainly in layer 1 and to a lesser extent in the deep layers (Witter & Groenewegen, 1986; Amaral et al., 1992; Suzuki & Amaral, 1994). In higher species, limbic connections are largely restricted to cortical areas that are far removed from the primary sensory and motor areas (Van Hoesen, 1982; Pandya & Yeterian, 1985; Insausti et al., 1987; Squire et al., 1989; Amaral et al., 1992). This raises the intriguing possibility that the limbic system lies at the top of the hypothetical hierarchy of cortical areas. There is much evidence that the limbic system is important for instinctive and emotional behaviour (Isaacson, 1982).

Nonspecific fibre systems. In addition to the thalamic, claustral, corticocortical and limbic inputs, every cortical area is also innervated by a number of fibre systems originating from relatively small clusters of neurons in the brain stem and the basal forebrain, each fibre system releasing a particular neurotransmitter (reviewed by Eckenstein & Baughman, 1987; Fallon & Loughlin, 1987; Foote & Morrison, 1987). Details are given in Table 2.2. Unlike the specific input fibres described above, each of which only makes contact with a localized cluster of cortical neurons, these
Figure 2.9: Areas and connections of the visual cortex in one cerebral hemisphere of the macaque monkey. Most if not all of the pathways are reciprocal: pathways between areas at different levels in the diagram consist of feedforward and feedback connections, and pathways between areas at the same level are of the lateral type. The pathways going off the left edge of the diagram are connections with sensorimotor cortex, itself hierarchically organized. The lateral geniculate nucleus (LGN) of the thalamus is shown at the bottom of the hierarchy, and the hippocampus (HC) at the top. Some subdivisions of the LGN and of areas V1 and V2 are also shown. (From D. C. Van Essen, D. J. Felleman, E. A. De Yoe, J. Olavarria and J. Knierim, 1990, Modular and hierarchical organization of extrastriate visual cortex in the macaque monkey, *Cold Spring Harbor Symposia on Quantitative Biology* 55: 679–696, Fig. 8. Copyright 1990 Cold Spring Harbor Laboratory Press. Reproduced with permission.)
nonspecific fibres are spread extremely widely throughout the cortex, often not even respecting area boundaries. They are present in all areas, although the density of their innervation differs from layer to layer and from area to area. The widespread nature of these fibre systems suggests that they in some way ‘modulate’ the processing of large populations of neurons. For example, they are involved in the control of arousal and the sleep/wake cycle.

**Intrinsic circuitry.** Detailed information about the circuitry within any cortical area is very scanty because of the immense technical difficulties involved in obtaining quantitative data on the connectivity of neurons. The best studied area is, as usual, area V1 of cats and monkeys, in which the circuitry is very complicated (for reviews see Lund, 1981, 1988; Gilbert, 1983; Martin, 1984; Valverde, 1985, 1991). However, I have attempted to abstract a simplified, basic scheme of information flow, which I will describe here. What little is known of the intrinsic organization of other areas and other species does seem to be roughly consistent with this basic pattern (Lund, Hendrickson, Ogren & Tobin, 1981; Mitani, Shimokouchi, Itoh, Nomura, Kudo & Mizuno, 1985; Jones, 1986; White & Keller, 1987; Ghosh, Fyffe & Porter, 1988; Burkhalter, 1989; Schwark & Jones, 1989; Levitt, Lewis, Yoshioka & Lund, 1993). The two main sources of input to any cortical area are feedforward connections (from the thalamus, claustrum, and from other cortical areas lower in the hierarchy) which terminate in the middle layers, and feedback connections (from the limbic system, from areas higher in the hierarchy, and from certain thalamic sources) which terminate in layer 1 and layer 6. Information is processed in layers 2–4 by circuitry consisting of excitatory and inhibitory components; the former are the collaterals of pyramidal and spiny stellate neurons, the latter the collaterals of other nonpyramidal cell types. The output generated here is the feedforward projection to higher cortical areas and to the limbic system. This output is also copied into layer 5 by means of collaterals from the axons of layer 2–4 pyramidal cells. Circuitry in layer 5 generates the output to subcortical motor centres. This output in turn is copied into layer 6 by means of further collaterals. The output from this stage is the feedback projection to the thalamus, claustrum and lower cortical areas. Finally, collaterals from layer 6 close the loop by projecting back up to the middle layers.

In summary, the structure of the neural circuitry underlying a small patch of the cortical surface appears to be roughly the same across the whole cortical sheet, in the following respects: (1) it has a basic six-layered structure; (2) it is composed of the same types of neuron; (3) it contains the same number of neurons and the same proportions of neuron types; (4) it has the same laminar distribution of the different types of inputs and outputs; (5) the same nonspecific fibre systems are present; and, more tentatively, (6) it has the same basic pattern of intrinsic circuitry. These statements are only approximately true: there are significant differences in structure from one area to another. As noted above, area V1 in higher species appears to be particularly specialized. Nevertheless, the degree of uniformity is remarkable. We badly need comparative studies of several different areas, especially regarding the detailed intrinsic connectivity, in order to better evaluate the uniformity hypothesis.

There are other sheet-like brain structures besides the cerebral neocortex in which a basic neural circuit is present throughout the sheet. The best example is the cortex of the cerebellum (Fig. 2.1), which is even more uniform in this sense than the neocortex (Eccles, Ito & Szentágothai, 1967). What makes the neocortex unique among these cases is that different areas of the sheet are involved in

<table>
<thead>
<tr>
<th>Source</th>
<th>Neurotransmitter</th>
<th>Distribution in cortex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantia nigra (pars compacta), ventral tegmentum</td>
<td>Dopamine</td>
<td>Dense in motor cortex, sparse in primary sensory areas</td>
</tr>
<tr>
<td>Locus coeruleus</td>
<td>Norepinephrine</td>
<td>Dense in sensorimotor cortex, sparse in temporal lobe</td>
</tr>
<tr>
<td>Raphe nuclei</td>
<td>Serotonin</td>
<td>Inverse of norepinephrine system</td>
</tr>
<tr>
<td>Nucleus basalis of Meynert</td>
<td>Acetylcholine</td>
<td>Most dense in paralimbic cortex</td>
</tr>
</tbody>
</table>

Table 2.2: Fibre systems that originate from certain small nuclei and innervate the whole cortex, releasing particular neurotransmitters. Regional variations in the density of the innervation are listed for monkeys. (Mainly from Foote & Morrison, 1987.)
a very diverse range of faculties that seem (from the perspective of both folk psychology and contemporary cognitive psychology) to be entirely different from one another. This is not the case in the cerebellum, for example, where it appears that the whole sheet is performing basically the same function, namely, the coordination of skilled movements (Brooks & Thach, 1981).

### 2.5 The Fundamental Hypothesis

The cortex is involved in an extremely broad range of faculties, with different faculties localized in different cortical areas. However, there is a substantial amount of anatomical uniformity, with the structure of the neural circuitry being very similar throughout the cortex. The uniformity is certainly not perfect; there are significant variations from area to area in such parameters as laminar organization and neuron numbers. Nevertheless, it is reasonable to suppose that the specialized features of particular areas (such as the spiny stellate cells in certain areas) are merely variations on a uniform underlying theme. We can abstract a canonical cortical circuit (I take the phrase from Douglas, Martin & Whitteridge, 1989), of which the actual circuitry in any particular cortical area is an approximation. The canonical circuit is an idealization that incorporates those aspects of the neural circuitry that are common to most or all areas.

I propose that the canonical cortical circuit provides a basic system of computational mechanisms that are applied in each of the faculties of cognition in which the cortex is involved—that is, that the cortex is a uniform cognitive architecture, in the sense defined earlier. The variations in anatomy from area to area presumably reflect variations in computational mechanisms, so the cortex is not a pure uniform architecture. However, it should be possible to formulate a general model of the mechanisms realized by the canonical circuit, a model which is a close approximation to the actual mechanisms used in any area. Such a general model of the architecture would not itself be a general theory of cognition, for two reasons. First, as pointed out before (p. 15), a general theory of cognition requires not only a model of the architecture but also a theory of how the mechanisms supplied by the architecture get applied in particular faculties—shape perception, language comprehension, speech production, visual imagery, and all the other important domains of psychological enquiry. Second, the cognitive system as a whole is more than just the cortex; the rest of the nervous system is also important. One cannot have a complete theory of vision without understanding the retina, or of memory without understanding the hippocampus. Nevertheless, the proposal is that a general model of the cortical architecture would provide an excellent foundation on which to build a unified theory of cognition. The model would be applicable not just to humans but to mammals generally, because the canonical cortical circuit appears to have been elaborated relatively little over the course of mammalian evolution. The human cortex seems to differ from that of the rat primarily in its vastly greater surface area and neuronal population, and in the number of distinct areas it contains.

It might be argued that the canonical circuit is of no real interest because it is merely an idealized abstraction, not actually existing in any particular cortical area. However, science is replete with examples of general models that have been abstracted from particular cases. Probably the best example is the ‘ideal cell’ of biochemistry. There are, of course, significant differences between the cells that make up the brain, kidney, liver, and muscle. But, to a remarkable extent, cell structure and the processes of metabolism, mitosis, and so on are common to all the cells in the body—and, for that matter, common to all plants and animals. This uniformity enables biochemists to routinely deal with an ideal cell, a general model which incorporates the common structures and processes. In the same way, I suggest, a general model of the canonical circuit of the cortex would be of great theoretical utility. This is not to deny, of course, that it is also important to understand the ways in which the circuit is specialized in particular cortical areas, just as it is important to understand the specialized biochemical characteristics that differentiate kidney cells from liver cells.

Chapter 3 introduces a theory of the canonical circuit. The details of this model are not directly relevant to the general argument of the present chapter, but I do wish to make a few suggestions here regarding some of the basic functions that the canonical circuit performs. In primitive mammalian species alive today, the whole of the cortex is dedicated to perceptual-motor processing (Lende & Sadler, 1967; Diamond & Hall, 1969)—activities that are regarded as ‘peripheral’ in human cognition—and it is reasonable to suppose that the cortex evolved in early mammals because it enhanced their abilities in these faculties. It is only recently, in evolutionary terms, that the cortex was
extended, apparently without major changes to the basic circuit, to support central faculties like language processing and complex problem solving. It is very recently indeed that it has been applied to such tasks as playing chess and doing arithmetic. This suggests (see below) that the cortical architecture is extremely well-adapted for perceptual-motor processing, but that it is comparatively inefficient for more central tasks. What basic functions should a perceptual-motor processor be performing? I propose that the single most important function of the cortical architecture is pattern recognition, the essence of perception. Pattern recognition presumably involves performing some kind of matching operation between the input information and stored representations. The fact that all cortical areas send output to subcortical motor centres suggests that a second basic function is motor production, the generation of output appropriate for patterns that have been recognized in the input. A third important requirement is plasticity, the ability to learn from experience. The following chapter presents specific proposals as to how these functions are realized in the canonical circuit.

Let us now turn to central cognition. The idea that the cortex is basically a perceptual-motor processor perhaps provides the reason why our performance in many central activities, like choosing a move in chess, solving an algebra problem, or writing a book, tends to be slow, effortful and error-prone, whereas our perceptual and motor abilities are fast and fluent. The canonical circuit of the cortex may be poorly-adapted and therefore comparatively inefficient for central cognition. A relevant observation here is the gradual transition from an ‘analytical’ to a ‘perceptual’ mode of cognition as we gain skill in a particular domain. The best-studied example of this phenomenon is in the domain of chess. It has been shown that chess Masters do not typically analyse more potential moves than much weaker players; but the moves that they select for analysis are all good moves, whereas weak players waste time analysing bad moves (de Groot, 1965). The Master’s superior performance appears to rely on the ability to recognize particular configurations of chess pieces, by matching these patterns against representations of such configurations, or chunks, stored in long-term memory (Chase & Simon, 1973a, 1973b). It has been estimated that a Master has a repertoire of between $10^3$ and $10^4$ chunks, which is about the same size as the recognition vocabulary of a good reader (Simon & Gilmartin, 1973). ‘Chunking’ might be a quite general method underlying skilled performance in many faculties, including central ones (Newell & Rosenbloom, 1981). Thus, in chess and other cognitive skills, the novice is forced to rely on conscious, serial thought processes that are slow and error-prone, whereas the expert is able to supplement these thought processes with a highly efficient, memory-based pattern recognition capability. I propose that this results from the architecture of the cortex being very well-adapted for pattern recognition, but very inefficient for serial, symbolic problem-solving algorithms. Just the opposite is true of the architecture of conventional digital computers.

The idea that a broad range of cognitive faculties are underlain by a roughly uniform architecture has two general implications for research in cognitive psychology. As pointed out earlier, it is generally assumed that the human cognitive architecture is heterogeneous—that the mechanisms used in one faculty have little in common with the mechanisms used in any other faculty. This assumption encourages the Balkanization of cognitive psychology: miniature theories are proposed and developed to account for performance in particular task domains (priming, visual search, verbal learning, speech perception, syllogistic reasoning, and so on) but little effort is made to link these theories into a single framework, with a few noteworthy exceptions (Anderson, 1983; Newell, 1990). If the architecture is, in fact, uniform, then cognitive psychology should be seeking to account for results from diverse experimental tasks using general mechanisms. The second point is rather more subtle. Cognitive psychologists tend to take a ‘design-oriented’ approach to theory construction: computational mechanisms are proposed that perform the task in question in a highly efficient and elegant manner, as if they had been designed for the task by an intelligent engineer. Perhaps the finest example of this approach is the elaborate problem-solving mechanism of the Soar architecture (Laird, Newell & Rosenbloom, 1987). Clark (1987) has argued that the design-oriented approach is a good one when applied to basic cognitive processes that we share with other species and which have high survival value, because natural selection will have ensured that the underlying mechanisms are indeed highly efficient. However, when considering more recently-evolved capacities, such as many of those in central cognition, we should be aware that the mechanisms may be rather inefficient and lacking in elegance. This is because the mechanisms in question will have evolved originally for some quite different or more primitive purpose, only recently being adapted to their present function. If this is the
case, the design-oriented approach is unlikely to arrive at a good characterization of them. Applying Clark’s argument to the cortex, the design-oriented approach to understanding the architecture is most likely to succeed in those cognitive faculties for which the canonical circuit first evolved, namely, perception and motor control. When we have gained an understanding of how the cortex works in these peripheral faculties, we will be in a much better position to tackle the central faculties, for which the architecture is probably much less well-adapted. This approach can also be justified on the grounds that there is vastly more neurophysiological data available in the peripheral faculties.

2.6 Some Objections to the Hypothesis

2.6.1 Modular Anatomy and Localization of Function

A modular architecture is defined as one composed of a number of separate processors, each processor dedicated to a particular faculty of cognition. There is overwhelming evidence that the cortex is an architecture of this type, with each area counting as one module. The gross organization of the cortex into areas, the connections between areas, and the connections between them and the rest of the nervous system are undoubtedly genetically-specified. This is so because the same basic parcellation is found in all individuals of a species. Physiological investigations of neuronal properties in animals, and neuropsychological studies of the cognitive deficits following cortical damage in humans, confirm that different areas are engaged in different faculties, and that the same areas are engaged in the same faculties in different individuals. Thus, visual perception is localized in one region of the cortex, auditory perception in another, language functions in another, and so on. Within these regions, there is further segregation of function in particular areas. In visual cortex, for example, area V4 is important for the perception of colour but not for motion, whereas the opposite is true for area MT (Van Essen, 1985).

Neuropsychological evidence strongly suggests that there is also segregation of function in the parts of the cortex responsible for central cognition (Shallice, 1988; Fuster, 1989).

Inspired by a ‘columnar’ organization of the functional properties of neurons in many areas (e.g., Asanuma, 1975; Hubel & Wiesel, 1977; Livingstone & Hubel, 1984; Mountcastle, 1984; Brugge & Reale, 1985; Zeki & Shipp, 1988), some neuroscientists have argued that the cortex is also modular at a finer scale: they propose that the cortex consists of a mosaic of column-shaped processors of fairly uniform size (about 0.5 mm in width) and internal organization (Mountcastle, 1978; Szentágothai, 1978; Eccles, 1984; Kaas, 1990). Whether this is the case (I argue against it later) is not directly relevant to the argument of the present chapter; the same issues are raised whether the modules are small (columns) or large (areas).

None of this is embarrassing for the proposal of architectural uniformity. As pointed out earlier, a modular architecture is not necessarily a heterogeneous one. A modular architecture, in which different modules are dedicated—perhaps unalterably so—to different cognitive faculties, can also be uniform, provided that all the modules apply the same set of computational mechanisms to their different tasks. The cortex, I claim, is an architecture of this type. Different areas of the cortex receive inputs from different sources, send outputs to different destinations, and are responsible for different cognitive faculties. In this sense, the cortex is manifestly non-uniform. The cortex may well be uniform, however, in the sense that all the modules apply basically the same computational mechanisms (the same kinds of representation and the same basic operations) to their respective inputs. It is heterogeneity, not modularity, which is incompatible with the view that the architecture is uniform. The anatomical, physiological and neuropsychological evidence certainly shows that the cortex is a modular architecture, but it does not prove that it is a heterogeneous one.

The idea that the cortex is both uniform and modular provides an answer to one possible objection to the uniformity hypothesis: why is it that damage to the cortex often causes irrecoverable loss of function, since another part of the cortex should be able to substitute for the damaged tissue? The reason is that the gross division of the cortex into areas is genetically determined and is not modifiable (at least, once maturity is reached). If an area A is destroyed, no other area can substitute for A, unless it receives the same inputs as A and is capable of sending its outputs to the same places as A did; that is, unless it is connected to other areas and other parts of the nervous system in much the same way as A was. However, if the uniformity hypothesis is correct, it would be possible in principle.
for another area to substitute for A if the appropriate input and output connections were somehow established.

A similar argument explains why humans have certain capabilities that other species lack, even though the same basic mechanisms are present in the cortex of all mammalian species. Consider the case of language, the faculty in which the case for specialized mechanisms has been made most strongly (Chomsky, 1980; Fodor, 1983). It is quite compatible with the hypothesis advanced in the present chapter to suppose that there are areas of the human cortex, vital for language processing, that have no homologues in the cortex of other species. If an animal has no equivalent of Broca’s area, say, then no other part of the animal’s cortex can substitute for it, for the reason explained above. Furthermore, because the modularity of the cortex is genetically-specified, the animal cannot ‘grow’ a Broca’s area, even if given heavy exposure to language during development. The argument of the present chapter is in conflict with Chomsky only with regard to the claim that the language modules make use of processing mechanisms fundamentally different from those used, say, in visual perception. Unfortunately, very little is known about the internal organization of the cortical areas involved in language; the possibility must be conceded that these areas may indeed have a significantly different architecture from the rest of the cortex. This seems unlikely to be the case, however.

2.6.2 Diversity of Functional Properties of Cortical Neurons

In the last few decades a great deal has been discovered about the functional properties of single neurons in many cortical areas. In their famous series of experiments on area V1 of the visual cortex, Hubel and Wiesel (1962, 1968) found that neurons were highly selective in their responses to visual stimulation, each neuron giving a strong response only if a stimulus pattern of a certain type was displayed in a certain part of the visual field (the receptive field of the neuron). They found, for example, that many neurons responded well only when a light edge of a certain orientation was presented. Since these early studies, neurons have been discovered in the visual cortex whose responses are selective for a variety of other stimulus parameters, including velocity, wavelength, visual texture, binocular disparity, and spatial frequency (see Chapter 4). In visual cortex beyond V1, there is a tendency for different visual parameters to be analysed in different areas (Zeki, 1978; Van Essen, 1985; Livingstone & Hubel, 1987). For example, area V4 contains a high proportion of neurons that are differentially sensitive to colour, whereas most neurons in area MT are unselective for colour but are very sensitive to stimulus motion. Stimulus-selective neurons have also been found in the auditory and somatosensory cortices (reviewed by Brugge & Reale, 1985; Mountcastle, 1984). It is likely that a similar diversity of neuronal physiology exists in the cortical areas responsible for central cognition, such as the prefrontal cortex, although little is known in this field as yet (Fuster, 1989).

How is this diversity to be reconciled with architectural uniformity of the cortex? It was pointed out in the previous section that different areas of the cortex receive input connections from different sources. It follows that different areas receive different types of input information: visual cortex receives visual information, auditory cortex receives auditory information, and so on. Within visual cortex, areas responsible for different stages in visual perception receive information that has undergone different amounts of processing: V1 gets fairly raw visual input from the lateral geniculate nucleus, V2 gets the output of V1, and so on. The same principle applies at an even finer level: different groups of neurons within a single area can receive information from different sources. For example, distinct regions within V1 are associated with different layers of the lateral geniculate (Livingstone & Hubel, 1987). In the prefrontal cortex, too, different areas receive different connections (Pandya & Yeterian, 1985). In short, each part of the cortex has its own unique combination of sources of input. Given this, it is not surprising that neurons in different places have different functional properties, even if the same basic computational mechanisms are applied throughout the cortex. Diversity of neuronal physiology from area to area is a reflection of diversity in the information received by those areas, not diversity in the mechanisms applied to the information. A similar proposal has been made by Barlow (1985).

However, there is more to the story than this. The differences in functional properties between neurons in different areas are also caused by differences in the fine structure of the neural circuitry.
within those areas—the detailed pattern of connections and synaptic strengths. Although there is uniformity in cortical structure at a coarse scale, there is diversity at a fine scale. One possibility is that this nonuniformity is genetically-programmed and unmodifiable, and therefore part of the architecture according to our definition of the term. On this view, the near-uniformity in the gross organization of the cortex is misleading: there are actually large architectural differences between cortical areas. However, a more interesting possibility is that the fine structure is not part of the architecture, but is the result of (or at least is modifiable by) a learning process. The hypothesis of architectural uniformity would be consistent with the physiological diversity provided there was plasticity in the pattern of synaptic connections, and that the laws governing that plasticity were uniform throughout the cortex. Because it is generally assumed that modification of synaptic strengths is the means by which information is stored in neural structures, the physiological differences between neurons in different areas are caused by differences in the information stored in those areas as well as by differences in the information they process.

There is much evidence for plasticity in the functional properties of cortical neurons. It is well-known that neurons in V1 of kittens develop properties very different from the normal adult pattern if the animal is raised in an abnormal visual environment, provided this is done during a certain ‘critical period’ in development (see Chapter 4). Compared with kittens, the V1 physiology of newborn monkeys is much more like that of the adult (Wiesel & Hubel, 1974), but it too has a critical period during which neuron responses can be greatly altered by abnormal visual experience (e.g., Blakemore, Garey & Vital-Durand, 1978). Logically, some plasticity must persist in at least some cortical areas throughout life if the cortex really is a site of storage for long-term memory and skills. Plasticity in an adult animal has been demonstrated in the somatosensory cortex of the monkey. This region contains several ‘maps’ of the body surface which become substantially rearranged over a period of weeks if the sensory input is altered in various ways (reviewed by Merzenich, Recanzone, Jenkins, Allard & Nudo, 1988; Kaas, 1991). For example, if a monkey engages in behaviour that results in heavy stimulation of a limited skin surface, the representation of that surface in the somatosensory cortex expands (Jenkins, Merzenich, Ochs, Allard & Guí-Robles, 1990; Recanzone, Merzenich, Jenkins, Grajski & Dinse, 1992). Thus, plasticity in this region is not limited to a critical period of development but persists, presumably, throughout life.

The hypothesis of architectural uniformity requires not only that the functional properties of cortical neurons be modifiable by experience, but that the modification follows laws of synaptic plasticity that are uniform throughout the cortex. Experimental evidence bearing directly on this question has only recently started to become available. These experiments (see Sur, Pallas & Roe, 1990, for review and discussion) have investigated the effects of feeding information from one sensory modality into parts of the cortex that would, in the normal animal, process information from another modality. As described earlier, sensory information is relayed through the thalamus before reaching the cortex, and each sensory modality has its own thalamic nucleus (Table 2.1). Sur, Garraghty and Roe (1988) experimentally induced fibres from the retina to grow into the auditory part of the thalamus in visually-inexperienced newborn ferrets. In these animals, therefore, visual information was passing to the auditory thalamus and then to the auditory cortex. When the animals had grown to maturity, neurons in the auditory cortex were found to respond to visual stimulation. In many respects, the response properties of these neurons were very like those found in the visual cortex of normal ferrets: similar receptive field classes were found in similar proportions (Roe, Pallas, Kwon & Sur, 1992). In certain other respects, neurons in the auditory cortex of the operated ferrets were significantly different from those in the visual cortex of normal animals. However, there was evidence that the retinal cells supplying the visual information in the operated animals were of a type (the W class) that does not normally supply the major part of the visual cortex input, and this might account for the differences found (Roe et al., 1992; Roe, Garraghty, Esguerra & Sur, 1993). Similarly impressive results were obtained in an analogous experiment by Mézin and Frost (1989), who directed retinal fibres into the somatosensory thalamus of newborn hamsters. When the animals had grown to maturity, neurons in their somatosensory cortex were found to have visual response properties very like those in the visual cortex of normal hamsters: again, similar receptive field classes were present in similar proportions. In a different type of experiment, Schlagger and O’Leary (1991) transplanted a piece of occipital cortex (cortex that would normally be visual) to the somatosensory region in newborn rats. Part of the somatosensory cortex in rodents is marked by the presence in layer 4 of
distinctive columnar groups of neurons called **barrels** that are not present elsewhere in the cortex (Woolsey & van der Loos, 1970). Schlaggar and O’Leary found that an apparently normal array of barrels developed in the transplanted cortex, indicating that the potential to develop barrels is not unique to somatosensory cortex. It is tempting to speculate that all cortical areas have this potential, and that the barrels result from the particular properties of the sensory information delivered by the rodent’s whiskers.

The results of these experiments are consistent with the view that neurons in different areas of the cortex develop different functional properties by the application of uniform principles of neuronal plasticity to different types of input information. However, one difficulty in making this interpretation is the problem of distinguishing between processes of maturation and those of learning. An alternative interpretation of the results is that there are very different computational mechanisms in different cortical areas—that is, the cortical architecture is very non-uniform—but that the genetic program that establishes the mechanisms in a given area is triggered by some characteristic of the information input to that area during its formation. Although this explanation might seem unlikely, we should not be confident in taking these experiments as demonstrating uniform learning mechanisms throughout the cortex until we have reason to believe that the same processes underlie learning in adult animals. Nevertheless, the results are intriguing.

Finally, I will re-state the caveat that the cortex is not claimed to be a perfectly uniform architecture. Just as there are significant differences in anatomy from area to area, so there are likely to be differences in neuronal physiology—differences that are not the result of plasticity but are innate and fixed and therefore part of the architecture. This might include, for example, variations in axonal propagation speed and dendritic membrane currents. The claim is only that the degree of architectural uniformity is sufficiently great that a general model of the idealized canonical circuit is possible.

### 2.6.3 From Uniform Neural Circuitry to Uniform Cognitive Architecture

The reader might object that all this neuroscientific stuff has little importance for psychology, because it is only relevant to the level of implementation (p. 13). The observation that the neural circuitry in the cortex is basically the same in all areas only means that the computations responsible for different faculties have similar implementations; this does not imply that the representations and computations are themselves similar. (This argument is directly analogous to the observation that a wordprocessor, a statistics package and a video game can be running on computers with identical hardware.) In short, uniformity of neural circuitry does not necessarily mean uniformity of cognitive architecture. In this section I will argue that there is good reason to suppose that there is a level of description of cortical function that is recognizably a level of representations but which is sufficiently close to the neural circuitry to inherit its uniformity. The existence of such a computational level justifies the conception of the cortex as a uniform cognitive architecture, and hence the possibility of a unified theory of cognition based on a model of this architecture.

Cognitive scientists have designed a huge number of different cognitive architectures, both as psychological models and as artificial intelligence programs. These models fall into two, reasonably well separated classes, which we can call **symbolist** architectures and **connectionist** architectures. The prototypical symbolist architecture, as described, for example, by Newell (1981, 1990) and Pylyshyn (1984), has a character very similar to the architecture implemented by an interpreter for the Lisp programming language running on a conventional digital computer. The representations it provides are discrete, word-like symbols which can be composed into sentence-like symbol structures, and the basic operations it provides are operations for creating, combining and comparing symbol structures. The prototypical connectionist architecture, as described, for example, by Rumelhart, Hinton and McClelland (1986), is quite different. It consists of a collection of neuron-like units joined by connections; each unit has a level of activity, which changes as a function of the activity levels of other units to which it is connected and of the strengths of its connections; the representations supported by the architecture are patterns of activity distributed over the units, and patterns of connection strengths; and basic operations provided include the activation functions of the units and rules for modifying the strengths of the connections. Clearly, symbolist architectures are inspired by conventional digital computers of the von Neumann type, whereas connectionist architectures are inspired by—or in some cases are biologically-detailed models of—the neural networks of the brain.
(Symbolist architectures are often referred to as ‘classical’ architectures in the literature; ‘symbolist’ is more descriptive. It is admittedly an ugly word, but ‘connectionist’ is even worse, and we are now stuck with that.)

Not all the architectures that cognitive scientists have investigated have all the characteristics of either prototype. For example, the visual imagery model proposed by Kosslyn (1980) is plainly a symbolist architecture, yet it uses representations that are not sentence-like symbol structures. Nevertheless, the great majority of architectures are much closer to one prototype than the other. That is, symbolist and connectionist architectures form two distinct clusters in the space of possible architectures. It is unknown whether these clusters are natural kinds, or whether they just result from lack of interest or imagination in exploring the whole space. However, we can only work from what we know, so I will assume that the architectures realized in different parts of the brain are recognizably symbolist or recognizably connectionist.

Now consider the architecture realized in one particular area of the cortex. Assume that this area, call it Area P, is dedicated to rather low-level perceptual processing; it may be engaged in early vision, for example. The cortex evolved originally for perceptual-motor tasks, so it seems certain that Area P is extremely efficient at perceptual processing. (As pointed out earlier, we cannot be so sure that areas of the cortex devoted to central faculties are highly efficient.) In particular, we can assume that Area P is extremely well-adapted for the task of pattern recognition. From this observation it is possible to predict that the architecture realized by Area P, whatever its type, has certain characteristics:

**Efficiency.** Area P must be capable of performing very fast pattern recognition. Humans are capable of performing many basic perceptual tasks in much less than a second; however, cortical neurons have membrane time constants typically about 10–20 msec (McCormick, Connors, Lighthall & Prince, 1985), meaning that they change state very slowly compared with the switching rates of electronic computers. This means that the pattern-recognition algorithm implemented in the cortex cannot consist of a large number of sequential operations (Feldman & Ballard, 1982; Thorpe & Imbert, 1989). Considering the huge information content of a visual image, for example, the implication is that the algorithm must involve a considerable amount of parallel processing. Another constraining factor is the vast number of patterns that a human can recognize quickly. This implies that the algorithm must give fairly direct access to the correct pattern in memory; a sequential search of possible matches is out of the question.

**Robustness.** The pattern-recognition algorithm must be robust as well as fast: it must give reasonable performance when the input pattern is noisy, incomplete, or otherwise degraded. This rules out algorithms that fail completely under these circumstances.

**Learning.** As well as recognizing patterns, the architecture of Area P must also provide a learning mechanism so that it can acquire new ones. This cannot merely involve storing copies of all input patterns, because there will be far too many of them. Area P is required, rather, either to abstract representations of categories from the patterns to which it is exposed, or to store particular patterns that will be useful to recognize in the future. Another difficulty is that there is no external agent informing Area P exactly what it should learn. Instead, Area P must **self-organize**, using only the information available in its input patterns. (Neuronal plasticity may be ‘modulated’ by extrinsic factors, such as the nonspecific fibre systems originating in the brain stem and basal forebrain, but these general influences fall far short of what would be required of a specific teaching mechanism.)

Now, it is a striking fact that connectionist architectures are **much** better suited to these requirements than symbolist architectures (McClelland, Rumelhart & Hinton, 1986), Fodor and Pylyshyn (1988), discussing these features of connectionist systems, argue that there is no reason, in principle, why a symbolist architecture should not also exhibit them. This claim would be a good deal more convincing if there were a few working demonstrations around of symbolist models having these characteristics; as Chater and Oaksford (1990) observe, the symbolist has work to do! On the other hand, there are many demonstrations of connectionist architectures having most or all of these properties. This strongly suggests that the architecture realized by Area P is a connectionist architecture.

There are two ways of interpreting the proposition that Area P realizes a connectionist architecture. The first is that there is a direct mapping from the units and connections of the network onto the physical structure of the neural substrate. A unit may correspond to a single neuron, a group
of neurons, or even to a piece of dendrite; similarly, a connection may correspond to a single synapse or several synapses. The only requirement is that each unit and connection has a fixed physical realization. The second interpretation is that the units and connections of the network are purely functional entities. On this interpretation, a single unit in the connectionist network can be realized by different physical states of the underlying neural network at different times, in the same way that a single data structure in a computer can be realized by different physical states of the computer’s circuitry at different times. This second interpretation, however, is difficult to take seriously. After all, the main motivation behind investigating connectionist architectures is that the units are very much like neurons. There is no reason to suppose that purely functional entities have properties anything like the physical entities that realize them, as the case of the data structure in the computer demonstrates. It is therefore more reasonable to assume the first interpretation.

We arrive at the conclusion that the representations and computations used in Area P are implemented very directly in the neural circuitry of the area. This allows us to make the crucial generalization: since the neural circuitry is approximately uniform throughout the cortex, the kinds of representations and computations that are used in Area P are also used in all other cortical areas. The cortex is a uniform cognitive architecture, and a general theory at this level would form a reasonable basis for a unified theory of cognition.

It might be argued that, although the computations at this particular level of description are uniform, they nevertheless serve very different purposes in different regions of the cortex. Specifically, in peripheral regions (such as Area P) they are themselves responsible for the perceptual and motor processing that goes on there, but in the regions subserving the central faculties they serve merely to implement higher-level, symbolist, faculty-specific architectures. Hence, a general theory at this low level would not really be a unified theory of cognition. It would provide a good account of peripheral cognition, but would miss important regularities in its account of the central faculties. I concede that this is possible. Indeed, something along these lines might well be happening in certain cases, because humans are capable of consciously following explicitly-stored rules—an activity in which we might be emulating a symbolist architecture (Smolensky, 1988; Clark, 1989, Chap. 7). But this does not seem to be the normal mode of human cognition. As argued earlier, a large proportion of skilled, fluent cognitive activity, in central as well as peripheral cognition, seems to be based on pattern recognition—the task for which, I suggest, the cortex is ideally suited. Therefore, a theory placed close to the neural level may eventually provide a good account of central as well as peripheral faculties.

Some psychologists might be uncomfortable with the very close relationship being advocated between computation and its neural implementation in the cortex, since psychological theory is usually placed at a much more abstract level of description. However, the close relationship is surely a virtue, because it enables a vast amount of anatomical and physiological data from experimental neuroscience to be brought to bear on constraining cognitive theory. This can help to overcome one of the traditional problems of cognitive psychology, whether behavioural and introspective data alone is sufficient to uniquely specify the cognitive architecture (Anderson, 1978). Of course, the formulation of a general theory at this low level would not make higher-level models (regarded as approximations to the low-level account) completely redundant. Higher-level theories, of the kind that dominate the current cognitive psychology literature, may well continue to be valuable, because they may be easier to understand or to apply than the low-level theory. However, the low-level theory will have the virtue of giving a unified account of an extremely broad range of phenomena, whereas high-level theories are likely to remain specific to particular faculties.

2.7 Summary

This chapter has examined the proposal that the cerebral neocortex is an approximately uniform cognitive architecture, and that a general model of the computational mechanisms realized by the canonical cortical circuit could form the basis of a unified theory of cognition. The cognitive architecture was defined as comprising all the structures and mechanisms of the cognitive system that are genetically-specified and not modifiable by learning. A uniform architecture, in which a single system of computational mechanisms is used in all cognitive faculties, was distinguished from a heterogeneous architecture, in which a fundamentally different system of mechanisms is used in each
faculty. Although the human cognitive architecture as a whole is certainly heterogeneous, it was proposed that the cerebral neocortex is a roughly uniform architecture. Anatomical evidence was reviewed in support of the proposal that the neural circuitry beneath a small patch of the cortical surface is very similar throughout the cortex. It was proposed that a general theory of the function of the canonical cortical circuit is possible. It was argued on evolutionary grounds that the cortex is basically a perceptual-motor processor, and that the basic functions performed by the canonical circuit are pattern-recognition, the production of motor output, and learning. Central cognition was examined in the light of this suggestion, and implications for cognitive psychology research were discussed.

Some objections to the proposal were considered. It was pointed out that the localization of faculties in different areas of the cortex is not incompatible with its being a uniform architecture, since the proposal is only that all cortical areas apply the same basic mechanisms to their respective inputs. It was argued that differences in the functional properties of neurons from area to area arise from the fact that different areas process and store different types of input information. Some experimental evidence was reviewed in support of the suggestion that the laws governing synaptic plasticity are similar in all cortical areas. Finally, an argument was presented that uniformity of neural circuitry does suggest uniformity of representations and of elementary operations: if the cortex is basically a perceptual-motor processor, then it is much more likely to be a connectionist than a symbolist architecture. This is not incompatible with the existence of higher-level, faculty-specific theories in addition to the low-level, uniform one.

The fundamental hypothesis is that the degree of uniformity in the cortical circuitry is sufficiently great for an idealized, general model of the canonical circuit to be theoretically useful. It is important to emphasize that this hypothesis may turn out to be incorrect. Detailed anatomical studies of presently-unexplored regions of the cortex may reveal substantial variations in intrinsic circuitry. Even if there is approximate uniformity in the anatomy, it is still possible that subtle, genetically-specified variations in physiology from area to area (say, in the distribution of neurotransmitters) could make substantial computational differences, in which case a general theory of cortical function would not be feasible. But if a general model is possible, there is good reason to suppose that a unified theory of cognition could be built upon it, although it would, of course, have to be supplemented with theories of other parts of the brain. The following chapter introduces a general theory of the canonical cortical circuit.
Chapter 3
A General Theory of the Neocortex

In this chapter I take it as a working hypothesis that it is possible to abstract a canonical cortical circuit which is a rather close approximation to the actual circuitry present in any part of the neocortex. I will describe several general-purpose computational mechanisms that, I propose, are realized in the canonical circuit. By assumption, therefore, these mechanisms will be found to be operating in all cortical areas, although there will certainly be some variation from area to area in the details of their operation.

As argued in the previous chapter, the fact that the cortex originally evolved for perceptual-motor processing suggests that it is highly efficient for perceptual-motor tasks. For the general presentation of the theory in this chapter the cortex is treated as if it is used only for perceptual-motor processing. That is, I will draw on physiological data exclusively from the sensory and motor regions of the cortex—primarily from the visual cortex, this being the best-understood region. For the reasons explained in the previous chapter, consideration of such central faculties as language and problem solving is unlikely to throw much light on the basic computational mechanisms provided by the cortex, and so these topics are ignored in this chapter.

It is customary in cognitive science to refer to large and complex theories by short and snappy names. I call the theory presented here the Magpie theory of the neocortex because it collects good ideas from a variety of other theories.

3.1 The Model Cortical Neuron

An essential prerequisite for understanding computation in the cortex is to have an appropriate model of the computational properties of single cortical neurons. A simple model is to be preferred, so that its properties can be easily understood at an intuitive level, and to minimize the computational burden in simulating large networks. On the other hand, the model must be sufficiently realistic to capture the essential computational properties of real neurons. Unfortunately, which properties are essential and which can be neglected is a question that cannot be answered a priori, but only in the context of a theory of cortical function. It also depends to a large extent on what aspect of cortical function the theorist is addressing. A very detailed model neuron that takes account of the various different species of membrane channel and the electrotonic spread of potential through the dendritic tree may be essential for providing an accurate explanation of intracellular recording experiments; but this amount of detail might make the model far more complex than it need be when trying to understand networks of thousands or even millions of neurons. This section introduces the model neuron that is used in the current computer simulations of the Magpie theory. I do not claim that aspects of real neurons omitted from this model are computationally irrelevant, only that they are of secondary importance in a model placed at this level.

The first simplifying assumption is to treat the whole soma and dendritic tree of the neuron as a single electrical compartment, characterized by a single membrane potential. This is a simplification because in real neurons the membrane potential is non-uniform—a local potential resulting from a current injected by a synapse has to spread through the dendritic tree before reaching the soma. Measurements made in vitro indicate that cortical neurons are electrotonically quite compact (Stafstrom, Schwindt & Crill, 1984; Stratford, Mason, Larkman, Major & Jack, 1989; Douglas & Martin, 1990a), so the attenuation suffered by a spreading dendritic potential is not too severe.
However, simulations suggest that the effective electrotonic distance is greater under normal physiological conditions, when the cell is constantly bombarded by synaptic events resulting from the activity of other neurons (Bernander, Douglas, Martin & Koch, 1991). In particular, attenuation can be expected to be quite severe in the apical dendrites of pyramidal neurons, which may be over 3 mm long (Stratford et al., 1989). In the present model the attenuation caused by the apical dendrite is ignored. However, there are indications that the effect has important computational consequences (p. 56), so it might prove necessary in a future model to treat the apical dendrite as one or more electrical compartments separate from the soma (see Segev, Fleshman & Burke, 1989). The isopotential simplification also prevents the model from being able to account for any local computations that might occur within the dendritic tree or on individual dendritic spines (e.g., Koch & Poggio, 1987). However, there is no proof at present that such local effects are important. It is also assumed that the soma-dendritic membrane does not generate action potentials. Dendritic spikes have not been reported in the neocortex under normal physiological conditions (Prince & Huguenard, 1988; Douglas & Martin, 1990a; but see Pockberger, 1991), although in vitro studies show that the machinery for producing such spikes does exist at least in the apical dendrites of pyramidal neurons (Amitai, Friedman, Connors & Gutnick, 1993; Kim & Connors, 1993; Stuart & Sakmann, 1994).

Under these assumptions, the electrical properties of the cell membrane can be modelled by the circuit illustrated in Figure 3.1B. The membrane capacitance is represented by a single capacitor; the capacitance current $i_{\text{cap}}$ is equal to $C \frac{dV}{dt}$ where $C$ is the constant capacitance and $V$ is the potential across the membrane. The membrane contains pores or channels that allow ions to flow into and out of the cell under chemical and electrical pressure. The permanently-open channels are represented in the circuit by a resistor in series with a battery. The flow of ions through these channels constitutes a current, called the leak current, which by Ohm’s law obeys the relation $i_{\text{leak}} = (V - E_r)g_r$ where $E_r$ is the rest potential and $g_r$ is the leak conductance of the membrane. (Conductance is the reciprocal of resistance, $g = 1/R$, and is measured in siemens or reciprocal ohms.) The various different kinds of ion involved are not treated separately in this simple model. Considering only these two currents ($i_{\text{cap}}$ and $i_{\text{leak}}$) for the moment, Kirchhoff’s law tells us that their algebraic sum is zero, giving the equation:

$$C \frac{dV}{dt} = (E_r - V)g_r$$

(3.1)
Clearly, the membrane potential achieves a steady state \((dV/dt = 0)\) when \(V = E_r\), the rest potential. At this potential the electrical and chemical forces acting on the ions are in equilibrium and there is no net flow of current. The rest potential of cortical neurons is typically about \(-75\) mV.

Consider now the effect of a single synaptic input on the membrane potential. When an action potential arrives at the presynaptic axon terminal a small amount of neurotransmitter is released which diffuses across the synapse to the postsynaptic membrane. There it binds chemically with specialized molecules called receptors, which in turn cause associated membrane channels to transiently open. The flow of ions through these channels constitutes an additional current across the membrane, \(i_{syn}(t)\), momentarily affecting the membrane potential. In the circuit (Fig. 3.1B) the synapse is represented by a battery \(E_{syn}\) in series with a variable resistor with conductance \(g_{syn}(t)\). The function \(g_{syn}(t)\) describes the change in conductance that results from the opening of the synaptic channels. It is assumed that the conductance change does not depend on \(V\). This is a simplification for excitatory synapses because they are likely to be mediated in part by NMDA receptors, which are voltage-dependent (p. 67). The synaptic current is given by \(i_{syn}(t) = (V - E_{syn})g_{syn}(t)\) and incorporating this current into Equation 3.1 gives:

\[
C \frac{dV}{dt} = (E_r - V)g_r + (E_{syn} - V)g_{syn}(t)
\]  

(3.2)

The effect of the synapse on the membrane potential depends on the constant \(E_{syn}\), the reversal potential of the synapse. If \(E_{syn} > E_r\), the synapse tends to raise \(V\) above \(E_r\), bringing \(V\) closer to 0 (remember \(E_r\) is negative); that is, it depolarizes the membrane and so the synapse is excitatory. If \(E_{syn} < E_r\), the synapse hyperpolarizes the membrane and so the synapse is inhibitory. If \(E_{syn} = E_r\) the synapse does not push \(V\) away from the resting potential, although by increasing the membrane conductance it makes it more difficult for excitatory synapses to depolarize the membrane; this is called shunting inhibition. Actually, there is little evidence for this latter effect in the cortex (Douglas, Martin & Whitteridge, 1988; Ferster & Jagadeesh, 1992).

Equation 3.2 is easily extended to the case of multiple synaptic inputs under the additional simplifying assumptions that all the excitatory synapses have reversal potential \(E_e\), that all the inhibitory synapses have reversal potential \(E_i\), and that the conductance changes caused by all the synapses sum linearly:

\[
C \frac{dV}{dt} = (E_r - V)g_r + (E_e - V)\sum_j g_j(t) + (E_i - V)\sum_k g_k(t)
\]  

(3.3)

where the summations \(j\) and \(k\) are over all the excitatory and inhibitory synapses, respectively, and \(g_n(t)\) is the conductance change caused by activation of the \(n\)th synapse. The reversal potentials \(E_e\) and \(E_i\) are effectively upper and lower bounds on the range of values that \(V\) can attain. This can be seen from Equation 3.3, bearing in mind that \(E_e > E_r > E_i\) and that capacitance and conductances are positive: if \(V < E_i\) then \(dV/dt > 0\) and if \(V > E_e\) then \(dV/dt < 0\). In the simulations \(E_e\) was fixed at 10 mV, and \(E_i\) was fixed at \(-80\) mV. These are typical values in the neural modelling literature. In real neurons the situation is more complicated because there are two kinds of inhibitory synapse (involving GABA_A and GABA_B receptors) which have different reversal potentials.

Now consider the output of the model neuron. In common with many simplified neural network models, individual action potentials are not used. Instead, the output is a graded, dimensionless quantity in the range [0, 1] called the firing rate, which can be considered to be a running average of the frequency at which action potentials are generated. The lack of individual spikes in the model means that certain effects observed in the cortex cannot be accounted for. For example, correlations between the spike trains generated by different neurons have recently been discovered in the cat’s visual cortex (Gray, König, Engel & Singer, 1989; reviewed by Engel, König, Kreiter, Schillen & Singer, 1992). In spite of much theoretical turmoil (e.g., Niebur, Koch & Rosin, 1993) it is still
unknown whether this has any computational significance or is as epiphenomenal as the hum of an engine. Also, the model cannot account for elaborate temporal patterns in spike sequences. Such patterns apparently exist (Richmond, Optican & Spitzer, 1990), although whether neurons are actually capable of making use of them to transmit information seems doubtful (Douglas & Martin, 1991b; Tovée, Rolls, Treves & Bellis, 1993). The firing rate of the model neuron is related to the membrane potential $V$ by the function:

$$f(t) = \tanh^+(\frac{V - \rho}{\gamma})$$

(3.4)

where $\rho$ and $\gamma$ are constants and the function $\tanh^+$ is the positive part of the hyperbolic tangent function: $\tanh^+x = \max(0, \tanh x)$. This function is shown in Figure 3.2: it has a threshold at $V = \rho$ and thereafter increases with a slope determined by $\gamma$ until levelling off at the asymptotic value 1. The spontaneous firing of the neuron (when $V < \rho$), which is very low in cortical neurons, is ignored in the model. A further assumption is that the production of action potentials does not affect the dendritic membrane potential. This is probably a severe simplification; a recent in vitro study suggests that voltage-dependent conductances in the soma-dendritic membrane enable action potentials initiated in the axon to propagate back into the dendritic tree (Stuart & Sakmann, 1994).

Different categories of cortical neuron have different spiking behaviours (reviewed by Prince & Huguenard, 1988). The majority of pyramidal neurons are regular spiking cells, producing a steady sequence of action potentials in response to sustained depolarization (McCormick, Connors, Lighthall & Prince, 1985). They show pronounced spike-frequency accommodation, although this effect is not incorporated in the present model. Neurons with smooth dendrites (nearly all inhibitory cells) are fast spiking. Compared with pyramidal neurons, they are able to fire at higher rates, are not spontaneously active at rest, and show little or no spike-frequency accommodation (McCormick et al., 1985). These differences in spiking behaviour between the two types of neuron are ignored in the model. A third category of neurons are bursting cells, which produce a burst of three or more spikes when depolarized. These neurons are a small subset of the pyramidal neurons in layers 4 and 5 (McCormick et al., 1985). This behaviour is not incorporated in the model. It would be interesting to know whether bursting occurs in the awake animal or whether it is connected with the slow-wave EEG seen in sleep, as is the case with some cells in the thalamus (p. 85).

It remains to specify how the change in conductance caused by activation of a synapse is related to the firing rate of the presynaptic neuron. First, we deal with the minor complication caused by the fact that, in general, there may be more than one distinct synapse formed between a presynaptic cell and a postsynaptic cell. Rather than treat these synapses separately, it is more convenient to lump them together into a single, abstract connection. Thus, there is at most one connection from any neuron to another. Connections can be excitatory or inhibitory.

Suppose there is a connection from neuron $u_j$ to neuron $u_i$. Then $g_{ij}(t)$, the change in membrane conductance of $u_i$ caused by $u_j$, is taken to be proportional to $f_j(t)$, the firing rate of $u_j$. 

Figure 3.2: The firing rate of the model neuron as a function of the membrane potential.
More specifically \( g_{ji}(t) = g \omega_{ji} w_{ji} f_j(t) \), where \( g \) is a constant conductance that is the same for all connections of that type, \( \omega_{ji} \) is a dimensionless non-negative number called the hard weight of the connection, and \( w_{ji} \) is another dimensionless non-negative number called the soft weight of the connection. The hard weight represents the fixed ‘strength’ of the connection which may, for example, be a function of the distance between the two neurons. The soft weight also contributes to the connection strength, but unlike \( \omega_{ji} \) it is modifiable: it can slowly increase or decrease over time. All long-term storage of information is assumed to occur through the modification of soft weights. Not all types of connection have soft weights, however; in particular, inhibitory connections have only hard weights and their strength cannot be modified.

The assumption that the conductance change is simply proportional to the presynaptic firing rate is certainly a simplification. In reality the channels opened by activation of a synapse can remain open for extended periods after the presynaptic activity has ceased, having prolonged effects on the postsynaptic potential. The most important example is the inhibitory channels associated with GABA \(_B\) receptors, which can hyperpolarize cortical neurons for some 200–300 msec (e.g., Douglas & Martin, 1991a). The delay caused by the time taken for an action potential to propagate from \( u_j \) to \( u_i \) and by synaptic transmission is ignored in the model. This seems reasonable since the total delay would probably not exceed 4 msec for two neurons within a local cortical circuit.

This concludes the initial description of the model neuron and its simplifying assumptions; further details are given when we consider the simulations in Chapters 4 and 5. Abstracting from the details, this model corresponds closely to the ‘classical neuron’ described in psychology textbooks: excitatory and inhibitory inputs are summed (albeit not strictly linearly) and translated into a firing rate which is propagated along the axon to other neurons. More esoteric phenomena, such as highly nonlinear interactions in the dendritic tree or oscillatory firing, are not included. This is not to say that such phenomena do not exist, only that they are not of fundamental importance in understanding cortical function. Whether this assumption will be substantiated remains to be seen.

### 3.2 Population Coding

Virtually all connectionist architectures make a basic distinction between two forms of information storage. The first, which can be called active storage, consists of the representations encoded by the activity of the units. The second, which can be called latent storage, consists of the representations encoded by the strengths of the connections between the units. In slightly less abstract terms, active and latent storage correspond to the firing rates of the neurons and the soft weights of the connections, respectively. Loosely, active storage holds the information currently being processed by the architecture, and latent storage holds the knowledge and skills that the architecture has acquired in the past. (One of the essential characteristics of connectionist architectures is that these two kinds of information are not stored in separate processing modules, as they are for example in von Neumann computers, but are represented by different aspects of the same physical network.) The Magpie theory adheres to this basic dichotomy, although it is probably over-simple; there are processes, such as neuronal adaptation, that are not easily conceptualized in terms of either kind of storage. Note that latent and active storage must not be identified with the long-term and short-term memories studied by psychologists, because they are both much more general; latent storage contains perceptual and motor skills as well as representations of events and facts, and active storage contains a great deal of information of which we are not conscious.

Consider a cortical neuron that receives \( N \) excitatory connections with modifiable weights. We ignore inhibitory and unmodifiable excitatory inputs for the time being. Let \( \omega_i \) and \( w_i \) be the hard and soft weights of the \( i \)th connection to the neuron, and let \( x_i \) be the firing rate of the corresponding presynaptic cell. It is convenient to denote the set of \( N \) weights by the vector \( \mathbf{w} = [\omega_1 w_1, ..., \omega_N w_N]^{T} \), and the firing rates similarly by \( \mathbf{x} = [x_1, ..., x_N]^{T} \), so that \( \sum_i \omega_i w_i x_i = \mathbf{w}^{T} \mathbf{x} \) where \( ^T \) denotes vector transpose. Under the assumptions of the model neuron described in the previous section, the membrane potential \( V \) of the neuron is governed by the equation:

\[
\frac{dV}{dt} = (E_r - V) g_r + (E_e - V) g_e \mathbf{w}^{T} \mathbf{x}
\]  

(3.5)
It is useful to consider the simple case in which the input firing rates are constant. $V$ then attains its asymptotic value:

$$V = \frac{g_rE_r + E_gw^T x}{g_r + g_ww^T x}$$

which is a monotone increasing function of $w^T x$ (note that $w^T x \geq 0$ because weights and firing rates cannot be negative). In fact $V$ increases approximately linearly with $w^T x$ provided $g_ww^T x$ is small compared with $g_r$. The firing rate of the neuron is also an increasing function of this inner product, albeit with a threshold nonlinearity (Fig. 3.2). This has an instructive geometrical analogy: by an elementary identity $w^T x = \|w\|\|x\|\cos A$, where $\|\cdot\|$ denotes vector length (Euclidean norm) and $A$ is the angle between $w$ and $x$, so for a given $\|w\|$ and $\|x\|$ the inner product is a measure of the similarity of the two vectors (cf. Grossberg, 1976). The inner product is maximal when $w$ and $x$ are parallel ($\cos A = 1$) and zero when they are orthogonal ($\cos A = 0$). The neuron can therefore be considered a simple pattern-recognition device: its weight vector $w$ can be taken to encode a particular pattern of activity that may occur in its modifiable excitatory input connections, and its firing rate can be interpreted as a measure of how well the current input $x$ matches that optimal input pattern. The pattern of activity in the input connections that maximally activates the neuron (for a given $\|x\|$) can be called the neuron’s characteristic pattern.

As proposed in the previous chapter, pattern recognition is the single most important function of the canonical cortical circuit. It is assumed that activation of neurons with appropriate characteristic patterns is the essential means by which the cortex performs this task. An immediate difficulty with this assumption is that a given input pattern will in general activate many different neurons. This is because neurons respond not just to their characteristic pattern but to a range of different patterns. (A neuron that gives a strong response only to an input pattern that closely matches its characteristic pattern is said to be sharply tuned, whereas one that gives a strong response to a relatively large range of inputs is said to be broadly tuned. The sharpness of tuning is strongly influenced by the local excitatory and inhibitory circuitry in which the cell is embedded; this is considered in the following section.)

Suppose all neurons in a certain group receive connections from the same source. A pattern of firing $x$ presented to the group causes different neurons in the group to respond to different degrees, according to how well $x$ matches their characteristic patterns. How is this group response to be interpreted? One possibility is to consider only the neuron with the highest firing rate, the neuron whose characteristic pattern matches $x$ most closely. This is called nearest-neighbour coding and has the virtue of simplicity. However it seems very inefficient to use only the activity of a single neuron to represent $x$ in active storage, when several other neurons can potentially contribute useful information about it. The Magpie theory assumes instead that it is the response of all the neurons in the group that constitutes the representation. Population coding, as this is called, has been the subject of several formal and simulation studies (Hinton, McClelland & Rumelhart, 1986; Hintzman, 1986; Baldi & Heiligenberg, 1988; Lehky & Sejnowski, 1990; Poggio, 1990; Poggio & Edelman, 1990; Vogels, 1990; Lukashin & Georgopoulos, 1993).

Population coding offers a number of advantages over nearest-neighbour coding. First, it provides a much more accurate representation of the stimulus for a given number of neurons, or, conversely, it requires far fewer neurons to attain a given level of accuracy (Erickson, 1968). To discriminate between $N$ different input patterns using a nearest-neighbour code requires $N$ neurons, whereas with a population code one only needs enough neurons for the $N$ patterns to give rise to $N$ distinguishably-different patterns of firing in the group. A second advantage of population coding is that it is damage-resistant: the death of one neuron would cause the loss of one characteristic pattern from latent storage, but any input pattern that activated that neuron could still be represented in active storage, albeit less accurately, provided there were other neurons with similar characteristic patterns that were also activated by that input. The old chestnut about millions of cortical neurons dying every day is almost certainly a myth—modern quantitative studies suggest that the rate of neuron loss is
very modest in the healthy cortex (Haug, 1985; Terry, DeTeresa & Hansen, 1987)—but damage-resistant representation is still a useful property to have.

It is instructive to examine the simple case in which each neuron in the group has a characteristic pattern encoding a simple scalar parameter. For example, each neuron may represent the orientation or the velocity of a visual stimulus as its characteristic pattern. (As we shall see later, the reality is more complex than this: even at the earliest stages of perceptual processing neurons actually encode a particular combination of parameter values.) Let the real number \( y_i \) denote the value encoded by the characteristic pattern of the \( i \)th neuron, let \( f_i \) be that cell’s firing rate when presented with input pattern \( x \), and let \( x \) be the relevant parameter value of \( x \). Exactly how does the pattern of firing rates in the group represent \( x \)? Nearest-neighbour coding would choose the parameter encoded by the most active neuron: \( y = y_k \) such that \( f_k = \max_i f_i \). The representation under population coding, however, is not immediately clear. The answer, in fact, depends on how other neurons make use of the information represented in this group; at this stage in the analysis we are not in a position to address this because we are considering the group in isolation. However, a reasonable interpretation is to assume that each neuron makes a contribution to the representation of \( x \) proportional to its firing rate: that is, \( y = \sum_i f_i y_i \). Hence, those neurons with characteristic patterns closely matching \( x \) make a large contribution to the representation, neurons with characteristic patterns that are poorly matched make little or no contribution, with other neurons intermediate. Baldi and Heiligenberg (1988) have analysed this scheme formally and have shown that the relation between the sum \( y \) and the input \( x \) is approximately linear, assuming that the values \( y_i \) are fairly evenly distributed in their range and that the input activates a reasonable number of cells. This result can be generalized to the more realistic case in which neurons represent multi-dimensional patterns (Baldi & Heiligenberg, 1988). One way to interpret this is that the brain is interpolating between the finite number of particular values encoded by the individual neurons (cf. Poggio, 1990).

In this representation it is the firing rate of each neuron relative to the firing rates of the other neurons in the population, rather than the absolute firing rate, that indicates how well the neuron’s characteristic pattern matches the current input. This is because if the firing rates \( f_i \) are all scaled by the same constant factor, the sum \( y \) is also scaled by that factor, but it remains linearly related to \( x \). (This means that the absolute firing rates of the cells could potentially be used to encode some other parameter.) Note that the population response gives a much more accurate representation than the response of any single neuron. Because population coding requires neurons to be fairly broadly tuned, each neuron gives a strong response over a range of different inputs, and therefore no individual cell can provide very accurate information about the identity of the stimulus.

A major problem with population coding is that of interference between the representations of two or more similar patterns that are simultaneously present in the stimulus. This problem and its proposed solution are discussed later in the chapter.

Circumstantial evidence that population coding is used in the visual cortex is considered in Chapters 4 and 5. At present, the strongest evidence for population coding in the cortex is from studies of the motor cortex; this evidence is reviewed in the discussion of motor production below.

### 3.3 The Local Circuitry

Because the Magpie theory assumes the cortical neuron to be a rather simple device, at least to a first approximation, much of the computational power of the cortex must derive from the circuitry in which each neuron is embedded (cf. Douglas & Martin, 1991b). I use the term local circuitry to mean the connections between neurons within a single layer of the cortex and with a lateral (within-layer) spread of about 1 mm or less; that is, the circuitry contained in a fraction of a cubic millimetre of cortical grey matter. Although the number of neurons beneath a square millimetre of the cortical surface is fairly uniform (p. 21), there is greater variation in the number per cubic millimetre of tissue because of differences in cell size and amount of cabling; Beaulieu and Colonnier (1989) found an average of about 27000/mm\(^2\) in the cat, with a much higher figure (49000/mm\(^3\)) in the primary visual area. Omitting a small fraction of specialized cells, cortical neurons are of two main classes: those which release the inhibitory neurotransmitter \( \gamma \)-aminobutyric acid (GABA), and those which are excitatory in function, probably using the neurotransmitter glutamate and/or aspartate (Douglas &
Martin, 1990a). The latter group mostly consists of pyramidal neurons, but in layer 4 of certain areas there are excitatory neurons that lack apical dendrites, the spiny stellate cells. By definition, the canonical circuit of the cortex omits features peculiar to particular areas, and so spiny stellate cells will not be considered further. However, there are reasons (p. 21) to treat spiny stellate cells as equivalent to pyramidal neurons at the abstract level on which we are working.

Inhibitory neurons make up between a fifth and a quarter of the population (Hendry, Schwark, Jones & Yan, 1987). All of them have axons that ramify locally (Fairén, DeFelipe & Regidor, 1984), forming synapses both with each other and with the pyramidal neurons (LeVay, 1973; Somogyi, Kisvárday, Martin & Whitteridge, 1983; Kisvárday, Beaulieu & Eysel, 1993). This class can be subdivided into several types on the basis of their axonal and dendritic distribution patterns (Fairén et al., 1984); some are illustrated in Figures 2.5 and 2.6. Two types are especially noteworthy. Chandelier cells (reviewed by Peters, 1984) have axon collaterals producing many short vertical segments that give the axonal plexus the appearance of a chandelier (Fig. 2.6D). Each ‘candle’ makes multiple synaptic contacts with the initial segment of the axon of a pyramidal neuron. Chandelier cells are known to release GABA (Freund, Martin, Smith & Somogyi, 1983). The strategic placement of their outputs means that they can exert a very potent inhibitory effect on the pyramidal neurons in their neighbourhood, although simulations suggest that the inhibition is not strong enough to prevent the postsynaptic cell from firing (Douglas & Martin, 1990b).

Basket cells (reviewed by Jones & Hendry, 1984) also have specialized outputs: their axon collaterals form several synapses onto the somas of pyramidal neurons, although they can also contact the dendrites of pyramidal neurons (Somogyi et al., 1983) and each other (Kisvárday et al., 1993). It is likely that many of the unremarkable neurons described as ‘smooth stellate cells’ in anatomical studies employing the Golgi technique were actually basket cells (Jones, 1975; Fairén et al., 1984). The Golgi method does not reveal the axonal plexus in all its richness and so the ‘baskets’ characteristic of this cell type (Figs. 2.5H, 2.6H) could easily be missed; also, basket cell axons are frequently myelinated to promote fast transmission of action potentials, and myelinated axons do not show up in the Golgi stain. The largest basket cells can have axonal distributions 1–2 mm in diameter (Somogyi et al., 1983; Jones & Hendry, 1984), although most of the axon terminals seem to be clustered near the soma. Here I will treat all of the regular inhibitory cells as components of the local circuitry.

Inhibitory cells are sometimes called ‘local circuit neurons’ but this is misleading as collaterals of the pyramidal neurons also make a large contribution to the local circuitry. Pyramidal neurons typically give rise to axonal arborizations within the vicinity of the soma as well as long-range collaterals that can extend for several millimetres within the cortex (reviewed by Feldman, 1984). I assume that a distinction can be made between these two distributions: the former are part of the local circuitry, whereas the latter are concerned with longer-distance communication and are discussed later in the chapter. Both collateral systems make synapses with both pyramidal and inhibitory neurons (Kisvárday, Martin, Freund, Maglóczky, Whitteridge & Somogyi, 1986; McGuire, Gilbert, Rivlin & Wiesel, 1991; Keller & Asanuma, 1993).

Figure 3.3: The spatial distribution of excitatory and inhibitory connections in the local circuitry. The graph shows the hard weights of the connections as a function of lateral distance from the soma of the source neuron.
The neurons at each point in the cortex are assumed to exert excitatory and inhibitory effects on nearby neurons according to the spatial distributions shown in Figure 3.3. The abscissa of the graph indicates lateral distance from the soma; the ordinate indicates the strength of the excitation or inhibition, corresponding to the hard weights of the connections involved (p. 42). Both distributions are assumed to be approximately Gaussian (bell-shaped), so that the effect is maximal on immediately adjacent neurons and declines with increasing distance from the soma. (Consideration of the axonal spread of inhibitory neurons suggests that the width of the inhibitory distribution, at least, can depend on the cortical layer; in particular, it is more compact in layer 4.) The excitatory effect is caused by localized collaterals of the pyramidal neurons; these can be called short excitatory connections to distinguish them from the longer collaterals. Both the short excitatory connections and the inhibitory connections are assumed to be of fixed and unmodifiable strength (that is, they have no soft weights). The most important assumption is that the inhibitory connections spread over a greater distance than the short excitatory connections, but that the short excitatory connections exert a more potent effect than the inhibitory connections within their limited range. Hence, a small cluster of active neurons (including both excitatory and inhibitory cells) exert a net excitatory effect on neurons in the immediate neighbourhood of the cluster, but a net inhibitory effect at a somewhat greater range. (It is conceivable that diffusion of a neurotransmitter in the extracellular space could contribute to the short-range excitation; this would reduce the amount of wiring needed, but would probably be too slow.) Recent intracellular recordings from pairs of neurons in deep-layer sensorimotor cortex in vitro appear to support the claim that excitatory interaction is very powerful over short distances but that inhibition has a greater lateral spread (Thomson & Deuchars, 1994).

Circuitry of this nature has been studied by Grossberg (1976; Ellias & Grossberg, 1975; Levine & Grossberg, 1976) and Kohonen (1982, 1988) and its effects are easy to understand. Assume that both excitatory and inhibitory neurons have similar pattern-selectivity properties, as described in the previous section. This is made plausible by studies of cat visual cortex that reveal no obvious differences between the two cell classes in receptive field characteristics (Gilbert & Wiesel, 1979; Martin, 1984). Assume also that nearby neurons in the cortex tend to have similar characteristic patterns. (We shall consider how this comes about in the discussion of plasticity later in the chapter.) Now, suppose that a particular input pattern causes neurons in a certain localized group to respond more strongly than other neurons in the region. The positive feedback of the short excitatory connections between the pyramidal neurons reinforces and accentuates the activity of the group, including the activity of the group’s inhibitory neurons; and the latter cells suppress the activity of neurons in the surrounding region. Hence, the local circuitry promotes islands of activity in the cortex.

Figure 3.4: (A) Feedforward excitatory input to a field of neurons (represented here as a one-dimensional line) causes them to respond to various degrees (represented by the bar heights) depending on how well their characteristic patterns match the input. (B) The effect of the excitatory and inhibitory interactions between the neurons is to amplify the responses of the well-matched neurons and to suppress the responses of the poorly-matched ones, resulting in discrete ‘islands’ of neurons with high firing rates.
containing perhaps 10–100 neurons, at places where the neurons have characteristic patterns closely matching the input to the region. This is illustrated schematically in Figure 3.4.

To understand the usefulness of this we need to consider its effect on the pattern-selectivity of the individual neurons. There are three main effects. First, the increase in the firing rates of the neurons in the islands of activity makes their response less dependent on the absolute magnitude of the input pattern. Even a rather small amount of excitatory input to a neuron can cause it to give a vigorous response, provided it matches the cell’s characteristic pattern (and the cell’s neighbours’ characteristic patterns) sufficiently well to engage the positive feedback effect of the short connections. In other words, the local circuitry performs a limited amount of gain control in the response of cortical neurons. Second, the inhibitory component of the local circuitry has the effect of sharpening the tuning of neurons. If the input pattern to a neuron is only a poor match to its characteristic pattern, the neuron is inhibited by other neurons that match the input better. Hence, the selectivity of each neuron’s response for its characteristic pattern is enhanced.

The third function of the local circuitry is to make the sharpness of a neuron’s tuning less dependent on the absolute magnitude of its excitatory input. The response of a neuron with soft weight vector \( \mathbf{w} \) to input pattern \( \mathbf{y} \) is related to the inner product \( \mathbf{w}^T \mathbf{y} = ||\mathbf{w}|| ||\mathbf{y}|| \cos A \), as shown in the previous section. Thus, in the absence of inhibition, the neuron’s response will be just as high to a strong but poorly matched input pattern (large \( ||\mathbf{y}||, \) small \( \cos A \)) as to a weak but well matched input pattern (small \( ||\mathbf{y}||, \) high \( \cos A \)). Hence, the tuning of the neuron becomes broader as \( ||\mathbf{y}|| \) increases. By amplifying the response of neurons that are relatively well-matched to \( \mathbf{y} \) and abolishing the response of neurons that are relatively poorly matched, the local circuitry causes the width of each neuron’s tuning curve to be less dependent on \( ||\mathbf{y}|| \). These three effects are demonstrated in the simulation described in Chapter 4.

Neurophysiologists have long believed that inhibition has an important function in determining the response properties of neurons, but the importance of intracortical excitation is less generally appreciated. Douglas and Martin (1991a) made intracellular recordings of the membrane potential changes in cat visual cortex neurons following electrical stimulation of the thalamocortical pathway. They found that the excitatory potential change following stimulation typically had two components: a short-latency excitation caused by direct input from the thalamus, followed by a more prominent and long-lasting excitation presumably caused by input from other cortical neurons. They proposed that under normal physiological conditions a neuron in the visual cortex that responds strongly to a stimulus is mainly driven not by input from the thalamus but by input from other cortical neurons with similar preferred stimuli. This explanation accounts for the seemingly paradoxical finding of intracellular experiments that the stimulus causing the maximal inhibitory input to a neuron is the same as (or very similar to) the stimulus that causes the neuron to fire most strongly (Ferster, 1986; Douglas, Martin & Whitteridge, 1991; Sato, Daw & Fox, 1991). The resolution of the paradox is that the strong excitation from neighbouring neurons more than offsets this inhibition. The inhibition is more effective for less optimal stimuli, for which the intracortical excitation is relatively weak.

I would like to conclude this section with a speculation about the different types of inhibitory neuron. The theory assumes that the magnitude of the inhibitory output of a localized group of neurons decreases with distance, approximately according to a Gaussian distribution (Fig. 3.3). Perhaps no individual inhibitory cell produces connections that have this distribution, but instead the distribution results from the combined effects of different types of cell. For example, chandelier cells, which have very potent effects on the pyramidal neurons within a restricted neighbourhood, may be responsible for the central peak of the distribution, whereas basket cells, which have much more widespread collaterals, may be responsible for the weaker periphery (see Fig. 2.6D & H).

### 3.4 Feedforward Activation in Layers 2–4

The previous two sections were concerned with information processing in local clusters of neurons within each cortical layer. This and the following three sections address the flow of information over longer distances, in particular between the different layers and between different cortical areas. Because inhibitory effects are assumed to be limited to the local circuitry, all of this longer-range communication is performed by the excitatory axons and axon collaterals of the pyramidal neurons.
As explained previously, spiny stellate cells are assumed to be equivalent to layer 4 pyramids for our purposes. The present section deals with the more important excitatory connections within layers 2–4, and the input and output pathways of the feedforward type. This thesis is primarily concerned with understanding the canonical cortical circuit, which is a simplified and idealized abstraction of the circuitry present in any actual cortical area. The canonical circuit of the Magpie theory is introduced in stages by a series of schematic diagrams, of which Figure 3.5 is the first and Figure 3.11 the last. This circuit is a simplification of the intrinsic structure of area V1 of the cat and monkey (Lund, 1981, 1988; Gilbert, 1983; Martin, 1984; Valverde, 1985, 1991). Unfortunately, little is known in detail about the circuitry within other areas of the cortex, and so the claim that the canonical circuit of Figure 3.11 applies to all areas is conjectural. However, it does appear to be broadly consistent with what is known of the anatomy of other areas (see references on page 28). Each of the cells in these figures represents the connectivity of all the pyramidal neurons in that layer. Some elements of the circuit are not shown, in particular the inhibitory neurons (which are assumed to have strictly local axon distributions, are omitted, although the short excitatory connections of the local circuitry are included. This diagram shows the feedforward inputs and outputs of layers 2–4.

(As explained previously, spiny stellate cells are assumed to be equivalent to layer 4 pyramids for our purposes.) The present section deals with the more important excitatory connections within layers 2–4, and the input and output pathways of the feedforward type.

This thesis is primarily concerned with understanding the canonical cortical circuit, which is a simplified and idealized abstraction of the circuitry present in any actual cortical area. The canonical circuit of the Magpie theory is introduced in stages by a series of schematic diagrams, of which Figure 3.5 is the first and Figure 3.11 the last. This circuit is a simplification of the intrinsic structure of area V1 of the cat and monkey (Lund, 1981, 1988; Gilbert, 1983; Martin, 1984; Valverde, 1985, 1991). Unfortunately, little is known in detail about the circuitry within other areas of the cortex, and so the claim that the canonical circuit of Figure 3.11 applies to all areas is conjectural. However, it does appear to be broadly consistent with what is known of the anatomy of other areas (see references on page 28). Each of the cells in these figures represents the connectivity of all the pyramidal neurons in that layer. Some elements of the circuit are not shown, in particular the inhibitory neurons (which are assumed to have strictly localized axonal distributions within each layer) and the nonspecific fibre systems. Also, the diagrams do not specify the numerical details of the connectivity, such as the horizontal distances and the numbers of fibres involved in each of the pathways. Much more detailed anatomical and physiological evidence from several different cortical areas will be required to refine the model of the canonical circuit proposed here.

It is now well-established that the areas of the macaque monkey visual cortex are hierarchically organized (Van Essen, 1985; Zeki & Shipp, 1988; Felleman & Van Essen, 1991; Kaas & Krubitzer, 1991; Young, 1992). This hierarchy is diagrammed in Figure 2.9. There is accumulating evidence for a similar organization of other regions of the cortex in a variety of species (Galaburda & Pandya, 1983; Symonds & Rosenquist, 1984; Pandya & Yeterian, 1985; Barbas, 1986; Friedman, Murray, O’Neill & Mishkin, 1986; Seltzer & Pandya, 1989; Rouiller, Simm, Villa, de Ribaupierre & de Ribaupierre, 1991; Coogan & Burkhalter, 1993). As described in the previous chapter (p. 25), the hierarchy is defined by the anatomical organization of the corticocortical pathways that interlink the
areas. It appears that nearly all of the pathways are reciprocal. Pathways carrying information from lower-level to higher-level areas in the hierarchy are called feedforward (Van Essen & Maunsell, 1983). These connections originate from neurons in layers 2–4 of the source area and terminate in layer 4 and lower layer 3 of the target area. The input to the cortex from the major nuclei of the thalamus also terminates mainly in the middle layers (Jones, 1985), and so these connections can also be regarded as feedforward. The pathways that reciprocate the feedforward connections, carrying information from higher to lower levels of the hierarchy, are called feedback. These connections have a different laminar organization and are discussed later in the chapter. A minority of the connections between different areas, which link areas at the same level in the hierarchy, have a laminar organization that is neither of the feedforward or feedback type. These lateral connections are also discussed later.

Note that the term ‘hierarchy’ is used in a rather liberal sense here. Each level of the hierarchy can send information to several higher levels, not just to the next level up. For example, area MT of the visual cortex receives information from V1 both directly and indirectly via V2 (Fig. 2.9). Similarly, input from the thalamus is not restricted to the lowest level of the hierarchy: although the lateral geniculate nucleus projects almost entirely to V1 in primates, the pulvinar nucleus provides input to most if not all visual areas (Jones, 1985). Note also that there does not appear to be a single area at the highest level of each perceptual system where all the processing in that system ultimately converges. Rather, there are several different polymodal areas, in which neurons respond to stimuli in more than one modality. Ultimately, the highest-level areas send feedforward connections into the limbic system, including the amygdala and the hippocampal formation (Amaral, Price, Pitkänen & Carmichael, 1992; Suzuki & Amaral, 1994).

Now consider the canonical circuit. Because the connections from the thalamus and from areas lower in the hierarchy terminate in layer 4 and lower layer 3, and the feedforward projection to higher areas arises from layers 2–4, the neural circuitry that mediates the transformation from feedforward input to feedforward output is located mainly in layers 2–4. The Magpie theory assumes that the intrinsic circuitry of these layers also has a basic hierarchical organization: there is a sequential flow of information from the input in the middle layers up towards the cortical surface (Fig. 3.5). This flow is carried by axon collaterals of the excitatory neurons in these layers, which typically ramify above the level of the soma (e.g., Fig. 2.5B & C). It is assumed that these connections (in combination with the amplifying effect of the local circuitry described earlier) exert a powerful excitatory drive on the neurons they contact, so that, for example, a neuron in layer 3 can be strongly activated by convergent input from neurons in layer 4. Again, the hierarchical scheme does not imply pure sequential processing: feedforward input connections can terminate at more than one level in layers 2–4, output connections to higher areas arise from all levels, and intrinsic connections may skip levels.

Let us now relate this pattern of connectivity to the scheme of representation introduced earlier. Recall that the contents of latent storage are the characteristic patterns of the neurons, and that the firing rates of neurons indicate the membership of their characteristic patterns in active storage. The hierarchical organization of the feedforward connections between areas and of the intrinsic circuitry in layers 2–4 imply that the representations in latent storage are hierarchical. That is, the characteristic pattern of a neuron at one level in the hierarchy is a representation of some pattern of activity that may occur in active storage at lower levels of the hierarchy. The neuron is activated when the appropriate pattern of activity occurs in its presynaptic, lower-level neurons. Thus, the content of active storage in these layers of the cortex is determined primarily by the excitation from the thalamus propagating up through the cortical hierarchy, sequentially activating neurons with appropriate characteristic patterns at each level. For example, visual information from the lateral geniculate nucleus activates neurons in the visual cortex that represent elementary features common in visual images, these neurons in turn activate neurons encoding more complex patterns, and so on, until neurons are activated that represent very elaborate patterns like faces.

Of course, the idea of achieving pattern recognition by means of a hierarchy of pattern-detecting units is an old one. The basic concept is usually attributed to Selfridge’s (1958) Pandemonium architecture. Models of this kind have been proposed repeatedly by psychologists (e.g., Treisman, 1960; Neisser, 1967; LaBerge, 1975; Shiffrin, 1976; McClelland & Rumelhart, 1981; McClelland & Elman, 1986). A good discussion of the advantages and limitations of Pandemonium-like systems from a psychological perspective can be found in Lindsay and Norman (1977, Chap. 7).
The most important objection to a scheme of this kind is its apparent inability to explain certain perceptual constancies, such as our ability to recognize a visual pattern regardless of the position, size and orientation of its image on the retina. Although these abilities may not be as impressive as has commonly been supposed (e.g., Nazir & O’Regan, 1990), the extreme view that there is one neuron for every distinct pattern that can be recognized is clearly absurd. The problem is mitigated to some extent by population coding: because neurons are broadly tuned they can tolerate some degree of distortion in their stimulus patterns. Hence, only a limited number of ‘exemplar’ patterns need to be represented in latent storage (as the characteristic patterns of individual neurons) to enable a continuum of patterns to be represented in active storage (by groups of active neurons). Despite this, there is still likely to be an unacceptably rapid increase in the number of neurons needed as one moves up through the hierarchy. Some additional mechanism is warranted.

A possible solution to this problem was proposed by Kunihiko Fukushima in a pattern-recognition architecture he called Neocognitron (Fukushima, 1980; Fukushima & Miyake, 1982). The details of this architecture need not concern us here, but its essential innovation was a distinction between two types of pattern-detector units, called S-units and C-units, which were organized in alternating layers of a hierarchy. The S-units in the first layer detect particular patterns in a position-specific manner. The C-units in the second layer detect the same patterns as the S-units in the first layer, but are not completely position-specific: each C-unit receives connections from several S-units tuned to the same pattern at slightly different locations within a small region, so that the C-unit responds to that pattern wherever it occurs in that region. S-units in the third layer detect particular patterns of activity in the second layer, and so represent larger and more elaborate patterns than the S-units in the first layer. C-units in the fourth layer respond to the same patterns as S-units in the third layer, but generalize their responses over a certain limited range. This continues for several more pairs of layers, until C-units in the final layer respond to very elaborate patterns irrespective of their position in the input array. The computation performed by this architecture can be seen as a neat compromise between two conflicting requirements in pattern recognition: on the one hand, the desire for position-invariant recognition, and on the other, the fact that recognition of a complicated pattern requires knowledge of the spatial organization of its component parts. Neocognitron achieves position-invariance, thus avoiding an explosion in the number of units at high levels of the hierarchy, but by doing so in a series of stages it (hopefully) preserves as much position information as it requires to recognize patterns at each stage. The architecture also has the practical advantage that all the connections between adjacent S and C layers are short.

The Neocognitron architecture was directly inspired by Hubel and Wiesel’s studies of ‘simple’ and ‘complex’ cells in the visual cortex (Hubel & Wiesel, 1962, 1965a); Fukushima generalized the properties apparently exhibited by these cells to multiple layers of units. In doing so he put into a detailed and quantitative form a conception of the visual cortex that had been hinted at in Hubel and Wiesel’s own discussions (1965a), and in a well-known (and frequently misrepresented) paper by Barlow (1972). Indeed, Neocognitron was initially proposed as a model of the visual cortex (Fukushima, 1980), although Fukushima later shifted his emphasis to practical applications of the architecture in the face of apparently contradictory neurophysiological evidence. We shall be examining this evidence in Chapter 4 when we consider how simple and complex receptive fields are generated in area V1. Note that the terms S-unit and C-unit are preferable to S-cell and C-cell, as the latter are sometimes used in the neurophysiology literature to refer to precise definitions of simple and complex cells (Orban, 1984).

The Magpie theory borrows the distinction between S-units and C-units, and the proposal that they are organized in alternating layers. Specifically, S-units are located in layer 4 of the canonical cortical circuit, where the bulk of the feedforward input arrives, and C-units are located in layers 2–3, to which layer 4 projects and from which the bulk of the feedforward output is produced. There are, however, numerous differences between the present theory and Neocognitron. First, the anatomy of layers 2–4 necessitates a more complicated pattern of connectivity than just two sequential levels of processing: some inputs go to layer 3, some outputs come from layer 4, there might be a hierarchy among S-units within layer 4, and similarly a hierarchy among C-units within layers 2–3. Second, the dynamics of the neurons, their interactions, and their plasticity properties differ substantially between the two theories. Third, C-units in the Magpie theory perform generalizations other than just position-invariance.
It might be thought that the distinction between S-units and C-units would necessitate significant differences in membrane properties. Actually, this is not the case: the properties of the model neuron described at the beginning of this chapter are assumed to be equally applicable to S-units and C-units. It is also assumed that the use of population coding and the operation of the local circuitry are qualitatively identical in layer 4 and layers 2–3. The crucial difference between S-units and C-units is their plasticity: the soft weights of S-units and of C-units are modified in different ways. I therefore postpone further discussion of the distinction until the section on plasticity later in the chapter. For now it suffices to say that S-units in layer 4 of the canonical circuit respond to patterns in the feedforward input, and that C-units in layers 2–3 respond to similar patterns but in a slightly more general way.

### 3.5 Motor Production in Layer 5

In all parts of the cortex, the pyramidal neurons in layer 5 project to a number of subcortical centres, such as the basal ganglia, superior colliculus, pons, red nucleus, brain stem reticular formation, and the spinal cord (Jones, 1984). Although different cortical areas project more heavily to some of these structures than others (the cortical projection to the superior colliculus, for example, may be substantial from certain areas and weak or absent from others) it is not unreasonable to regard the output of layer 5 generally as motor output, since all the destinations are heavily involved in motor control. As noted in Chapter 2, it is traditional to regard the cortex as being partitioned into ‘sensory’ and ‘motor’ areas, but according to the perspective of the present theory it is probably better to make this division on a laminar basis, with layers 2–4 being concerned with the analysis of sensory input and layer 5 with the generation of motor output (see Diamond, 1979, for a similar view). Having said this, neurophysiological and neuropsychological evidence indicate that the posterior part of the frontal lobe is particularly important in the generation of movement, and so it is perfectly reasonable to retain the traditional term *motor cortex* for this region. The motor cortex actually consists of several distinct areas, including the primary motor area (M1), the supplementary motor area, and the premotor area. These three fields are defined by physiological rather than anatomical methods, but M1 is roughly Brodmann’s area 4, and the supplementary and premotor areas are roughly the medial and lateral parts of area 6, respectively (Fig. 2.3). The precise parcellation of the motor cortex is not yet settled, so this tripartite division may have to be revised in the future.

As noted above, many subcortical centres are involved in the control of movement. In fact, nothing illustrates the Darwinian origin of the nervous system more starkly than the design of the motor system: over the course of evolution, new structures did not replace older ones but merely supplemented them. The result in higher mammals is a system of enormous complexity, in which processing centres with a variety of architectures work together to produce movement (see Brooks, 1986, for general review). Most of the subcortical motor centres have direct access to sensory input, many are capable of modifying their function by processes of neural plasticity, and some at least are not mere slaves of the cortex but are quite capable of initiating actions themselves—consider, for example, the reflexes mediated by the spinal cord. Because of the sophisticated and quasi-autonomous nature of the subcortical motor centres, it follows that a given pattern of neural activity in cortical layer 5 may result in different movements at different times, or may not result in any overt behaviour at all, depending on the state of the rest of the motor system. The relation between cortical activity and movement is most direct in the motor cortex: this region has a large population of neurons in layer 5 that project directly to the motor circuits in the spinal cord, some of them (in primates) even synapsing directly with the spinal neurons that innervate the muscles (Kuypers, 1981). (Note that corticospinal neurons are not restricted to area M1, as used to be believed.)

The reader might be thinking that the above discussion does not fit well with the proposal that a unified theory of cognition could be built on a general theory of the cortex. I emphasize again the two caveats of Chapter 2. First, a theory of the cortex will need to be supplemented with theories of other parts of the nervous system; this is probably more true in the field of motor control than in any other domain of psychology. Second, a general theory of the computational mechanisms provided by the canonical cortical circuit is not sufficient for a theory of cognition: it is also necessary to have an understanding of how the mechanisms are applied in particular tasks. The fact that different cortical
areas have layer 5 neurons projecting to different destinations indicates that they are responsible for different aspects of motor control; it follows that the theory of how the cortex actually controls movement is likely to differ from area to area, even if the architecture is the same in all areas. My claim is that a general theory of the cortical architecture could be used as the basis of a unified theory of cognition, not that it would itself be such a theory.

Having made these preliminary points, we now consider layer 5 of the canonical cortical circuit. The axons of the pyramidal neurons in layers 2–4 that carry the feedforward projection to areas higher in the cortical hierarchy give off an extensive system of horizontal collaterals as they pass through layer 5. I assume that these collaterals provide the main excitatory drive to layer 5 neurons; this is shown in Figure 3.6. If layer 5 is concerned with motor production whereas the upper layers are responsible for sensory analysis, then it might be suspected that there are fundamental differences in architecture between these two laminar divisions of the cortex. On available evidence, however, this does not appear to be the case. It is true that pyramidal neurons that produce bursts of action potentials are concentrated in layer 5, but even in this layer they make up a rather small fraction of all pyramidal cells, and they are not included in the present version of the theory (p. 41). The Magpie theory assumes, therefore, that layer 5 represents and processes information in the same way as the upper layers: each layer 5 neuron has a characteristic pattern, defined by the weights of its modifiable input connections, to which it responds maximally, but with a broad tuning curve. The local circuitry in layer 5 is also assumed to be qualitatively identical to that in the upper layers.

The question of whether layer 5 neurons are S-units or C-units has no obvious a priori answer, but I assume that they are C-units on the grounds that cells with complex receptive fields are exceptionally common in layer 5 of the cat’s visual cortex (Martin, 1984; Orban, 1984) and are more common than simple receptive fields in layer 5 of monkey area V1 (Schiller, Finlay & Volman, 1976a). Unlike in layers 2–3, however, the fact that layer 5 neurons have complex receptive fields cannot be taken as unequivocal evidence that they are C-units. This is because they are likely to be second- or third-order neurons, obtaining their main input from the upper layers. An S-unit that receives its input from complex cells might itself have a complex receptive field according to the criteria normally used in physiological experiments.

Because upper-layer neurons are generally in the business of analysing sensory information, it follows that layer 5 neurons, which receive their principal input from those upper-layer neurons, also

Figure 3.6: The same as Figure 3.5 but with the connections of layer 5 added.
represent perceptual patterns. (Note, however, that layer 5 is unlikely to be very important in perception per se, because its output is directed to subcortical motor centres rather than to higher-level cortical areas.) This might seem paradoxical, given that layer 5 is concerned with the production of motor output. But the view that layer 5 neurons encode perceptual patterns is not really incompatible with the view that they encode motor commands, given the sophistication of the subcortical motor centres that was emphasized above. For example, many areas of the visual cortex have layer 5 neurons projecting to the superior colliculus (Gilbert & Kelly, 1975), which is involved in controlling the direction of gaze (Wurtz & Albano, 1980). Whether activity in layer 5 of the visual cortex actually generates eye movement presumably depends on the state of the superior colliculus, which in turn is influenced by the various other sources of information that this nucleus receives.

The claim that cells in layer 5 as well as the upper layers represent perceptual patterns appears plausible enough in the visual, auditory, and somatosensory cortices, but is it really tenable in the motor cortex? Activation of neurons in M1 has, after all, been shown to cause movements in a systematic fashion (see below), and it might seem contrived to regard these cells as representing percepts. I argue that the claim is legitimate, although it is wise to regard layers 2–4 of the canonical circuit as performing a general pattern recognition function rather than a specifically perceptual function. Although this pattern recognition mechanism is indeed concerned with perception in large regions of the cortex, in that it is used to analyse sensory data, in the motor cortex much of the information received from the thalamus actually derives from the subcortical motor centres, specifically from the cerebellum and basal ganglia. The motor cortex is therefore applying the pattern recognition function of the canonical circuit to motor input. Having said this, it is worth pointing out that the motor cortex does receive genuine sensory input, in the form of somatosensory information. This input derives from the adjacent somatosensory cortex, and probably also directly from the spinal cord via the ventral lateral posterior nucleus of the thalamus (Jones, 1986). Physiologically, neurons in the motor cortex are found to respond to somatosensory stimuli, the information deriving from receptors in the skin, joints, and muscles, in a similar manner to neurons in somatosensory cortex (Asanuma, 1975). Thus, instead of regarding somatosensory and motor cortices as being fundamentally distinct regions, the former concerned with perception and the latter with the production of movement, it might be profitable to regard the whole of the cortex surrounding the central sulcus as sensorimotor cortex. The whole of this region is responsible for somatosensory analysis (in layers 2–4) and with the production of motor output (in layer 5), but with segregation of different types of information in different areas, and with greater emphasis on motor and somatosensory functions in the precentral and postcentral areas, respectively. Evolutionary support for this perspective is provided by the fact that the motor and somatosensory representations overlap substantially in primitive mammals like the hedgehog (Lende & Sadler, 1967), which are thought to have brains similar to those of early mammals. A large degree of overlap is conserved in several monotreme and marsupial species (Rowe, 1990).

Studies of area M1 in the behaving macaque monkey by Apostolos Georgopoulos and others have thrown considerable light on the nature of representation in the motor cortex (reviewed by Georgopoulos, 1990; Georgopoulos, Taira & Lukashin, 1993). It is pertinent to consider these studies at this point, although the results were not obtained exclusively from neurons in layer 5 but were from all layers of M1 (Georgopoulos, 1990). These researchers have attempted to elucidate the relation between neural activity and movement, by recording from neurons in the region of M1 concerned with the arm while the monkey makes specific reaching movements to obtain a food reward. For a large majority of neurons in this region, a consistent relation is found between the firing rate of the cell and the direction of the arm movement from a starting location to a target. Each neuron fires at a maximum rate with movement in one direction, the preferred direction of that cell, and fires at progressively lower rates as the angle between the direction of movement and the cell’s preferred direction increases (Georgopoulos, Kalaska, Caminiti & Massey, 1982; Georgopoulos, Schwartz & Kettner, 1986; Schwartz, Kettner & Georgopoulos, 1988). More precisely, if the preferred direction of the i-th neuron is the three-dimensional vector $c_i$ and the direction of movement is $m$, in a Cartesian coordinate system of the space in front of the monkey, the cell’s firing rate $f_i(m)$ is a roughly linear function of the cosine of the angle between $c_i$ and $m$ (Schwartz et al., 1988). This implies that neurons have broad tuning curves, and that a very large number of neurons, with different preferred
directions, fire with each movement. The actual direction of movement $\mathbf{m}$ is very near to the population vector $p(\mathbf{m}) = \sum_i w_i(\mathbf{m}) \mathbf{c}_i$, where $w_i$ is a weighting function derived from the firing rate $f_i$ of the $i$th cell (Georgopoulos, Caminiti, Kalaska & Massey, 1983; Georgopoulos et al., 1986; Georgopoulos, Kettner & Schwartz, 1988). Georgopoulos et al. (1986) used the simple weighting function $w_i(\mathbf{m}) = f_i(\mathbf{m}) - b_i$, where $b_i$ is a constant; various other choices of $w_i$ make little difference to the accuracy of the correspondence (Georgopoulos et al., 1988). In short, the direction of reach is population coded in the arm region of M1: neurons with preferred directions close to the direction of movement make a large contribution, neurons with more poorly-matching preferred directions make a smaller contribution, but no one neuron is essential. This is illustrated diagrammatically in Figure 3.7. Measurements of the evolution of the population vector over time reinforce this interpretation of the evidence: most cells increase their firing rates well before the onset of movement and significantly before the onset of recorded muscle activity, and the direction of the population vector is a good predictor of the direction of the subsequent movement (Georgopoulos et al., 1988; Georgopoulos, Crutcher & Schwartz, 1989). The magnitude of the population vector, which depends on the absolute firing rates of the cells (as opposed to the direction of the vector, which depends on the relative firing rates) may code the speed of the movement, the force exerted, or some
related variable (Schwartz, 1993). Finally, results similar to those described here for the motor cortex have also been obtained in area 5 of the somatosensory cortex (Kalaska, Caminiti & Georgopoulos, 1983), in support of the above arguments against a fundamental division between these two regions of the cortex.

These results provide strong support for the claim that the motor cortex makes use of population coding. However, it cannot be claimed at present that the results are explained in detail by the theory of representation in active storage presented earlier in the chapter. It is important not to confuse the preferred direction of a neuron with its characteristic pattern. The preferred direction of a cell is a vector representing a movement in three-dimensional space; the characteristic pattern is the pattern of input firing rates to which the neuron gives its strongest response. The characteristic pattern is assumed to change only slowly, as a result of synaptic plasticity; the preferred directions of motor cortical cells, on the other hand, have been shown to change dynamically as a function of the location of the arm in space (Kettner, Schwartz & Georgopoulos, 1988; Caminiti, Johnson & Urbano, 1990; Caminiti, Johnson, Galli, Ferraina & Burnod, 1991). To show that the above results can be explained by the Magpie theory will require detailed modelling, and this in turn will require more experimental data on the information that the motor cortex receives as input.

3.6 Feedback Activation From Layer 6

Nearly every feedforward pathway between cortical areas is reciprocated by a feedback pathway. In the canonical circuit assumed in the Magpie theory, feedback connections originate from pyramidal neurons in layer 6 of the source area, and terminate in layers 1 and 6 of the target area. Layer 1 contains few neuron somas; it consists mainly of the apical dendritic trees of pyramidal neurons in all the lower layers as well as a plexus of axonal fibres. The feedback connections from cortical areas higher in the hierarchy probably make up the bulk of the layer 1 plexus, where they contact the apical
dendritic trees of the pyramidal cells of the target area. Layer 6 in each area also gives rise to a substantial projection to the nuclei in the thalamus that provide that area with its main subcortical input (Jones, 1984, 1985). Finally, layer 6 provides an intrinsic projection rising vertically to the upper layers, terminating especially in layer 4. All this is illustrated in Figure 3.8.

Layer 6 is similar to layer 5 in its main source of input: rather extensive horizontal collaterals from the descending axons of neurons in upper layers, including layer 5 neurons. Layer 6 also receives some connections from the thalamus, and from the feedback projections of higher-level cortical areas. Assuming that these collaterals are the principal excitatory connections to layer 6, what is this layer doing? One clue is that layer 6 is a late stage of intracortical processing: although it gets some direct thalamic input, most of the information it receives has already been processed in the upper layers. A second observation is that the collaterals within layer 6 are quite broadly spread (Gilbert & Wiesel, 1979), and so a single neuron in this layer can potentially integrate information from a relatively broad column of the cortex. This is corroborated by physiological studies of area V1 which reveal layer 6 neurons to have exceptionally large receptive fields (Gilbert, 1977; Gilbert & Wiesel, 1979). Finally, as noted above, the feedback outputs of the layer are directed back to lower-level cortical areas.

It is plausible to assume that the feedback corticocortical connections originating from a layer 6 neuron provide excitatory input to neurons in lower cortical areas that are commonly active at the same time as it is. Because the layer 6 neuron is higher in the hierarchy than the neurons to which it projects, its characteristic pattern will generally represent a larger and more elaborate stimulus. Therefore, the activity of the lower-order neuron will be increased if the pattern it encodes is occurring in a context, encoded by the layer 6 neuron, in which it has often occurred in the past. The facilitatory effect of the layer 6 neuron on the lower-order cells can be seen as contextual enhancement of their activity.

As the word ‘enhancement’ suggests, it is assumed that the excitatory effect of the feedback connections on their target neurons is weaker than that of the feedforward inputs. In general, feedback inputs are not sufficiently powerful to activate the neuron strongly by themselves, but their influence is important when they act in concert with the feedforward inputs. In particular, if the membrane potential has been brought close to the action potential threshold by the feedforward inputs, the feedback inputs can raise it above threshold and cause the neuron to fire. The relative weakness of the feedback connections seems plausible on a priori grounds. As a general rule, a pattern-recognition architecture should not make an interpretation of its input only on the basis of what has occurred in that context in the past. It should make such an interpretation only if there is at least some evidence in the input to support it. Put more plainly, an animal should see what is there, not what it thinks should be there, otherwise it will suffer from hallucinations.

The anatomy of the feedback connections is consistent with their being comparatively weak: they terminate mostly on the apical dendritic trees of pyramidal neurons, and so their greater electrotonic distance from the soma would make them less effective than the feedforward inputs, which are located on the basal and proximal apical dendrites (Fig. 3.8). (Unfortunately, the model neuron introduced at the beginning of this chapter cannot treat distant synapses differently from ones close to the soma, because of the isopotential assumption. Instead, the distinction between the two kinds of excitatory connection is modelled by the feedback connections having low-valued hard weights.)

Contextual enhancement can serve two purposes. First, suppose a localized part of the sensory input (for example, a patch of the retinal image) is ambiguous, such that that part of the stimulus matches two characteristic patterns in latent storage equally well. The feedforward input to those two neurons will then activate both neurons equally. However, the neuron with the characteristic pattern most consistent with other parts of the sensory input (that is, with the surrounding context) will receive more feedback activation from higher-order cells, and so will be more strongly activated. This difference in firing rate may be enhanced by the inhibitory action of the local circuitry. Thus, the ambiguous part of the stimulus is interpreted in the way most consistent with context. Second, suppose the feedforward input to a neuron is weak, noisy, or incomplete, such that it can only weakly activate the neuron by itself. Neurons analysing other parts of the stimulus may be activated more strongly, and they will collectively be able to activate higher-order neurons. Feedback from these
higher-level cells may increase the activity of the weakly-active neuron, if its characteristic pattern is consistent with the context, as illustrated in Figure 3.9. Hence, feedback can help to interpret noisy or incomplete sensory input.

If the feedforward input to a neuron is unambiguous, strong, and complete, as will often be the case, the comparatively weak feedback input will have little effect on the neuron’s activity. The only effect it could have in this case would be to decrease the latency with which the neuron responds, provided the feedback input arrived before, or in synchrony with, the feedforward input. This ‘priming’ effect could occur only if some higher-level neurons were activated first, as might happen if there were significant latency differences between different neurons.

There have been very few physiological studies of the function of the feedback pathways. Mignard and Malpeli (1991) studied the pathway from V2 to V1 in cat visual cortex. They temporarily inactivated the layer of the lateral geniculate nucleus (LGN) projecting to V1. This caused a substantial reduction in the responsiveness of neurons in layer 4 to visual stimuli, but cells in layers 2–3 were still well driven. Activity in the upper layers was greatly reduced, however, when the experiment was repeated after destruction of area V2. Mignard and Malpeli interpreted this as showing that activity in layers 2–3 of V1 could be supported either by feedforward input from layer 4 or by feedback input from V2. However, there are two difficulties with this experiment. First, inactivation of the LGN did not completely succeed in abolishing activity in layer 4; it is quite possible that weak residual activity in layer 4 and weak feedback from V2 were in combination sufficient to engage the amplifying effect of the local circuitry, resulting in the robust activity in layers 2–3. This interpretation of the results is consistent with the above proposal, that feedback helps to interpret degraded input. Second, the hierarchical relationship between the first two visual areas in the cat is not nearly as clear-cut as in primates: the pathway from V2 to V1 originates largely from the upper layers rather than from layer 6 (Symonds & Rosenquist, 1984), and there are reasons to regard both V1 and V2 as primary visual areas in this species (Orban, 1984, Sec. 2.2). This is therefore not the best system in which to investigate the function of the feedback connections.

As well as its feedback corticocortical outputs, layer 6 also projects to the parts of the thalamus that provide that cortical area with input. Parsimony suggests that this projection serves the same purpose: it causes enhancement of the activity of the thalamic relay cells whose activity is consistent with the context. This is consistent with a report that activation of the pathway from the visual cortex to the LGN in guinea pigs can induce a long-lasting depolarization in thalamic relay cells, which can facilitate their transfer of information to the cortex (McCormick & von Krosigk, 1992). Although this proposal seems feasible in the case of those thalamic nuclei that relay sensory and motor information to the cortex, it is less feasible in the case of the so-called association nuclei such as the medial dorsal nucleus and the pulvinar, in which the cortical inputs appear to be relatively more important than the subcortical inputs (e.g., Bender, 1983). The function of the thalamic association nuclei is one of the
places in the Magpie theory where the ice is thinnest (although see pp. 78, 173). However, uniformity of the cortex does not logically imply uniformity of the thalamus; it is quite consistent with the theory for different thalamic nuclei to use their cortical inputs in different ways.

Finally, layer 6 sends intrinsic connections up to layer 4 (Fig. 3.8). Although this is also a ‘feedback’ projection in the sense that it passes from higher-order to lower-order neurons, there is reason to believe that its function is rather different from that of the feedback pathways between different cortical areas. Two observations support this. First, these connections do not terminate in layer 1 but in layer 4, and so presumably they contact their target neurons at sites closer to the soma. This makes it less plausible to regard their effects as weak. Second, there is evidence from the cat’s visual cortex that a large proportion of the connections from layer 6 to layer 4 terminate on neurons with smooth dendrites, which are presumably inhibitory (McGuire, Hornung, Gilbert & Wiesel, 1984). These connections may provide the primary excitatory input to a subset of the inhibitory neurons in layer 4, as illustrated in Figure 3.10. These particular inhibitory neurons should therefore possess similar properties to layer 6 cells, including a preference for larger or more complex patterns than other layer 4 neurons. Such properties have not yet been found in layer 4 of V1, but they could have been missed because smooth neurons have small somas and so are probably undersampled in electrode recordings. Neurons that receive connections from this population of inhibitory cells should therefore be suppressed, rather than facilitated, when their preferred stimulus pattern occurs in a particular context. Such an effect is suggested by a study by Bolz and Gilbert (1986) on the ‘end-stopping’ property of some V1 neurons, and is discussed further in Chapter 4. Further anatomical and physiological studies are needed to show that this particular mechanism is not peculiar to area V1 of the cat, and that it is therefore justifiable to include it in the canonical circuit.

An issue still to be addressed is whether layer 6 neurons are S-units or C-units. The proposed function of layer 6 does not make a clear prediction on this point. The empirical evidence from the visual cortex is even harder to interpret than for layer 5. In the cat, simple receptive fields (indicating S-units) dominate layer 6 (Martin, 1984; Orban, 1984), which is consistent with the fact that this layer receives direct input from the LGN, but difficult to reconcile with the input from collaterals of layer 5 pyramidal neurons, or with the input (terminating close to the somas rather than in the apical dendritic tree) from feedback connections from higher areas (Fig. 3.8). This must be regarded as an open question. One possibility is that layer 6 contains different populations of cells participating in separate circuits (e.g., Katz, 1987).

Figure 3.10: Intrinsic connections from layer 6 to layer 4 preferentially terminate on inhibitory neurons. Activation of the layer 6 cells tends to inhibit some neurons in layer 4 that would otherwise be active.
Figure 3.11: The complete canonical circuit. Same as Figure 3.8 but with lateral connections in layers 2–4 added.
3.7 Lateral Activation in Layers 2–4

Anatomical studies of the intrinsic circuitry of the cortex have traditionally emphasized its vertical or columnar organization (e.g., Lorente de Nó, 1949). However, modern studies have revealed quite extensive horizontal or lateral connections within the upper layers of the cortex. The axonal distributions of inhibitory neurons generally extend no more than 500 µm in diameter, except for some basket cells which may have collaterals with a spread of 1–2 mm. Longer horizontal connections, extending up to a few millimetres (Gilbert & Wiesel, 1983; Rockland & Lund, 1983; Martin & Whitteridge, 1984; Matsubara, Cynader & Swindale, 1987; Yoshioka, Levitt & Lund, 1992; Lund, Yoshioka & Levitt, 1993), are collaterals of pyramidal neurons and are excitatory in function (Ts’o, Gilbert & Wiesel, 1986; LeVay, 1988). The target cells of these long excitatory connections are mostly, but not exclusively, other pyramidal neurons (Kisvárday, Martin, Freund, Maglóczky, Whitteridge & Somogyi, 1986; McGuire, Gilbert, Rivlin & Wiesel, 1991; Keller & Asanuma, 1993).

There is a tendency for lateral connections to spread over longer distances in areas at higher levels of the visual cortex hierarchy compared with lower-level areas (Amir, Harel & Malach, 1993); this is presumably related to the greater average size of receptive fields in higher-level areas. Adding these connections gives the final schematic diagram of the canonical circuit proposed by the Magpie theory (Fig. 3.11).

It is assumed in the theory that these long excitatory connections in layers 2–4 are similar in function to the feedback connections between different areas: they connect neurons that are often simultaneously active, and they are responsible for a contextual enhancement effect. That is, if several of the neurons sending long lateral connections to a given cell are active, then they facilitate the activation of that cell. Because of the comparatively long horizontal distances between the neurons that are interconnected in this way, these neurons will generally represent patterns in a different part of the sensory input, which is why the facilitation can be regarded as contextual enhancement. The cell receiving the connections has its activity enhanced when other patterns in the input are consistent with its characteristic pattern.

Some direct evidence that lateral connections interlink neurons that tend to be co-active has been obtained by Löwel and Singer (1992). They raised kittens whose eyes had been misaligned by sectioning an eye muscle, a manipulation known as artificial strabismus. In strabismic kittens most of the cells in area V1 of the visual cortex come to respond to stimulation of only one eye or the other, not both (p. 123). Because of the misalignment of the optical axes in these kittens, there is a much greater likelihood of correlated activity between two neurons responding to stimulation of the same eye than between two neurons responding to stimulation of different eyes. Therefore, if lateral connections join neurons that tend to be simultaneously active, one would expect them to develop between same-eye pairs of neurons but not between different-eye pairs of neurons. Löwel and Singer confirmed this prediction.

As explained in the previous section, feedback connections are assumed to be weak relative to the feedforward connections. The same is assumed of the lateral connections: they are taken to have hard weights with comparatively low values. It should be noted, however, that an in vitro study of cat visual cortex by Hirsch and Gilbert (1991) suggests a more complicated mechanism than just low hard weights; they found evidence for voltage-dependent conductances in the postsynaptic neurons activated by these connections. This finding needs confirmation under more natural in vivo conditions, but it would be consistent with the principle of contextual enhancement, that the excitatory effect of the lateral connections is greater when the postsynaptic neuron is already depolarized by its feedforward inputs.

Lateral connections are not all intrinsic to the cortex. Some of the pathways between different cortical areas have laminar distributions intermediate between the feedforward and feedback types (Van Essen & Maunsell, 1983). These pathways are generally between areas at the same level in the hierarchy (Felleman & Van Essen, 1991). These corticocortical connections are assumed in the Magpie theory to be equivalent in function to the intrinsic lateral connections. Many of the commissural pathways that link symmetrical areas in the two hemispheres through the corpus callosum are of this type (Innocenti, 1986). However, many commissural pathways do not link symmetrical areas; these have laminar distributions similar to the feedforward and feedback pathways.
within each hemisphere (Kennedy, Meissirel & Dehay, 1991), and so should not be regarded as lateral connections.

### 3.8 Plasticity

In his classic treatise on visual perception David Marr (1982) insisted that complex computational systems should be analysed at three fairly independent levels. The first level is concerned with the properties of the functions performed by the system, without regard to how the system actually works; the second level addresses the computations used by the system to carry out those functions; and the third level specifies the physical mechanisms used to implement those computations. In my view this three-level ideal is usually unattainable for biological systems, basically because their Darwinian origin means that their functions, computations and mechanism are inextricably linked. Cortical plasticity, however, appears to be a happy exception to the rule, and so this section is divided into three subsections along Marr lines.

#### 3.8.1 Properties of Cortical Plasticity

Plasticity, which from our perspective may be defined as the acquisition or adaptation of representations in latent storage, is one of the most important aspects of the cortex. We are able to learn and recognize huge numbers of patterns (the appearance and feel of objects, the sounds of words, and the like) and it is obvious that much of this knowledge cannot be built into the cortex but must be acquired through experience. In the context of the Magpie theory this means that the characteristic patterns of many or all neurons in the cortex must be modifiable. One basic constraint on this process is that there does not appear to be any teaching mechanism outside the local neighbourhood of the neuron that can instruct the neuron what characteristic pattern it should adopt (although see p. 146). There would seem to be two possible anatomical substrates for such a teaching mechanism: the monoaminergic and cholinergic fibre systems that innervate the whole cortex, and the feedback connections from higher-level cortical areas and from the limbic system. Although the monoamine and acetylcholine systems do indeed have some modulatory influence on plasticity, as we shall see, the widespread nature of their fibres makes it highly implausible that they could selectively instruct individual neurons. As for the feedback connections, I have already put forward the hypothesis that these are relatively weak connections that are responsible for a contextual enhancement effect; if this is correct, it is unlikely that they have a strong effect on the plasticity of the neurons they contact. In any case, each of the feedback connections must contact large numbers of neurons, and so cannot give specific instructions to individual cells. Therefore, each neuron must acquire a characteristic pattern that is useful to the system as a whole, using only information that is locally available to it: its own activity and the activity of its input connections. A plausible solution to this problem is that neurons adapt their characteristic patterns to match what Barlow (1985) calls ‘suspicious coincidences’ in the activity of their input connections. If a particular input pattern occurs more often than it would by chance, it is likely that the pattern represents something meaningful in the stimulus, and so is worth storing for future recognition. Storing the pattern means that the neuron adjusts its soft weights so as to respond strongly and selectively to that pattern.

If this is correct, then the contents of latent storage in the perceptual areas of the cortex should reflect the statistical properties of ecologically-normal sensory input. Patterns that occur frequently in the observer’s perceptual experience should be frequently represented as the characteristic patterns of neurons in the cortex, whereas infrequent patterns should not. I call this the Empiricist Principle. In Chapters 4 and 5 this principle will be used as a heuristic for explaining some of the receptive field properties that are found in the visual cortex. Besides the properties of natural stimuli, the significance of particular stimuli to the observer should also influence the contents of latent storage, since attentional factors also play a role in cortical plasticity, as proposed later in the chapter. For example, highly-significant visual patterns such as faces may be more heavily represented in the visual cortex than their ecological frequency alone would suggest.

It is certain that not all neurons in the perceptual cortex have to be trained by normal sensory input for them to have normal functional properties. For example, the properties of neurons in area V1 of the newborn, visually-inexperienced monkey are very similar to those of the adult (Wiesel & Hubel, 1974). The visual cortex of the newborn kitten is much less mature, and visual experience is necessary
to establish fully adult-like properties; however, many neurons do have well-established properties at the time of birth (reviewed by Frégnac & Imbert, 1984). Although these findings demonstrate that the neonatal cortex is not a tabula rasa, they do not necessarily imply that neurons have genetically-specified characteristic patterns. It is conceivable that the normal functional properties at the lowest levels of the perceptual cortex are sufficiently simple that neurons can be trained on patterned activity generated in subcortical centres (such as the retina, in the case of vision) before birth (p. 118). However, it is quite possible that at least some neurons do have genetically-programmed characteristic patterns (or pre-specified soft weights, in mechanistic terms). Even if this is the case, there is much evidence that the functional properties of neurons can be substantially modified if the animal experiences an abnormal sensory environment during a period after birth. Furthermore, these modifications seem to be in line with the Empiricist Principle. Evidence in support of this claim from the visual cortex is reviewed in the following chapters. Here, I wish to mention some observations on plasticity in the somatosensory cortex that appear to be consistent with the principle.

As noted in the previous chapter, the maps of the body in the somatosensory cortex of adult monkeys can reorganize in response to long-term alterations of the sensory input (reviewed by Merzenich, Recanzone, Jenkins, Allard & Nudo, 1988; Kaas, 1991). One characteristic of the body map in area 3b of normal monkeys (Fig. 3.15) is that very few neurons respond to tactile stimulation of more than one finger. That is, neurons in the region of the map representing the digits are rarely found to possess receptive fields (RFs) extending across the skin surface of two digits. Clark, Allard, Jenkins and Merzenich (1988) surgically connected the skin surface of two fingers in each of three adult owl monkeys. The monkeys were left for several months to allow the somatosensory cortex to reorganize, and RFs were then examined in the digit region of the map in area 3b. Clark et al. found that the area contained a region within which many neurons responded to stimulation of both of the fused fingers. The RFs of these cells included parts of the skin surface of both digits. The explanation proposed by Clark et al. was that temporal correlations of sensory stimuli are important in determining the RFs acquired by somatosensory neurons. In the normal monkey, the tactile stimulation of two of the digits is presumably not strongly correlated in time, and so the Empiricist Principle suggests that RFs extending across more than one finger should be comparatively rare. In the monkey with surgically fused fingers, there is a much greater probability that stimulation of one of the fingers will be accompanied by simultaneous stimulation of the other, and so RFs extending across both of the fingers should be more common. Another way in which the somatosensory map in area 3b can be manipulated is by conditioning monkeys to perform a task that involves repeated stimulation of a limited skin surface. This results in an expansion of the region of the map analysing that skin surface, at the expense of neighbouring regions (Jenkins, Merzenich, Ochs, Allard & Guic-Robles, 1990; Recanzone, Merzenich, Jenkins, Grajski & Dinse, 1992). In other words, a greater number of neurons acquire RFs covering the stimulated skin. This, too, is consistent with the Empiricist Principle, according to which the statistical distribution of patterns in the sensory input should be reflected in the distribution of characteristic patterns in latent storage. Interestingly, Recanzone et al. (1992) found that substantial map reorganization occurred only if the animal attended to the tactile stimulus; equivalent stimulation while the animal was concentrating on another task was much less effective.

We now consider the difference between S-units and C-units. It was suggested above that neurons adjust their soft weights to respond selectively to ‘suspicious coincidences’ in the activities of their modifiable excitatory input connections. Because of inevitable differences in the latencies of different connections, as well as the discrete nature of action potentials, it would not be reasonable for neurons to learn simultaneous coincidences. Instead, there must be a ‘time window’ of a certain duration, such that excitatory inputs arriving within this window are treated as simultaneous for the purpose of adapting the weights. I propose that the crucial distinction between S-units and C-units is the duration of this period: it is short (on the order of tens of milliseconds) for S-units, and long (on the order of hundreds of milliseconds) for C-units. One loose way to conceptualize this is that S-units learn spatial characteristic patterns, whereas C-units learn spatiotemporal characteristic patterns.

The implications of this are most easily seen in the case of vision. Learning patterns that occur frequently is useful because of the spatial continuity of the world: objects have visual attributes (aspects of shape, colour, texture, etc.) that tend to occur together in the retinal image, and so the
simultaneous occurrence of many of those attributes can be used by the observer to detect the
presence of the object with high reliability. But objects also have temporal continuity: the visual
appearance of an object changes gradually as the object moves relative to the observer. As földiá
(1991) and Rolls (1992) point out, this fact can be used to give neurons a limited ability to generalize
their responses over different views of an object. During the fraction of a second that a C-unit
integrates its inputs, an object can move significantly, enabling the C-unit to acquire a set of weight
values that allow it to respond to that object over a range of different positions. In contrast to the C-
units in the Neocognitron (Fukushima, 1980), which are hard-

3.8.2 Weight Modification Rules

The learning of frequent patterns could be achieved by this principle: if a certain input pattern on the
excitatory connections to a neuron succeeds in depolarizing the neuron above a certain threshold, then
the soft weights of the active connections are increased, so that the neuron is activated more strongly
if the same pattern re-occurs in the future. Weight modification rules of this kind, requiring a
conjunction of presynaptic and postsynaptic activity for weight increase to occur, are called Hebbian
rules after the proposal of the idea by Hebb (1949), although in fact the suggestion goes back to
William James’s discussion of his ‘Elementary Law of Association’ (James, 1890, Chap. 14). A great
many self-organizing connectionist architectures have been proposed based on various different
Hebbian rules (e.g., von der Malsburg, 1973; Grossberg, 1976; Fukushima, 1980; Kohonen, 1982;
Rumelhart & Zipser, 1986).

The plasticity scheme in the Magpie theory is adapted from a theory developed over many
years by Leon Cooper and colleagues. Bienenstock, Cooper and Munro (1982) introduced the
following weight modification rule:

\[
\frac{dw_{ji}}{dt} = \phi(f_i)f_j
\]

where \(w_{ji}\) is the weight of the connection from neuron \(u_j\) to neuron \(u_i\), \(f_i\) and \(f_j\) are the firing rates of
\(u_i\) and \(u_j\), and \(\phi\) is a function of the form shown in Figure 3.12. Note that \(\phi\) changes sign at the value
\(f_i = \theta_M\). In accordance with Hebb’s principle, if the postsynaptic neuron is strongly activated
\((f_i > \theta_M)\) then all the weights are increased in proportion to the presynaptic firing rates. If the
postsynaptic neuron is activated weakly \((f_i < \theta_M)\) then all the weights are decreased in proportion to
the presynaptic firing rates. As it stands, this rule allows connection weights to grow indefinitely; to
cope with this problem Bienenstock et al. allowed the modification threshold \(\theta_M\) to vary as a
nonlinear function of the postsynaptic activity averaged over a period of minutes to hours. When the
average activity of the neuron is higher than some optimum, its \(\theta_M\) is raised, making it more likely
that its weights will decrease. If average activity is lower than the optimum, \(\theta_M\) is lowered, increasing
the tendency for the weights to increase. Bienenstock et al. showed that this rule causes a neuron with
weaken initial pattern-selectivity to become highly selective for a pattern that occurs with reasonable frequency in the input, and that the sliding threshold ensured stability of the weights. The rule of Equation 3.7 is now called the BCM rule. The theory has a well-developed formal basis (Intrator & Cooper, 1992) and gives a very good account of most of the developmental effects on orientation selectivity and ocular dominance that have been observed in the kitten’s visual cortex (Bienenstock et al., 1982; Bear, Cooper & Ebner, 1987; Bear & Cooper, 1990; Clothiaux, Bear & Cooper, 1991). It has also been used to account for plasticity in the somatosensory cortex of adult rats (Beňušková, Diamond & Ebner, 1994).

Before introducing my modification of the BCM theory it is pertinent to review some evidence on long-term potentiation (LTP) in the cortex. LTP is a long-lasting (hours or more) increase in the efficacy of excitatory synaptic connections that was first reported in the hippocampus (Bliss & Lømo, 1973) and has since been the subject of intense study in that structure (reviewed by Bliss & Collingridge, 1993). LTP is typically induced in the hippocampus by delivering a volley of electrical stimuli to the pathway being studied and measuring the resulting change in the effectiveness of those connections in activating the postsynaptic neurons. The potentiation is specific to the connections that are activated, and only occurs if the postsynaptic dendrite is strongly depolarized—that is, it has Hebbian properties. LTP also occurs in the neocortex. For example, using an in vitro preparation of adult rat visual cortex, Artola and Singer (1987) found that high-frequency stimulation of the thalamocortical pathway could induce LTP of the activated synapses of pyramidal neurons in layers 2–4. In the majority of cells LTP could only be induced after the cells had been disinhibited by injection of bicuculline, a GABA_A receptor antagonist, probably because inhibitory circuitry was normally preventing many neurons from responding to the stimulation sufficiently for the potentiation to occur. If the stimulation is delivered to layer 4 rather than the white matter, LTP is reliably induced in layer 3 neurons without the need for disinhibition (Kirkwood & Bear, 1994a). This potentiation satisfies the Hebbian requirements: it is specific to the activated synapses, and can be induced even with low-frequency stimulation if the postsynaptic neurons are strongly depolarized by current injection (Kirkwood & Bear, 1994a). Besides rat and kitten visual cortex, LTP has been reported in vitro in rat sensorimotor (Bindman, Murphy & Pockett, 1988; Lee, Weisskopf & Ebner, 1991) and prefrontal cortex (Hirsch & Crepel, 1991), and in vivo in the motor cortex of adult cats (Iriki, Pavlides, Keller & Asanuma, 1989).

The study of long-term depression (LTD), a decrease in the effectiveness of excitatory connections with properties similar to those of LTP, has lagged behind the study of LTP somewhat (reviewed by Artola & Singer, 1993). LTD has been reported in vitro in rat sensorimotor (Bindman et al., 1988), visual (Artola, Bröcher and Singer, 1990) and prefrontal cortex (Hirsch & Crepel, 1991). LTD in layer 3 of rat and kitten visual cortex has been shown to be specific to the activated synapses (Artola et al., 1990; Kirkwood & Bear, 1994b). Artola et al. (1990) discovered that LTD, like LTP, depends on the postsynaptic membrane being depolarized above a certain level, but that the threshold was lower than that for LTP. This result suggests that the sign of the change of connection weight is a function of the postsynaptic membrane potential: if the cell is at rest there is no change, if it is depolarized above a threshold then LTD is induced, and if it is further depolarized above a higher threshold then LTP causes a net increase in the weight (Artola & Singer, 1993). This is, of course, strongly suggestive of the $\phi$ function of the BCM rule (Fig. 3.12) and Bear, Cooper and Ebner (1987) have explicitly proposed that $\theta_M$ corresponds to the LTP threshold.

The BCM theory cannot be incorporated into the Magpie theory without modification because it is based on a model neuron much simpler than the one introduced at the beginning of the present chapter: the postsynaptic variable in Equation 3.7 is the firing rate, which is taken to be a linear (or roughly linear) function of the neuron’s input. The experiments reviewed above seem to indicate that it is the postsynaptic membrane potential that is relevant (or a variable closely related to the potential—see the following section). In the computer simulation to be described in Chapter 4 I have used the following modification rule for the soft weights of the S-units:

$$\frac{dw_{ji}}{dt} = (P(V_i) - D(V_i)) f_j$$  (3.8)
where \( w_{ji} \) is the soft weight of the connection from neuron \( u_j \) to neuron \( u_i \), \( f_j \) is the presynaptic firing rate, and \( P(V_i) \) and \( D(V_i) \) are the magnitudes of LTP and LTD, respectively, both functions of the postsynaptic potential \( V_i \). Soft weights are constrained to lie in the range \([0, 1]\); a weight that declines to 0 is not reduced further, and a weight that reaches 1 is not further increased. I have used the following definitions of \( P \) and \( D \):

\[
P(V_i) = 2\lambda \tanh^+ \left( \frac{V_i - \theta^+_i}{5 \text{ mV}} \right) \\
D(V_i) = \lambda \tanh^+ \left( \frac{V_i - \theta^-_i}{5 \text{ mV}} \right)
\]

where \( \tanh^+ \) is the positive part of the hyperbolic tangent function and \( \lambda \) is a scaling constant. These functions are illustrated in Figure 3.13. The precise form of the definitions is not crucial; the important point is that LTP and LTD both have definite thresholds (\( \theta^+_i \) and \( \theta^-_i \) respectively) and increase monotonically with \( V_i \) above the threshold (cf. Artola & Singer, 1993). I assume that the LTD threshold \( \theta^-_i \) is constant. The weights are prevented from increasing indefinitely by their being limited to the range \([0, 1]\), as noted above. Some means of adjusting the sensitivity of each neuron to LTP is necessary, otherwise some neurons acquire a set of weights that causes them to respond too often whereas other neurons never respond to any input. This is achieved by allowing the LTP threshold \( \theta^+_i \) to vary as a function of the average postsynaptic potential, in a similar manner to the sliding threshold of the BCM rule. Specifically:

\[
\theta^+_i = \theta_0 + (\bar{V}_i - V_0)\psi
\]

where \( \bar{V}_i \) is the average value of \( V_i \), and \( \theta_0, V_0 \) and \( \psi \) are constants. The average potential is calculated in accordance with:

\[
\beta \frac{d\bar{V}_i}{dt} = V_i - \bar{V}_i
\]

where \( \beta \) is a time constant that is long compared with the rate that \( V_i \) normally changes.

Figure 3.13: The magnitude of LTP (graph above the abscissa) and of LTD (below the abscissa) as a function of the postsynaptic potential (Equations 3.9). The net effect of LTP and LTD is indicated by the dashed curve.
The meaning of Equation 3.10 can be understood as follows. If the average potential of the neuron $\bar{V}$ is below its target value of $V_0$ then the LTP threshold $\theta_i^+$ declines, making it easier for the weights to be increased in value. If $\bar{V}$ is above $V_0$ then the threshold increases, making it harder for the weights to be increased and more likely for them to be decreased (because of the fixed LTD threshold). The constant $\psi > 1$, so that the LTP threshold varies at a faster rate than the average potential. Notice that if the average postsynaptic potential is very low then the LTP threshold can actually fall below the LTD threshold, so that any significant postsynaptic response causes the weights to grow. The sliding threshold mechanism might appear complicated, but it has the computational advantage that each neuron independently adjusts its plasticity so that it eventually acquires a useful pattern. In some other models it is possible for some neurons never to respond to any input pattern and never to acquire a selective response.

Equation 3.8 governs plasticity in S-units: the weight modification is proportional to the presynaptic firing rate, which is assumed not to stay high for long. The distinguishing characteristic of C-units, however, is that they correlate inputs arriving during an extended time window. This implies that the neuron has some mechanism for ‘remembering’ for a certain limited period which of its input connections have been active. In abstract terms I assume that associated with each modifiable input connection to the C-unit there is a trace of the presynaptic firing rate (cf. Sutton & Barto, 1981), with the behaviour shown in Figure 3.14. As the firing rate rises the trace rises with it, but as the firing rate declines the trace falls more slowly. Formally the value of the trace at time $t$ is assumed to be related to the trace at time $t - \Delta t$ by the relation:

$$\tilde{f}_j(t) = \max(f_j(t), \tilde{f}_j(t - \Delta t)e^{-\Delta t/\alpha}) \quad (3.12)$$

where $f_j$ is the presynaptic firing rate, $\tilde{f}_j$ is the trace, and $\alpha$ is a time constant. The weight modification rule for C-units is then:

$$\frac{dw_{ji}}{dt} = (P(V_i) - D(V_i))\tilde{f}_j \quad (3.13)$$

which is identical to the formula for S-units (Eqn. 3.8) except that the trace of the firing rate is used as the presynaptic variable. (In mechanistic terms the trace would almost certainly be stored postsynaptically in real neurons.) The LTP and LTD functions $P$ and $D$ and the behaviour of the LTP threshold are assumed to be identical for both types of unit. The use of a trace mechanism to ‘smear’ activity in time has been used in two recent models of the development of complex receptive fields in the visual cortex (Földiák, 1991; Barrow & Bray, 1992). However in Földiák’s model the trace is a record of postsynaptic activity, while Barrow and Bray’s model has both presynaptic and postsynaptic components.

Until now the discussion has been limited to the feedforward excitatory inputs to a neuron. Plasticity of feedback connections have not yet been investigated in simulations; purely on the grounds of parsimony the soft weights of the feedback connections are assumed to be modified by the
same rules as the feedforward inputs. (Recall that the total weight of a connection is the product of the unmodifiable hard weight and the modifiable soft weight; feedback connections are distinguished by relatively low hard weights.) As for the inhibitory and short excitatory connections that make up the local circuitry, they are assumed to be unmodifiable. The local circuitry does, however, have a strong influence on the characteristic patterns acquired by cortical neurons. If several of a neuron’s close neighbours are active then the short excitatory inputs boost the neuron’s membrane potential, making it more likely that the LTP threshold will be reached. On the other hand, if cells a certain distance away are active then the inhibitory inputs will make it more likely that the neuron will either suffer LTD or not be modified at all. The effect of the local circuitry, therefore, is to impose a strong tendency for nearby neurons to acquire similar characteristic patterns, and for characteristic patterns to vary in a fairly smooth way in the horizontal direction across the cortex (cf. von der Malsburg, 1973; Kohonen, 1982; Swindale, 1982; Barrow & Bray, 1993).

The weight-adjustment processes proposed above can only operate on pre-existing connections between neurons, and so the initial connectivity of the cortical network places important constraints on the characteristic patterns that can potentially be acquired. It is assumed in the Magpie theory that the large-scale organization of thalamocortical and corticocortical pathways are genetically specified. It is also a reasonable assumption that general aspects of the circuitry within each area are established independently of experience. For example, for two neurons of given morphological types and laminar locations, separated by a given distance, there might be a certain fixed probability that a connection forms from one to the other. However, it is clearly out of the question for the detailed connectivity of each neuron to be innate. It is reasonable to assume that the initial connectivity within a localized patch of the cortex is random, within the general probabilistic constraints laid down by genetics. Because processes of weight modification can only operate on connections that already exist, an important issue is whether there is any means by which entirely new connections can be created between neurons in the adult cortex (as opposed to a mere increase in the number of synapses linking two neurons that are already connected). Unfortunately, there is little evidence with which to address this question at present. One relevant datum is that rearrangements of the layout of receptive fields in the primary sensory areas (such as the organization of the map of the body surface in the somatosensory cortex) do not seem to reach beyond the extent of the pre-existing network of excitatory fibres in the area (e.g., Merzenich, Recanzone, Jenkins, Allard & Nudo, 1988; Pons, Garraghty, Ommaya, Kaas, Taub & Mishkin, 1991; Darian-Smith & Gilbert, 1994). This suggests that sprouting of new connections is not very extensive in adults. It does appear to occur locally, however (Greenough, Withers & Anderson, 1992; Darian-Smith & Gilbert, 1994). No mechanism for this is included in the Magpie theory.

### 3.8.3 Molecular Mechanisms

The study of the molecular mechanisms responsible for LTP and LTD is a young and very rapidly developing field, so I will limit my remarks in this section to a very brief summary of the conventional wisdom (for more detail see Bliss & Collingridge, 1993; Artola & Singer, 1993) and then propose a new hypothesis regarding the role of NMDA receptors and intracellular calcium.

Synaptic transmission at many excitatory synapses in the nervous system involves a combination of two classes of glutamate receptor, called NMDA and non-NMDA receptors (the name is from the substance N-methyl-D-aspartate, a selective agonist of the former type of receptor). Whereas the non-NMDA receptors mediate the conventional excitatory effect of the synapse, the activation of the membrane channels associated with the NMDA receptors is dependent on the postsynaptic membrane potential. At the rest potential, these channels are blocked by Mg$^{2+}$ ions; the membrane must be depolarized above a certain threshold to relieve this block (Mayer, Westbrook & Guthrie, 1984; Nowak, Bregeestovski, Ascher, Herbet & Prochiantz, 1984). Thus, the channels come into effect only when the NMDA receptor is activated by neurotransmitter and the membrane is sufficiently depolarized. This dependence on a conjunction of pre- and postsynaptic activity is just what is required for a Hebbian mechanism. The NMDA channels allow an influx of Ca$^{2+}$ ions into the cell (Ascher & Nowak, 1986; MacDermott, Mayer, Westbrook, Smith & Barker, 1986), and it has been proposed that the resulting increase in intracellular Ca$^{2+}$ concentration [Ca$^{2+}$] is the trigger for LTP, both in hippocampus (Collingridge & Bliss, 1987) and neocortex (Bear, Cooper & Ebner, 1987;
Singer, 1987). This is consistent with reports that cortical LTP cannot be induced in the presence of 2-amino-5-phosphonovalerate (APV), an NMDA receptor antagonist (Artola & Singer, 1987; Hirsch & Crepel, 1991; Lee, Weisskopf & Ebner, 1991). In contrast to LTP, LTD can be induced in the presence of APV (Artola, Bröcher & Singer, 1990; Hirsch & Crepel, 1991), although like LTP it does appear to depend on elevation of [Ca\(^{2+}\)], (Bröcher, Artola & Singer, 1992; Hirsch & Crepel, 1992). This has led to the hypothesis that it is the amplitude of the Ca\(^{2+}\) surge following activation of the synapse that determines the sign of modification: LTP is triggered by a strong surge, generally involving activation of NMDA receptors, whereas LTD is triggered by a more moderate surge, perhaps resulting from voltage-dependent Ca\(^{2+}\) channels or release from intracellular stores (Artola & Singer, 1993).

A fashionable hypothesis is that dendritic spines act as ‘calcium compartments’, reducing diffusion of Ca\(^{2+}\) away from the synapse and so maximizing its effect at that synapse and reducing modifications of other, inactive synapses (Müller & Connor, 1991; Koch & Zador, 1993). A recent theoretical study has suggested that changes in certain spine parameters can have dramatic effects on the peak value of [Ca\(^{2+}\)], in the spine following activation of the synapse; this might provide a mechanism for a sliding LTD threshold (Gold & Bear, 1994). If spines do have this function, it suggests that synaptic plasticity in inhibitory neurons, which lack spines, has different properties from that in pyramidal neurons. It is likely that the excitatory input connections of inhibitory neurons are indeed modifiable, since there are no obvious differences between the two classes of neuron in receptive field properties in the visual cortex (Gilbert & Wiesel, 1979; Martin, 1984) and both classes can receive connections directly from the thalamus in layer 4 (White, 1986). This issue deserves experimental investigation.

I will now propose a rather speculative hypothesis regarding the mechanism of the firing rate trace. Recall that C-units are assumed to keep a trace of the average input firing rate of each modifiable synapse (Eqn. 3.12) and that it is this trace, rather than the firing rate, that determines the magnitude of change when the synapse is modified in strength (Eqn. 3.13). In S-units the firing rate is assumed to determine the magnitude of change (Eqn. 3.8); remember, however, that the firing rate is itself defined as a running average of the frequency of action potentials, so in reality S-units may have a similar trace mechanism to C-units but with with a faster time constant. The most obvious candidate for the trace is the intracellular calcium concentration [Ca\(^{2+}\)]. This is attractive for two reasons: first, the compartmentalization of Ca\(^{2+}\) by dendritic spines can keep a separate trace for each excitatory synapse, at least in pyramidal neurons; second, the Ca\(^{2+}\) signal can persist for a much longer duration than the burst of presynaptic firing that caused it, especially as the channels associated with NMDA receptors can remain open for up to 500 msec (Forsythe & Westbrook, 1988).

But there are two difficulties with this hypothesis. The first is the proposed difference between the two types of unit: the traces in S-units should decay much more quickly than the traces in C-units. This leads to the prediction that NMDA receptors, with their long-duration channels, should be more important in causing the Ca\(^{2+}\) surge in C-units than in S-units. The Ca\(^{2+}\) influx in S-units should mainly result from short-duration sources (probably voltage-dependent Ca\(^{2+}\) channels). This may provide an explanation for the laminar variations in the density of NMDA receptors in the cortex: NMDA receptors are most densely concentrated in layers 1–3 and are more sparse in layers 4 and 6, with layer 5 intermediate (Cotman, Monaghan, Ottersen & Storm-Mathisen, 1987). This is a remarkably close match to the laminar distribution of C-units, proposed on independent grounds earlier in the chapter. It is true that NMDA levels are much higher in layer 4 of the kitten’s visual cortex during the ‘critical period’ of plasticity (Fox, Sato & Daw, 1989); but there is a greater concentration of these receptors in all layers during this period, and layers 1–3 are consistently more rich than layers 4–6 (Bode-Greuel & Singer, 1989). The most important prediction with regard to the visual cortex is that NMDA receptors should make a greater contribution to the Ca\(^{2+}\) surge in cells with complex receptive fields than in those with simple receptive fields, subject to the caveat that S-units receiving inputs from C-units (as might be common in the deep layers) might appear to have complex receptive fields by standard experimental criteria (p. 92). Fox, Sato and Daw (1989) failed to find a correlation between strong NMDA-mediated currents and receptive field properties, although they did not address this question systematically. On the other hand, Nishigori, Tsumoto and Kimura (1990) found that NMDA-mediated currents are less prominent in neurons activated directly by input from the lateral geniculate nucleus than in higher-order neurons; unfortunately they did not assess...
receptive field properties. This prediction is still to be properly tested. A difference in the importance of NMDA receptors is not the only predicted distinction between the two types of unit, however: the intracellular mechanisms that remove the Ca\(^{2+}\) should also act more quickly in S-units than in C-units.

The second difficulty is more profound. If [Ca\(^{2+}\)], serves as a trace of the presynaptic firing rate, as suggested here, then it cannot simultaneously perform the Hebbian function ascribed to it by Artola and Singer (1993), of determining the sign of the weight modification. This can be seen from Equation 3.13: [Ca\(^{2+}\)], is suggested to correspond to the factor \(f_i\), whereas the sign of change (whether LTP, LTD or no change occurs) is determined by another factor, a function of the membrane potential. Mechanistically this factor could be a pair of chemical signals, each triggered by depolarization above a threshold, one causing LTP at every synapse with a high trace and the other causing LTD at every synapse with a high trace. According to this hypothesis it is the threshold-dependent activation of these intracellular signals that endow the neuron with Hebbian properties, not the NMDA receptor. Indeed, the suitability of the NMDA receptor as a Hebbian mechanism has been disputed (Fox & Daw, 1993), primarily on the grounds that NMDA-mediated currents are detectable even with very weak stimuli in the visual cortex in vivo (Fox, Sato & Daw, 1990). I suggest instead that the function of the NMDA receptor, acting in concert with voltage-dependent Ca\(^{2+}\) channels, is to establish a postsynaptic [Ca\(^{2+}\)], trace of amplitude proportional to the presynaptic firing rate. In contrast to the proposal of Artola and Singer (1993), the present hypothesis predicts that high [Ca\(^{2+}\)], by itself is necessary but not sufficient for the induction of either LTP or LTD. There is no evidence bearing on this prediction from the neocortex. A few experiments have been done in the hippocampus, with inconclusive results (Bliss & Collingridge, 1993); the main problem is ensuring that [Ca\(^{2+}\)], is raised where it really matters, inside spines.

### 3.9 Large-Scale Organization of the Cortex

So far we have considered single neurons, local neural circuits, and the flow of information through the cortical layers. We now move another step up and examine the organization of the cortex on a large scale, that is with the division of the cortex into areas and with the mapping of information across each area. Because the cortex has the same organization in all individuals of a species at this coarse scale, it is reasonable to assume that it is innate. Plasticity is assumed to affect the organization of the cortex at a finer scale (within distances of a millimetre or two) at which much greater variation between individuals is found.

The length and breadth of a cortical area are often significantly greater than the length of the connections making up the circuits within the area. This observation leads to the conception of a cortical area as a sheet-like processor in which the computations of the canonical circuit (or its areaspecific specializations) are performed in parallel throughout the sheet. It is natural to suppose that some parameter is mapped across the sheet in a systematic fashion, hence the description of cortical areas as computational maps (Knudsen, du Lac & Esterly, 1987). Most areas of the visual cortex are retinotopically organized: the mapped parameter is direction in the visual field, so that neighbouring neurons respond to stimuli at nearby locations in the visual image (Van Essen, 1985). That is, the receptive fields (RF) of neighbouring neurons are overlapping. In area V1, the mapping of the visual field is not precise at a small scale, since there is a presumably random scatter in the RF positions of neighbouring neurons, but it is clearly discernable over distances in excess of the average RF width (Hubel & Wiesel, 1977). The map is typically distorted, with the central part of the visual field receiving proportionally a much larger region of cortex than the periphery. Average RF size is proportionally smaller in the centre than in the periphery, so that the amount of RF overlap is roughly constant throughout the area (Hubel & Wiesel, 1977). In higher visual areas such as MT, the retinotopic map is obscured by the much greater RF size, but it is still present (Albright & Desimone, 1987). However, the highest areas in the visual cortex appear not to be retinotopic (Gross, 1992), and it is unknown whether any parameter is systematically mapped across these areas.

In the low-level areas of the sensorimotor cortex, neurons respond to somesthetic stimuli that are localized in the body (for example, tactile stimulation of a limited skin surface). The mapped parameter in these areas is location on the body. Thus, the areas are somatotopically organized: the surface of the body is systematically laid out across each area. Figure 3.15 illustrates the body map in
areas 1 and 3b of the primary somatosensory cortex of a macaque monkey (Nelson, Sur, Felleman & Kaas, 1980). As in the visual cortex, the maps are not geometrically accurate, since some parts of the body are mapped onto relatively larger regions of cortex than others. For example, the hands (especially the digits, D1, D2, etc.) are mapped across a larger region than the trunk. Also as in the visual cortex, average RF size is scaled in such a way that the amount of RF overlap is roughly constant (Sur, Merzenich & Kaas, 1980); thus, RFs of neurons analysing the fingers are much smaller than those of neurons analysing the trunk.

In the primary auditory cortex, most neurons are tuned to a particular stimulus frequency, and neurons with the same preferred frequency lie along an isofrequency line across the area (Brugge & Reale, 1985). Orthogonal to the isofrequency lines there is a gradient of preferred frequency, with neurons tuned to very low frequencies at one end of the area and neurons tuned to very high frequencies at the other end. Thus, the mapping here is tonotopic, with frequency as the mapped parameter. This derives from the mapping of frequency along the basilar membrane, the auditory

Figure 3.15: The layout of the ‘body map’ in the primary somatosensory cortex (areas 1 and 3b) of the macaque monkey. Note the proportionately much greater area of cortex devoted to the hands and feet (especially the digits, D1, D2, etc.) than to the trunk. (From R. J. Nelson, M. Sur, D. J. Felleman and J. H. Kaas, 1980, Representations of the body surface in postcentral parietal cortex of Macaca fascicularis, Journal of Comparative Neurology 192: 611–644, Fig. 1. Copyright 1980 Alan R. Liss Inc. Reproduced with permission of John Wiley and Sons.)
receptor surface in the cochlea. Auditory neurons also show various types of binaural interaction, which is presumably related to our ability to localize the direction of sound on the basis of phase and intensity differences between the two ears (Brugge & Reale, 1985). Along each isofrequency line, neurons with similar binaural interaction properties are aggregated into clusters (Middlebrooks, Dykes & Merzenich, 1980).

As well as the coarse mapping of parameters described above, many areas of the cortex also have an interesting organization at a finer scale. Neurons with similar characteristic patterns tend to be clustered together in column- or slab-like groups extending through the depth of the cortex. The clustering of neurons with similar binaural interaction properties in the auditory cortex was mentioned above. In the somatosensory cortex, neurons that give a sustained response to stimulation and those that rapidly adapt are segregated in columns (Sur, Wall & Kaas, 1981). In the primary motor cortex, neurons that activate particular muscle groups are aggregated together (Ansomna, 1975). Several examples of columnar organization in the visual cortex are described in Chapter 4. On the basis of these observations, some researchers have suggested that the cortex is constructed as a mosaic of distinct columnar modules, each module extending through the depth of the cortex and having a width of about 0.5 mm (Mountcastle, 1978; Szentágothai, 1978; Eccles, 1984; Kaas, 1990). This proposal receives some support from anatomical observations of patchiness in the distribution of thalamocortical, corticocortical, and intrinsic connections (e.g., Jones, 1981; Goldman-Rakic & Schwartz, 1982; Gilbert & Wiesel, 1983; Matsubara & Phillips, 1988; Lund, Yoshioka & Levitt, 1993). However, the proposal is difficult to reconcile with several facts: there is a large diversity in column width (Swindale, 1990); there are extensive horizontal connections within the cortex, which argues against the columns being distinct modules; in many cases, there are no sharply-defined boundaries between columns, since neurons near the column borders have intermediate properties (e.g., Livingstone & Hubel, 1984); there are substantial differences between animal species in columnar organization (Purves, Riddle & LaMantia, 1992); and columnar organization can be grossly modified by processes of neural plasticity during development. These facts support the view that columnar organization of functional properties is not caused by genetically-specified modularity in the anatomy, but that it arises from the mapping of information across a sheet-like processor. In particular, columnar organization would result from two factors: the need to represent several different types of information within a small patch of cortex, and a tendency for neighbouring neurons to represent similar patterns. These two requirements will together cause clustering of neurons with similar properties within each patch of cortex. The second factor results from the influence of the local circuitry on plasticity, which was discussed earlier.

The layout of computational maps can be related to the organization of latent storage. Recall that latent storage is hierarchically organized: the characteristic pattern of a higher-order neuron is formed from the modifiable excitatory connections it receives from lower-order neurons. Hence, a characteristic pattern representing the conjunction of a number of sub-patterns can be formed only if connections from neurons representing those sub-patterns converge onto at least one higher-order neuron. If such convergence onto a single neuron does not occur, that conjunction cannot be represented in latent storage. The space of patterns that can be represented in latent storage is therefore limited by the wiring of the cortex.

As noted above, the gross layout of computational maps and the general organization of the pathways between them must be genetically specified. Therefore, the explanation of this organization must be a Darwinian one: the layout of the computational maps is such that, for all the sensory patterns that an animal is likely to need to represent in latent storage, given an ecologically-normal environment, it is very probable that the appropriate convergence of fibres occurs in at least one area of the cortex. That is, natural selection has seen to it that there is a computational map somewhere in the animal’s cortex in which any given ecologically-likely pattern can be stored. For example, faces are visual patterns of great significance to primates, and presumably natural selection has produced a region of the visual cortex in which the appropriate convergence of fibres occurs for the representation of faces by single neurons. The prevalence of retinotopic maps at the lower levels of the visual cortex reflects the fact that spatially-localized patterns are the most important to recognize at the earliest stages of visual perception. (Actually, economy of wiring might be an equally important consideration.) Similarly, somatotopic maps indicate that somesthetic stimuli that are spatially localized in the body are represented in the lower-level sensorimotor areas.
On the other hand, stimulus patterns that are ecologically unlikely may not be capable of being represented in latent storage at all, except as a collection of separate fragments, because no neuron receives the appropriate convergence of fibres. I call this the conjunction problem. The conjunction problem may not be as important as it at first appears. First, the problem occurs only in latent storage, not in active storage. I expand on this below. Second, there may be ways of getting around the problem. Humans, in particular, can re-code a perceptual pattern into a verbal description, or a combined verbal-and-imaginal description, for example. Third, the hippocampus may play a role in overcoming the conjunction problem. Several theorists have suggested that the hippocampus, with its widespread connections to many high-level areas of the neocortex, might act as an associative memory device, storing associations between items of information represented in different cortical areas (Marr, 1971; Teyler & DiScenna, 1986; Rolls, 1989; Squire, Shimamura & Amaral, 1989; Alvarez & Squire, 1994). These theories are based on the proposal that the activation of some of the items in a stored association causes the hippocampal formation to return signals to the neocortex re-activating the remaining items of that association. Indeed it is possible to give a partial account of the amnesic syndrome that follows hippocampal damage in terms of the loss of such a mechanism (Squire, 1992).

The conjunction problem must not be confused with the binding problem. The latter is a problem that arises in perception. It is not my view that the representations underlying perceptual experience consist of the activities of a relatively small number of neurons at the very highest levels of the perceptual hierarchies. I adopt instead the plausible assumption that information represented in many different cortical areas at all levels (except perhaps the very lowest) can contribute directly to on-going thought and behaviour. The binding problem then arises as the problem of linking together all the information relating to a particular object, such as the shape, colour and motion of a cat, which may be distributed around several different areas of the cortex (Crick & Koch, 1990). (If the cat is familiar then there may indeed be a unitized representation of the cat at some particular place in the cortex, or perhaps in the hippocampus, but this could not be the case for, say, a green cat, should such a creature be encountered.) This is different from the conjunction problem because it relates to representations that are computed quickly and transiently, in active storage, rather than slowly and permanently in latent storage. Actually, I am not convinced that the binding problem is as ubiquitous as it appears at first glance. Recall that all of the lower-level areas in each of the perceptual hierarchies appear to have the same topographic organization (retinotopic, somatotopic, etc.) and so all of the attributes relating to a given object, at least within each modality, are implicitly associated by their common topographic location in each area. If the observer is performing a task that only requires information represented in these lower-level areas then there seems to be no logical need for a mechanism to explicitly bind the attributes together. On the other hand, a mechanism is required to deal with the binding problem in higher-level areas, where the receptive fields are large and the topographic mapping breaks down; such a mechanism is proposed in the following section.

3.10 The Focusing Mechanism

Attention clearly has an important influence on perceptual processing, and any comprehensive theory of the cortex must account for its effects. It is now generally recognized that attention is a complex phenomenon, having several different effects, and it is quite possible that several distinct neural mechanisms are involved. In this section I propose one attentional mechanism. There might well be other systems, perhaps below the level of the cortex, that contribute to attentional phenomena besides this one. For example, it is plausible that the automatic orienting reaction that typically follows the sudden onset of a bright or fast-moving stimulus in the periphery of the visual field is mediated by a direct pathway from the retina to the superior colliculus. The mechanism discussed in this section, the focusing mechanism, is proposed to be the main system underlying selective perceptual attention—the selective processing of one stimulus pattern or dimension at the expense of others. For general reviews of the psychology of selective attention see Broadbent (1982), Johnston and Dark (1986), and Shiffrin (1988).
3.10.1 Properties of the Mechanism

Selective attention has traditionally been considered to result from some capacity limitation in the cognitive system, either a communication ‘bottleneck’ somewhere in the system which will only allow information to flow through it at a limited rate, or a limited availability of general ‘cognitive resources’, the nature of which is seldom specified in detail. The most influential of the former kind of theory is the filter model proposed by Broadbent (1958) to account for the results of dichotic listening experiments on the auditory system. In these experiments, the subject receives different spoken messages in each ear through headphones, and is found to be unable to attend to both messages simultaneously. Broadbent proposed that the auditory system consists of two processing stages: a large-capacity first stage in which simple features are extracted from the auditory stimulus, followed by a low-capacity channel in which higher-level processing, such as understanding a spoken sentence, is carried out. Transfer from the output of the feature-analysis stage to the limited-capacity channel is controlled by a filter mechanism. The filter can be set to select information from the first stage on the basis of simple physical attributes, like ear of origin or voice quality. Selection on the basis of more complex attributes, such as the meaning of a message, is not possible because they are not explicitly available at the stage at which the filter operates. It follows from this model that unattended stimuli, information not selected by the filter, undergoes only a limited amount of analysis. This aspect of Broadbent’s model was soon challenged by experiments showing that unattended messages could apparently be processed to quite a deep level (e.g., Deutsch & Deutsch, 1963). For example, while attending to a message in one ear, the subject can sometimes notice the occurrence of his or her own name in the unattended ear (Moray, 1959). (A familiar instance of this phenomenon is attending to one conversation at a party, but becoming aware of a conversation in the background upon hearing one’s name mentioned.) If selection is performed on the basis of simple physical features only, name recognition should occur after the operation of the filter. To accommodate such effects, Treisman (1960) proposed a revision of the filter model in which the processing of unattended stimuli was merely attenuated relative to that of the attended stimulus, rather than prevented completely. Thus, all stimuli may be processed to quite a deep level, but the processing of unattended stimuli is in some way less intense. If an unattended stimulus is highly significant (e.g., one’s own name) or contextually relevant, even this attenuated processing is sufficient for it to be brought into consciousness and become the basis for a behavioural response.

An effect of this kind is assumed in the Magpie theory. As explained previously, perception is considered to involve the spread of activation up through the hierarchy of latent storage. This process is modified by the local circuitry and by feedback and lateral contextual enhancement, but it is basically a feedforward spread of excitation. Suppose that the feedforward activation to each area is a little too weak to maximally activate the neurons in the area. As activation spreads up through the hierarchy of areas, it becomes progressively attenuated, as envisaged by Treisman (1960). For example, neurons in area V1 of the visual cortex are strongly activated by inputs from the lateral geniculate nucleus, neurons in V2 are less strongly activated by their inputs from V1, and so on. Neurons in the higher echelons of the perceptual hierarchies are therefore activated relatively weakly by sensory stimuli. One way this progressive attenuation could occur is if the hard weights of the feedforward connections to each area are limited such that they cannot maximally activate their target cells. This does not seem plausible, however, because the gain-control properties of the local circuitry (p. 47) would counteract the effect. A more likely possibility is that the cortex is subjected to a continuous inhibitory influence by one of the nonspecific fibre systems (Table 2.2). This inhibition could actively dampen the activity across large regions of the cortex, having a relatively greater effect at higher levels of the perceptual hierarchies. A prime candidate for such a mechanism is the fibre system that originates in the raphe nuclei in the brain stem. This system innervates the whole cortex quite densely, and at least in the primate visual cortex it is most dense in layer 4, where the feedforward inputs terminate (Morrison & Foote, 1986). Individual fibres of this system follow very widespread trajectories in the cortex, suggesting a global modulatory function, and the effect of the neurotransmitter released by these fibres, serotonin, appears to be inhibitory on cortical neurons (reviewed by Fallon & Loughlin, 1987).

Another traditional analogy for selective attention in the visual system is a spotlight which can be directed to different regions of the visual field, independently of eye movements (e.g., Posner,
Snyder & Davidson, 1980; Treisman & Gelade, 1980; LaBerge, 1983). A similar analogy compares selective attention with a zoom lens that can take in the whole visual image or can be focused onto a small part of it (Eriksen & St. James, 1986). A somewhat more physiologically plausible version of the idea is that there is a continuous distribution of attention across the visual field but with a peak concentration at some location (LaBerge & Brown, 1989). The spotlight analogy is suggested, for example, by experiments investigating subjects’ reaction times when identifying target stimuli appearing at random locations in the visual field (e.g., Posner et al., 1980). If the location of the target is indicated to the subject by a cue stimulus (e.g., an arrow), appearing just prior to the onset of the target, reaction time is speeded relative to the condition in which a neutral cue (e.g., a plus sign) is given. If the cue indicates a wrong location on a minority of trials, reaction time to the target is increased relative to the neutral cue condition. This can be explained by the idea that attention can be moved around the visual field like the beam of a spotlight, and that the effect of the cue is to direct the spotlight to a particular location. Movements of attention can occur independently of eye movements (Remington, 1980). The size of the attended region of the image (the width of the spotlight beam) is under the control of the observer (Eriksen & St. James, 1986). It appears that only a spatially-contiguous region can be attended—that is, the spotlight beam cannot be split (Posner et al., 1980; Eriksen & St. James, 1986; but see Egly & Homa, 1984).

The focusing mechanism of the Magpie theory is proposed to underlie these effects. The function of the focusing mechanism is to counteract the nonspecific attenuation described above: it facilitates the firing rate of neurons in such a way that activation spreads much more effectively up through the hierarchy of areas, enabling neurons in the highest areas to be strongly activated. Focusing increases the response of affected neurons to their characteristic patterns. Also, by enhancing the suppressive effects of the local circuitry (p. 47), focusing makes the affected neurons’ tuning curves sharper, so that they respond more weakly to inputs that match their characteristic patterns poorly. The effect of focusing is therefore to enhance the selectivity, not just the firing rates, of the affected neurons. The use of this enhancing mechanism is proposed to be mandatory when performing any high level perceptual task, such as object recognition, because in its absence neurons in the higher levels of the perceptual cortex will not be sufficiently strongly activated for the observer to perform these tasks. On the other hand, the mechanism might not be necessary for more low-level tasks, such as discriminating the orientation of a line segment, because these tasks are presumably based on representations at lower stages of the cortex, where neurons are always quite strongly activated by their feedforward inputs (cf. Posner & Dehaene, 1994).

As the name ‘focusing’ indicates, it is assumed that this enhancement need not uniformly affect all neurons in the cortex, but can selectively affect certain groups of neurons. In particular, in the case of the visual cortex, I propose that the mechanism can be made to affect just those neurons, in all visual areas, whose receptive fields (RFs) cover a small, contiguous region of the visual field. This is illustrated in Figure 3.16. In the case of the lower-level areas of the visual cortex, which are all retinotopically mapped and in which neurons have fairly small RFs, this is equivalent to affecting a narrow, column-like group of neurons within each area. In higher areas, where the retinotopic map is more diffuse and RFs are larger, a broader region of each area must be affected. All neurons within the selected regions of each area have their pattern-selectivity enhanced. Because the affected neurons all analyse the same patch of the visual field, the effect is like a spotlight, as suggested by the experiments described above. Stimuli falling within the selected patch of the image are analysed to a high level, because neurons with RFs covering that patch are facilitated. The remainder of the image is only analysed to a rather shallow level, because the neurons with RFs not covering the selected patch are not facilitated, and consequently the spread of activation up through the hierarchy from these parts of the image is attenuated. The observer (or more precisely, the central part of the observer’s cognitive system) has voluntary control over the direction of the focusing mechanism. That is, central cognitive processes can choose to focus on a particular region of the visual field. It is further assumed that the width of the column of neurons affected by the focusing mechanism is variable, as suggested by the experiments described above investigating the spatial characteristics of the visual spotlight. Thus, in the lowest-level areas, the mechanism can affect either a narrow column of cells or a relatively broad region of the area, or perhaps even the entire area.
Practically all of the ‘spotlight’ experiments on selective attention have been concerned with vision. It would be interesting to know whether comparable effects occur in other perceptual modalities. With regard to somesthesis, it is intuitively plausible that one can choose to focus attention on a restricted region of the body surface, and that the size of the selected region is under voluntary control, within limits. If this is the case, I propose that the focusing mechanism can affect the somatosensory system in the same way as the visual system: it can enhance just those somatosensory neurons with RFs covering a restricted part of the body. Because the low-level somatosensory areas are somatotopically mapped, this effect can be achieved by the mechanism enhancing activity in a fairly narrow column of neurons in each area, exactly as in the visual cortex (Fig. 3.16).

It is less easy to extend the theory to the auditory system, because the mapping of parameters across the auditory areas is somewhat less straightforward than the retinotopic and somatotopic maps in the visual and somatosensory cortices. By analogy with the visual cortex, the focusing mechanism should be able to enhance processing in a narrow column of neurons in each low-level auditory area, as illustrated in Figure 3.16. This implies that it should be possible to attend selectively to sounds within a particular band of frequencies, because frequency is one of the dimensions mapped across these areas. This might be able to account for one of the findings of the early experiments on dichotic listening, that it is possible to selectively attend to one of two simultaneously-delivered messages if one is spoken by a male and the other by a female speaker (Broadbent, 1958). But the cue that is most effective for selective attention in dichotic listening is a difference between the two voices in spatial direction of the sound source (Broadbent, 1958). If two messages are delivered simultaneously by the same speaker but from different locations, or to different ears through headphones, the listener can easily listen to one and ignore the other. The basis of spatial localization in audition is differences between the sound sensed by the two ears (primarily differences in phase at low frequencies and intensity at high frequencies, although there are also monaural contributions from the subtle effects of the shape of the head and ears). As noted in the previous section, there are columnar groups of neurons with similar binaural interaction properties in the auditory cortex (Middlebrooks, Dykes &
However, the computation of spatial location in the auditory system is not well understood. This parameter does not appear to be mapped in any straightforward manner across the primary auditory cortex (Middlebrooks & Pettigrew, 1981), in the way that it is, for example, in the tectum of the barn owl (Knudsen, du Lac & Esterly, 1987). In fact, it is likely that spatial location is computed in the subcortical auditory system in mammals (Masterton, 1992). Further research on this issue and on the characteristics of auditory spatial attention would be welcome.

What is the purpose of the focusing mechanism? Why should it be necessary to focus on particular parts of the sensory input and ignore the rest? I suggest that the principal benefit it provides is to overcome the problem of interference between representations in active storage. This is illustrated in Figure 3.17. Suppose that two stimuli, X and Y, when presented separately and in isolation, give rise to two patterns of activity in active storage in one of the higher-level perceptual areas of the cortex (Fig. 3.17A & B). Suppose that X and Y are similar stimuli, such that X and Y have similar representations in active storage in this area. That is, many of the same neurons respond strongly to both X and Y, partly because population coding requires neurons to be quite broadly tuned and partly because of the generalizing properties of C-units. Now consider the simultaneous presentation of X and Y. The representations in active storage overlap, since many neurons are now responding to both stimuli (Fig. 3.17C). The observer cannot distinguish between X and Y, because the pattern of activity in this high-level area effectively represents a single item, the properties of which are a confused mixture of the properties of X and Y. This problem can be overcome if the observer uses the focusing mechanism to selectively process one of the two stimuli. The representation of the other stimulus is now attenuated, so active storage contains an accurate representation of the selected stimulus (Fig. 3.17D). Note that the focusing mechanism can only select one of the two stimuli in this way if they are represented at significantly different locations in the lowest-level sensory areas (Fig. 3.16).

A second use of focusing is to provide a solution to the binding problem introduced in the previous section. This is the problem of linking together all the information in active storage relating to a particular stimulus; focusing solves it by restricting a full perceptual analysis to a single stimulus. As suggested earlier, binding only becomes a serious problem at the higher levels of each perceptual hierarchy, which is where the influence of focusing is more pronounced. At lower levels, the topographic mapping of parameters across each area means that, in general, all the attributes of a given object are implicitly bound together by their common location in each area.

Attention has long been considered to be a necessary condition for the storage of permanent memories (e.g., Moray, 1959). The focusing mechanism might therefore be involved in the regulation of plasticity. Actually, this function of focusing might be secondary to the enhancement of neural activity. As explained earlier, plasticity is assumed to occur only if the neuron’s membrane potential exceeds a threshold, and this would clearly be more likely to occur in the presence of enhancement by focusing, especially in the higher-level areas.
The focusing mechanism gives central cognitive processes some control over the analysis of sensory data. Does this mean that the theory is proposing a homunculus—a sort of black box, labelled ‘central cognition’, that is actually doing most of the work? Not at all. Interference of representations in active storage is a definite problem, selective attention is one of the most striking aspects of our psychology, and it is not implausible that there are computations in the limbic system or prefrontal cortex, say, that are controlling the direction of selective attention. All the focusing mechanism is doing, I propose, is enhancing activity and pattern-selectivity in certain regions of the cortex, and (indirectly) influencing plasticity. Central cognition is selecting from the analyses performed by the perceptual cortex, not doing the analysis itself. I should also make clear that I do not regard the focusing mechanism as the substrate of consciousness, although it is an important influence on the contents of consciousness. My intuition—and that of William James (1890)—is that I am at least weakly conscious of things not presently selected by focal attention. See Iwasaki (1993) for discussion.

We now consider the results of some neurophysiological studies of the effects of attentive fixation on the response properties of neurons in the visual cortex of behaving monkeys. Many of these experiments suggest that neurons in several higher-level visual areas, such as V4, IT, and area 7 of the parietal cortex, respond much more strongly to their preferred stimulus when the animal is attending to that stimulus than otherwise (Mountcastle, Andersen & Motter, 1981; Mountcastle, Motter, Steinmetz & Sestokas, 1987; Haenny & Schiller, 1988; Motter, 1993). This effect may be due in part to a general increase in the responsiveness of visual neurons caused by an increase in the animal’s alertness to visual stimuli. However, at least in some cases the facilitation appears to be specific to the neurons that analyse the attended stimulus (Spitzer, Desimone & Moran, 1988; Motter, 1993). The facilitation also occurs in area V1, although it is less pronounced (Haenny & Schiller, 1988; Motter, 1993), as would be expected from the proposal that the spread of activation up through the hierarchy of areas becomes progressively more attenuated in the absence of focusing. The tuning of affected neurons is sometimes sharper in the presence of attention than in its absence (Spitzer et al., 1988; Haenny & Schiller, 1988).

One widely-cited study of the effects of selective attention on the visual cortex of the behaving monkey is that of Moran and Desimone (1985). They found that V4 and IT neurons with two stimuli within their RFs could respond differently depending on which stimulus the animal was attending to. For example, a V4 neuron that responded only to a red stimulus anywhere in its RF gave a much more vigorous response when the animal attended to a red stimulus than when it attended to a green stimulus, although both stimuli were within the cell’s RF in both cases. That is, the strength of the cell’s response to a red stimulus was determined by whether the animal was attending to that stimulus. Surprisingly, the attenuation of the neuron’s response to the red stimulus was strong only when the animal attended to a green stimulus inside the cell’s RF. When the attended green stimulus was outside the cell’s RF, the cell’s response to the unattended red stimulus in its RF was not greatly attenuated. Moran and Desimone’s findings might be explainable by the assumption that attention was actually affecting a processing stage prior to area V4 in these experiments. Consider two V4 neurons with overlapping RFs, one responding to a red stimulus and the other responding to a green stimulus. Suppose that the animal is focusing on the red stimulus. By assumption, both neurons are facilitated by the focusing mechanism, because both have RFs covering the focused stimulus (the red pattern). However, neurons in areas below the level of V4 are also being facilitated by focusing. In these lower areas, RF sizes are smaller, and because the red and green items are at different locations the neurons responding to them have non-overlapping RFs. Thus, only the neurons responding to the red item are facilitated in the lower areas, not the neurons responding to the green item. The V4 neuron responding to the red item will therefore receive greater excitatory input than the one responding to the green item. This in itself may not cause a great difference to the two neurons’ activity. However, the local circuitry will enhance the difference, assuming the cells are separated by a distance at which inhibition is effective. Conversely, when the animal is focusing on the green stimulus, neurons in V4 responding to it will receive greater excitatory drive from the lower cortical areas than ones responding to the red stimulus, and the latter cells will be more inhibited. The fact that the V4 neuron responding to the unfocused stimulus is more strongly attenuated when the focused stimulus is within its RF than when the focused stimulus is outside its RF may be explained by the limited spatial range
of inhibitory effects. If the focused stimulus lies outside the neuron’s RF, inhibitory cells responding to the focused stimulus may be too distant to affect the recorded neuron.

This explanation of the Moran and Desimone (1985) result requires that focusing should significantly affect neuronal activity in areas below V4 in the visual hierarchy. Significant attentional facilitation of firing rates was not found in V1 by Moran and Desimone (1985), but it has been detected in V1 by Haenny and Schiller (1988) and in V1 and V2 by Motter (1993). The explanation is similar to that proposed by Crick and Koch (1990) and modelled by Niebur, Koch and Rosin (1993), but with one difference: their proposal is that attention affects the temporal structure of spike sequences but not mean firing rates in visual areas below V4, whereas one of the assumptions of the present theory is that the mean firing rate is the only important variable. (It would not be surprising to find that the truth lies between these extremes.)

3.10.2 Anatomical Substrate

Of course, an anatomical substrate is required for the focusing mechanism. I will suggest two candidates. The thalamus has often been proposed to have an attentional function, and there is evidence linking one particular thalamic nucleus, the pulvinar, with visual attention (reviewed by Chalupa, 1991; Robinson & Petersen, 1992). For example, Desimone, Wessinger, Thomas and Schneider (1990) inactivated the lateral pulvinar of an awake monkey by infusion of muscimol, an agonist of GABA<sub>A</sub> receptors, and tested the monkey’s ability to perform a colour discrimination task. They found that when the target stimulus was displayed in isolation, the monkey’s performance was nearly as good with pulvinar inactivation as without. However, when the target was presented together with a distractor stimulus in the other half of the visual field, the monkey made many more errors when the pulvinar was inactivated. One way to interpret these results is that the monkey had difficulty focusing on the target and ignoring the distractor when the pulvinar was inactivated.

Unfortunately, it is difficult to make concrete proposals regarding how the focusing mechanism could be realized by the pulvinar, because very little is known about the physiology of this nucleus. The pulvinar is very large in primates and contains several distinct parts with connections to and from different areas of the visual cortex; most but probably not all of these divisions also receive inputs from the superior colliculus (Jones, 1985). Even in the portions of the pulvinar that receive substantial collicular input, it appears that the input from the visual cortex is relatively more important in controlling the functional properties of neurons (Bender, 1983; Chalupa, 1991). This is, of course, in stark contrast to the other visual nucleus of the thalamus, the lateral geniculate nucleus, which is dominated by the retina. The pulvinar projects to the whole of the visual cortex, although only diffusely to area V1 (Jones, 1985). It would be interesting to know whether these fibres form conventional excitatory synapses, or whether they have some modulatory effect on cortical neurons. I have proposed that the focusing mechanism can affect more than just visual cortex; if this is the case then certain other thalamic nuclei would have to perform a function similar to that of the pulvinar for non-visual cortex.

As an alternative, I suggest that the fibre system originating in the nucleus basalis of Meynert (NBM) in the substantia innominata of the basal forebrain is a plausible candidate for the anatomical substrate of the focusing mechanism. This fibre system (reviewed by Eckenstein & Baughman, 1987; Foote & Morrison, 1987), which uses the neurotransmitter acetylcholine (ACh), innervates all cortical areas. In higher species, the NBM is the only major source of ACh in the cortex (Mesulam, Mufson, Levey & Wainer, 1983). In the rat, individual fibres from the NBM innervate column-like cortical territories less than 1.5 mm in diameter (Price & Stern, 1983; Baskerville, Chang & Herron, 1993). Assuming that a similar specificity holds in higher species, at least in the lowest-level areas of the cortical hierarchy, this would seem to provide the critical requirement that the mechanism is capable of restricting its effects to the neurons that analyse a small patch of the sensory periphery. (The theory allows that in cortical areas at higher levels of the hierarchy, where neurons have large RFs, each cholinergic fibre may innervate a much broader patch of cortex; see Figure 3.16). The cholinergic fibres form synaptic contacts predominantly on dendritic shafts of both pyramidal and smooth-dendrite neurons (de Lima & Singer, 1986; Beaulieu & Somogyi, 1991) in all layers of the cortex (Foote & Morrison, 1987). In primates, although the NBM sends output to the entire cortex, it receives input only from restricted cortical regions that are far-removed from the primary sensory and
motor sites in the amygdala, septum, and hypothalamus (Mesulam & Mufson, 1984; Russchen, Amaral & Price, 1985). All of these centres are involved in either instinctive or central cognitive processes, and so the NBM would seem to be very well placed to have the function of an ‘attention director’.

It might be thought implausible that a minor nucleus like the NBM could be responsible for so important a function. The nucleus has, however, become progressively larger and more differentiated over the course of evolution; in humans it is about 15 mm across (Mesulam & Geula, 1988) and contains about $2 \times 10^5$ cholinergic neurons per cerebral hemisphere (Arendt, Bigl, Tennstedt & Arendt, 1985). Because the surface area of the cortex of one human cerebral hemisphere averages 1200 cm$^2$ with much variation between individuals (Haug, 1987), there are about 1–2 cholinergic cells for each square millimetre of cortex. The number of cholinergic cells actually innervating a given location in the cortex may well be much higher than this, for two reasons. First, as one ascends through the hierarchy of cortical areas the size of the patch of cortex innervated by each cholinergic fibre may increase, as noted above, and so the amount of overlap of these patches should also increase. Second, the theory allows that a single cholinergic cell can send collateral branches to more than one cortical area, provided that all the fibres from a given cell innervate parts of the cortex that analyse the same portion of the sensory periphery (Fig. 3.16). For example, a single cholinergic cell might send collaterals to several areas of the visual cortex; the theory predicts that the innervated portion of each area would have to be analysing the same part of the visual field. Note that the organization of the ACh innervation of the cortex may be innate, because it is assumed that the layout of the computational maps at a coarse scale is genetically specified.

The effects of ACh on cortical neurons (reviewed by Foote & Morrison, 1987; Sillito & Murphy, 1987) appear to be consistent with the present proposal. Sillito and Kemp (1983) studied the effects of iontophoresis of ACh on the stimulus-evoked responses of neurons in cat visual cortex. They found that the great majority of neurons were affected by ACh, two-thirds being facilitated and one-third inhibited. Of the facilitated cells, all showed an increase in the firing rate evoked by their optimal stimulus without any decrease in stimulus-selectivity, and in some cells tuning was actually sharpened. These facilitatory effects are similar to those found in some of the physiological studies of attention, described above. In the case of the inhibited cells, it is likely that the ACh iontophoresis was facilitating neighbouring neurons that inhibited the recorded cells (McCormick & Prince, 1986; Müller & Singer, 1989). The ionic mechanism for the facilitatory effect of ACh appears to be a reduction in the magnitude of certain inhibitory currents (McCormick & Prince, 1986). This would not directly depolarize the neuron, but it would enhance the effect of other excitatory currents by raising and sustaining the neuron’s firing rate. ACh has also been implicated in the regulation of plasticity: Bear and Singer (1986) found that depletion of ACh (together with norepinephrine) in kitten visual cortex severely reduced ocular dominance plasticity (p. 124).

Studies of NBM lesions in animals have not yielded clear-cut results (reviewed by Dunnett, Everitt & Robbins, 1991). The lesion causes behavioural deficits in a variety of tasks, and interpretation is complicated by the possibility that the lesion is not restricted to the cholinergic cells themselves. However, Dunnett et al. conclude that the evidence is most consistent with the lesion producing an attentional dysfunction. A recent study of monkeys with NBM lesions found no deficit in a variety of learning and memory tasks, but did find a deficit in a visual selective attention task (Voytko, Olton, Richardson, Gorman, Tobin & Price, 1994). Finally, degeneration of the NBM and a severe reduction in ACh levels in the cortex are two of the symptoms of Alzheimer’s disease (Coyle, Price & DeLong, 1983). Dysfunction of selective attention has been found in some Alzheimer’s patients (Freed, Corkin, Growdon & Nissen, 1989; Parasuraman, Greenwood, Haxby & Grady, 1992). Alzheimer’s disease has a complex pathology, however, involving much more than ACh depletion.

This outline proposal leaves unanswered many important questions regarding the detailed workings of the focusing mechanism. For example, what is the minimum and maximum width of the affected column of neurons? Are the effects of the mechanism more intense when the affected column is narrow than when it is broad, as Eriksen and St. James’s (1986) zoom lens analogy suggests? If the focus is moved from one location to another does it have to pass through intermediate locations? What are the precise effects of the mechanism on the properties of cortical neurons and circuits? In the absence of a computer simulation it would be inappropriate to make more detailed proposals here.
The focusing mechanism is considered further in Chapter 5 where I attempt to give a qualitative account of some data from experiments on visual search.

3.11 Summary

This concludes the presentation of the general theory. A number of general-purpose computational mechanisms have been proposed for the canonical cortical circuit. For reasons explained in Chapter 2, the cortex was considered purely a perceptual-motor processor for the purposes of this presentation; the function of the cortex in central cognition was ignored. A comparatively simple model of a cortical neuron was proposed to be adequate for the purposes of the theory. Each neuron has a characteristic pattern, encoded by the soft weights of the excitatory connections it receives, which is the input pattern to which it gives a maximal response. Latent storage consists of the characteristic patterns of all the neurons. The firing rate of each neuron, relative to the firing rates of all the other neurons in the local population, is a measure of the match between the neuron’s characteristic pattern and the current input. The firing rates of all the neurons together make up active storage. The representation in active storage of any particular stimulus is population coded, since neurons are quite broadly tuned and several different neurons respond to the stimulus to different degrees.

The circuitry within each layer and within a horizontal spread of about a millimetre is called the local circuitry. It consists of short-range excitatory and longer-range inhibitory components and has the effect of amplifying the responses of neurons to weak inputs, sharpening their tuning, and making the sharpness of tuning less dependent on the magnitude of the input excitation.

The areas of each perceptual system are hierarchically organized, as is the circuitry within layers 2–4 of each area. Activation flows up through this hierarchy along feedforward connections, activating neurons with appropriate characteristic patterns; this constitutes an analysis of the sensory input. S-units in layer 4 encode spatial patterns in their feedforward inputs. C-units in layers 2–3 recognize the same patterns in a more general way. The feedforward output of each area is copied into layer 5, where neurons project to various subcortical motor centres. Output is also copied into layer 6, which is the source of feedback activation to lower-level areas. This feedback, by means of relatively weak connections, causes contextual enhancement, resolving ambiguities and helping to interpret noisy sensory input. Lateral activation within layers 2–4 and between areas at the same hierarchical level has a similar effect.

Plasticity of latent storage is governed by the Empiricist Principle: neurons modify their characteristic patterns to match patterns that occur frequently in their input. This occurs by a modified version of the BCM weight modification rule. C-units acquire their more generalized characteristic patterns by means of a synaptic trace mechanism; the molecular basis of the trace was discussed. Many cortical areas are computational maps, in which certain parameters are topographically organized across the surface of the area. This organization is genetically specified on a coarse scale, and is refined on a smaller scale by the influence of the local circuitry on plasticity. A focusing mechanism, realized by certain thalamic nuclei or by the cholinergic fibre system originating in the nucleus basalis of Meynert, enhances activation and pattern-selectivity in restricted regions of the cortex. This is needed to overcome the problem of interfering representations in active storage, and for solving the binding problem.

Clearly, this theory is highly eclectic: many of its components have been proposed before, although not in their present form and not as parts of an integrated architecture. As Newell says, “Developing a unified theory … is a synthetic enterprise. It is not an enterprise of showing that what psychology has already discovered is wrong. Rather, it is an attempt to synthesize the existing understanding into one unified whole, so we can go forward with the task of fashioning it into a better whole” (Newell, 1990, p. 37).

Although the theory accounts for most of the known anatomical components of the canonical circuit, some things have been left out, including: the diffuse fibre systems that release norepinephrine and dopamine (Table 2.2); the bipolar cells (Figs. 2.5G, 2.6E), which release peptide neurotransmitters; the double bouquet cells (Fig. 2.6B), which have inhibitory collaterals spreading in a narrow column across several layers of the cortex; and the connections with the claustrum, the amygdala, and the nonspecific nuclei of the thalamus. There are many aspects of the functional
properties of neurons which I chose to omit, including adaptation, active dendritic conductances, and oscillatory spiking. It remains to be seen how damaging these omissions prove to be.

In the following two chapters the Magpie theory is applied to the physiology and psychology of visual perception.
Chapter 4

Early Visual Processing

Primates are highly visual animals. Visual perception is the principal means by which we obtain information about our surroundings. Not surprisingly, the visual cortex is a system of immense computational power and complexity. The visual cortex as a whole consists of many distinct areas that are heavily interconnected with one another. Figure 2.9 shows the areas and pathways that make up the visual cortex of the macaque monkey; these areas comprise about 55% of the entire cortex in the macaque (Fellem & Van Essen, 1991). Faced with such complexity, the task of obtaining a general theoretical understanding of how the visual cortex works appears daunting. However, a good way to start is to ask: what is the nature of the basic computational mechanisms provided by the visual cortex? That is, what kinds of representation does it use for encoding information, and what are the elementary computational operations that are realized by the neural circuits of the cortex? In particular, are these computational mechanisms uniform throughout the visual cortex, or are fundamentally different mechanisms used in different cortical areas?

With regard to this last question, it is generally assumed that different modules within the visual system make use of highly specialized mechanisms, tailored by natural selection to be very efficient for their particular tasks. In computational studies of visual architectures it is conventional to incorporate task-specific algorithms into separate modules that process different types of visual information—form, colour, motion, depth, texture, and so on (e.g., Barrow & Tenenbaum, 1978; Marr, 1982). In neurophysiological studies, it is well established that different areas of the visual cortex, and even different column-like subdivisions within each area, are engaged in processing different types of information (Zeki, 1978; Van Essen, 1985; Livingstone & Hubel, 1987; Maunsell & Newsome, 1987; DeYoe & Van Essen, 1988). For example, area MT is known to be important in the processing of visual motion, but apparently not colour, whereas colour is processed in area V4. The segregation of different types of information in distinct areas is not absolute, but it is certainly present to a large extent. It is natural to assume, therefore, that different parts of the cortex contain different and highly specialized computational mechanisms. However, the fact that the visual cortex has a modular construction does not necessarily imply that the basic computational mechanisms are completely different in each module, as explained in Chapter 2. In this and the following chapter I explore the alternative possibility, that there is an approximately uniform system of mechanisms, implemented by the canonical cortical circuit, that are responsible for the analysis of all types of visual information.

I treat visual perception as the main test-bed for the general theory of the canonical circuit described in the previous chapter. There are at least three reasons why vision is the best domain of psychology to use for this purpose. First, vision is important, as noted above. Second, there is vastly more neurophysiological information available on the function of the visual cortex than there is on any other cortical region. Third, as argued in Chapter 2, there is reason to believe that the cortex is extremely efficient for perceptual processing, but is rather less efficient for more central cognitive processes. It is sensible, therefore, to try to understand perceptual processing before moving on to more central faculties. Moreover there is ample evidence from the psychology of mental imagery that vision is not just used for perception: visual representations and processes can also be used in thinking and problem solving (e.g., Metzler & Shepard, 1974; Kosslyn, 1980; Shepard & Cooper, 1982). A thorough understanding of the neural basis of visual perception may therefore provide a good foundation for tackling central cognition.
The structure of the chapter is as follows. The first section is a brief review of the subcortical visual system, summarizing the visual responses of neurons in the retina and the lateral geniculate nucleus. This is followed by an overview of the early processing of visual information in the cortex from the perspective of the Magpie theory. I then consider in some detail how the theory accounts for the properties of the ‘simple cells’ and ‘complex cells’ discovered in area V1 by Hubel and Wiesel (1959, 1962, 1968), and introduce a detailed computer simulation of these receptive field types. The remainder of the chapter provides a more brief and qualitative account of how the low-level visual cortex processes other aspects of the visual input, including motion, spatial frequency, binocular disparity, texture, colour and lightness.

4.1 Review of the Subcortical Visual System

The major parts of the subcortical visual system are illustrated schematically in Figure 4.1. As everyone knows, the eye is a camera containing a lens that forms an image on the retina. The retina contains two kinds of light-sensitive cell or photoreceptor, the rods and the cones. The rods are extremely sensitive and are responsible for vision in poor light. The cones are less sensitive, but they mediate high-acuity vision in good light. Cones are also required for colour vision: in humans and some other primates, there are three types of cone with different wavelength-sensitivities. The photoreceptors feed into some very intricate neural circuitry, which is not considered here (see Sterling, 1990). The output of the retina is generated by the ganglion cells, the axons of which form the optic tract. Each ganglion cell pools information from rods and cones in a small patch of the retina, and so can be influenced by visual stimuli in only a small region of the visual image, the receptive field (RF) of the cell. Each point in the visual field is covered by the RFs of many ganglion cells. In primates, there is a small part of the retina called the fovea in which cones are concentrated and ganglion cells have extremely small RFs, giving very high acuity. Average RF size increases, and acuity correspondingly decreases, with increasing eccentricity (angular distance from the centre of the fovea).

Ganglion cells project mostly to the lateral geniculate nucleus (LGN) in the thalamus, and to a lesser extent to the superior colliculus and a few other subcortical centres. Some fibres cross to the other side of the brain before reaching the LGN, others remain on the same side. This crossing of fibres is organized in such a way that all ganglion cells with RFs on the left side of each retina (analysing the right half of the visual field) project to the left LGN. Similarly, all ganglion cells with RFs on the right side of each retina (analysing the left half of the visual field) project to the right LGN.
Thus, each LGN receives information from the opposite side of the visual field, from both eyes. There is, however, very little mixing of the information from the two eyes within the LGN. Neurons in the LGN project to the visual cortex. In primates, nearly all of this projection terminates in V1, the primary visual area; in the cat, a substantial fraction also goes directly to V2 and V3 (Orban, 1984). Although this pathway through the LGN is by far the most important, there are other routes by which the cortex can receive visual information. In particular, the superior colliculus sends output to another thalamic nucleus, the pulvinar, which in turn projects to the whole of the visual cortex, although only diffusely to V1 (Jones, 1985).

The retinocortical pathway in primates consists of two main parallel channels which are largely independent of each other and in which the neurons have rather different RF properties (for review see Shapley & Lennie, 1985; Livingstone & Hubel, 1987). They are referred to as the M and P streams, because they are segregated in different layers of the LGN, the two magnocellular layers which contain relatively large cells and the four parvocellular layers which contain relatively small cells. Compared with cells in the P channel, cells in the M channel have larger RFs at a given eccentricity, have faster-conducting axons, are more sensitive to moving and low-contrast stimuli, and give more transient responses to stationary stimuli (Dreher, Fukada & Rodieck, 1976; Schiller & Malpeli, 1978; Derrington & Lennie, 1984). Most P cells are also much more selective for the spectral composition of light than M cells (see below). The retinocortical pathway in cats is similarly divided into two main channels, labelled X and Y, in which RF properties are similar to those of the P and M channels in primates, respectively (reviewed by Stone, 1983; Shapley & Lennie, 1985). There are important species differences, however, and it is probably incorrect to identify the X and Y streams with the P and M streams. In particular, Y cells in the cat have some complex, nonlinear summation properties (Hochstein & Shapley, 1976), which are not possessed by most of the M neurons in primates (Derrington & Lennie, 1984). Also, neurons in cats are not strongly selective for the spectral composition of light; cats are nocturnal animals and have poor colour vision. Finally, it should be noted that there are retinal ganglion cells in both cats and primates which have RF properties not clearly belonging to either channel. Most of these cells, called W cells in the cat, project to the superior colliculus (Stone, 1983).

Figure 4.2: (A) The distribution of sensitivity in the receptive field of a retinal ganglion cell can be approximated by a Mexican hat shaped function (solid profile), obtained as the difference of two Gaussian or bell shaped functions (dashed profiles). The cell responds to a pattern of light that activates the ON-subfield (+) more than the OFF-subfield (−), as in B and C but not D.
4.1.1 Luminance Processing in Retina and LGN

Retinal ganglion cells generally respond poorly to uniform light, however strong its intensity. On the other hand, a ganglion cell will give a vigorous discharge if a structured pattern of light lies in its RF. Kuffler (1953) discovered that most ganglion cells in the cat retina have roughly circularly-symmetric RFs, consisting of a central region and a surrounding annulus. So-called ON-centre cells increase their firing rates in response to light increment in the RF centre, and decrease them in response to light decrement. OFF-centre cells give the opposite response, increasing their firing rates to light decrement in the centre and decreasing them to light increment. The RF surround has opposite properties to the centre: ON-centre cells have OFF-surrounds, and vice versa. The two regions are antagonistic, so that uniform light increment or decrement over the whole RF causes little or no response. About half the neurons in each of the two channels (P and M or X and Y) are ON-centre, the other half OFF-centre.

Each of the neurons in the LGN that relays information to the cortex appears to receive input from only a few ganglion cell fibres—perhaps only one in many cases. These relay cells operate in two modes: when the animal is awake and alert, they faithfully relay the activity of their input fibres to the cortex; when the animal is drowsy or in slow-wave sleep, depolarization causes the relay cells to produce rhythmic bursts of action potentials (Steriade & Llinaés, 1988). The mode of operation is probably controlled by the acetylcholine and norepinephrine fibres that the LGN receives from the brainstem (McCormick, 1989). The LGN contains intrinsic inhibitory neurons in addition to the relay cells (LGN circuitry is reviewed by Sherman & Koch, 1990). Like all the major thalamic nuclei, the LGN also receives a large number of fibres from layer 6 of the cortex, the function of which is unclear. In the previous chapter I suggested that these connections produce a contextual enhancement effect on the thalamic relay neurons, similar to that of the feedback corticocortical connections. In spite of these complexities, the relay neurons have RF properties very similar to those of retinal ganglion cells. The principal difference is that LGN cells respond even less well than ganglion cells to a uniform light field (Hubel & Wiesel, 1961); this probably results from the inhibitory effects of the LGN’s intrinsic neurons.

The distribution of light-sensitivity in a ganglion or LGN cell’s RF can be described by a ‘Mexican hat’ shaped function, with a positive central dome and negative annulus for an ON-centre cell, as shown in Figure 4.2. The signs are reversed for an OFF-centre cell. This RF structure can be understood as a tall, narrow, positive dome superimposed on a broad, shallow, negative dome (Fig. 4.2A). Quantitatively, both domes are closely approximated by Gaussian functions, so the RF structure can be modelled by a difference-of-Gaussians (DOG) function (Rodieck, 1965; Enroth-Cugell & Robson, 1966). Qualitatively, it is easy to see what the response will be to various simple patterns. In Figure 4.2B, for example, a spot of light covers the ON-centre, leaving most of the OFF-surround in relative darkness; clearly, the cell gives a vigorous response. The cell also gives a good response to the pattern in Figure 4.2C: the surround is now partly in darkness, so the magnitude of its response is less than that of the centre, and the total response of the cell is positive. The cell does not, however, respond to the pattern of Figure 4.2D, because here the response of the surround outweighs that of the centre. Because of the low spontaneous firing rate of the cell, it cannot signal negative values. However, an OFF-centre cell would respond to this pattern. Note that because of the roughly circular symmetry of the RFs, cells below the level of the cortex are not very sensitive to the orientation of a contour.

4.1.2 Wavelength Processing in Retina and LGN

As mentioned above, in the primate retina there are three kinds of cone with different sensitivities to the spectral composition of light. R, G, and B cones have peak sensitivities for long, middle and short wavelengths, respectively. Their sensitivities are graphed in Figure 4.3. R cones are more common than G cones, and both are much more common than B cones. The ‘RGB’ nomenclature is, of course, meant to suggest that the three cone types respond to red, green and blue light, but there are at least two important reasons why this should not be taken too literally. First, although the three types have peak sensitivities in different parts of the spectrum, their tuning curves are broad and overlapping (Fig. 4.3), so that light of a given wavelength evokes different degrees of activity in all three types. Second,
school physics textbooks notwithstanding, the colour of an object in a complex scene does not depend in any simple way on the spectral composition of the light received from that object (see later). Thus, for example, light that maximally activates R cones does not necessarily appear red. Provided these points are remembered the RGB labels do have a useful heuristic value.

Wavelength sensitivity beyond the level of the photoreceptors has been most thoroughly studied in the LGN (Wiesel & Hubel, 1966; Derrington, Krauskopf & Lennie, 1984; Reid & Shapley, 1992), where the relay neurons presumably have properties similar to those of retinal ganglion cells (DeMonasterio & Schein, 1982). Wiesel and Hubel (1966) placed most LGN cells in the macaque into four classes, with RFs illustrated in Figure 4.4. Type I cells, which make up about 80% of cells in the P channel, have circularly-symmetric, ON-centre or OFF-centre RFs of the kind described in the previous section, but the centre and the surround have different and opposing spectral sensitivities, resulting from input from different cone types (Fig. 4.4A). Four combinations are found. The most common Type I cells have a centre driven by R cones and surround driven by G cones, or vice versa. The less common Type I cells have a centre driven by B cones and a surround driven by the pooled outputs of R and G cones, or vice versa. For brevity, a combination of R and G cone outputs can be denoted Y for ‘yellow’. Thus, there are R-versus-G cells and B-versus-Y cells. Because some of these cells are ON-centre and the others OFF-centre, there are eight possible RF organizations in total. Note that, when stimulated with ‘white’ light, which contains all wavelengths and activates all cone types about equally, the Type I cells have simple ON- or OFF-centre RFs as described in the previous section. About 10% of the cells in the P channel are Type II, which do not have a centre-surround RF organization. Instead, their RFs contain just a single region which combines the outputs of different cone types (Fig. 4.4B). The most common Type II cells give an ON-response to R cone input and an OFF-response to G cone input, or vice versa. Thus, light that stimulates both R and G cones equally, including ‘white’ light, evokes little or no response. The less common Type II cells give an ON-response to B input and an OFF-response to Y input, or vice versa. The remaining 10% of cells in the P channel, and many of the cells in the M channel, are Type III. These have opponent centre-surround RFs, either ON-centre or OFF-centre, in which the centre and surround have identical spectral sensitivities, usually pooling the outputs of all three cone types (Fig. 4.4C). Thus, Type III cells are not very wavelength-selective, and have RFs as described in the previous section. The remaining cells
in the M channel, the Type IV cells, have rather peculiar properties. Their RFs are similar to those of Type III cells, but they have high spontaneous activity, respond with rhythmic bursts of action potentials, and are powerfully suppressed by large fields of long-wavelength light. Perhaps these are Type III cells that are operating in the bursting mode described above.

The RF descriptions given here are actually somewhat idealized: there are a few cells that cannot be placed in any of these categories, and there are markedly different proportions of the various possible combinations of RF properties. Also, many of the cells that receive pooled inputs from R and G cones, denoted here simply as Y input, appear to receive greater input from one type of cone than the other. Ignoring these complexities, it seems that the P channel consists of three subchannels with different wavelength selectivities, one in which R responses are opposed to G responses (containing both Type I and Type II cells), one in which B responses are opposed to Y responses (Type I and Type II cells), and one in which the outputs of all three cone types are pooled (Type III cells).

4.2 Overview of Early Processing

The cortical areas at the lowest levels of the visual hierarchy are responsible for early processing—analysis of the retinal image that is largely unconscious but which is a necessary prelude to such higher-level tasks as object recognition. Actually, the fundamental claim of the present theory (that basically similar mechanisms are operating at all stages of visual perception) implies that there is no fundamental distinction between early and late processing. The best criterion to use is probably RF size: early processing is concerned with relatively localized analysis of the retinal image, whereas later processing stages integrate information from broad regions of the image. Even with this criterion, the choice of exactly where to draw the line between early and late stages is fairly arbitrary. The choice is not important for our purposes, since in the discussion that follows I will not be discussing particular cortical areas systematically. An area-by-area account of visual perception is not feasible at present, because not enough is known about any area except V1. The discussion is organized instead in terms of stimulus dimensions, and I draw upon data from various cortical areas as appropriate.
Following many earlier authors, especially Marr (1976, 1982), I assume that the function of early processing is to generate explicit representations in active storage of useful features of the retinal image. A feature is a simple, spatially-localized visual pattern. In particular, many features are small fragments of contours in dimensions such as luminance, colour, and visual texture. A contour is a place in the image where there is a discontinuity, or step change, in the value of a dimension. For example, the border between a relatively bright and a relatively dim region of the image is a luminance contour. (Strictly, many such features are not discontinuities but regions where there is a steep gradient in the quantity in question.) Contours are useful to represent because they are generally (although not necessarily) caused by important attributes of the viewed scene, such as the edges of objects and the borders between adjoining surfaces. A vast number of features are stored as the characteristic patterns of neurons in the low-level visual areas, and the flow of activation through these areas constitutes a massively parallel analysis of the retinal image.

It is important to realize that features are features of the image, not necessarily features of the viewed scene. The edge of an object or the border of a surface will not necessarily be marked by any contours in the image on the retina, at least not all along its length, because the luminance happens to be approximately the same on both sides of the edge. On the other hand, a contour fragment in the input to the visual cortex may have many different causes, perhaps including noise introduced by the retinal circuitry. From this imperfect data the visual cortex must compute representations that have a high probability of being veridical (that is, of correctly corresponding to aspects of the viewed scene). Such representations are probably not arrived at until the end of early processing. Another important point is that a feature is a visual pattern that typically has values on several different visual dimensions (e.g., a feature may have a particular orientation, a particular size and a particular direction of motion). This usage of the term differs from its use in much of the psychology literature, where a feature is usually a value on a single dimension.

4.3 Luminance

The principal source of information in visual perception is the variation in luminance across the retinal image. The main goal of early processing is to compute representations of the luminance contours and their characteristics. As described above, this process begins in the subcortical visual system: a retinal ganglion cell or relay neuron in the LGN responds well only if a contour lies in its RF. However, these cells are not very fussy about what kind of contour they respond to. For example, they are not very selective for the orientation of the contour. Hubel and Wiesel were the first to investigate the RF properties of neurons in area V1 of the cat and monkey (Hubel & Wiesel, 1959, 1962, 1968). They found that, apart from some neurons in monkey V1 that have circularly-symmetric RFs similar to those of the LGN, most V1 neurons have RFs that are different in a number of ways from those found subcortically. The most striking difference is that most neurons are quite selective for the orientation of a contour. A typical cell in macaque V1 responds strongly to a contour of a particular orientation, but the response declines as the contour is rotated, falling to 70% of the maximal firing rate at an orientation about 20° from the optimal (Schiller, Finlay & Volman, 1976b). At the orientation orthogonal to the optimal one, the contour evokes no response at all.
Hubel and Wiesel discovered that orientation-selective cells in V1 fell into two broad categories, which they termed simple and complex. Simple cells are selective for the position as well as the orientation of a contour: the contour has to be at a particular location within the RF to evoke a maximal response. Complex cells are less selective for position, responding well to a contour of the correct orientation that is positioned anywhere in the RF. Subsequent quantitative investigations (e.g., Schiller, Finlay & Volman, 1976a; Henry, 1977; Henry, Harvey & Lund, 1979; Orban, 1984) have clarified the distinction between the two RF types, and have introduced additional types that have properties intermediate between the two main ones, but the original simple/complex classification has stood up remarkably well. We will first consider the simple type.

The RF structure of simple cells can be plotted using small spots of light, and is found to consist of a number of non-overlapping subfields, each of which gives either an ON- or OFF-response in the same manner as the subfields of retinal ganglion and LGN cell RFs (Hubel & Wiesel, 1959, 1962, 1968). (The RFs of complex cells do not consist of non-overlapping subfields, since ON- and OFF-responses can be evoked in the same parts of the RF. This is the main criterion for distinguishing the two types.) Typical simple RFs are illustrated in Figure 4.5. The number of subfields is generally 1–3, although there can be as many as 6 (Schiller et al., 1976a; Orban, 1984), and they are always elongated. The response of a simple cell to a more elaborate luminance pattern is quite well predicted from its RF map: the optimal stimulus pattern is one consisting of high intensity in the ON-subfields and low intensity in the OFF-subfields. In other words, the subfields of the RF summate in a roughly linear fashion. Thus, the RF in Figure 4.5A responds best to a contour oriented horizontally, with higher intensity at the top of the RF, whereas the RF of Figure 4.5B responds best to a light ‘bar’ on a darker background (that is, two parallel contours of opposite polarity of contrast) oriented at about 30° to the vertical.

Since Hubel and Wiesel’s original discovery there has been considerable controversy regarding the neural circuitry by which simple cells’ response properties are generated (reviewed by Ferster & Koch, 1987; Martin, 1988). The first, now classical model was put forward by Hubel and Wiesel (1962) themselves. They proposed that simple cells received convergent excitatory input from several LGN neurons whose RF centres lie along a line, as illustrated in Figure 4.6A. A slightly modified proposal is that the ON-subfields and OFF-subfields are built from ON-centre and OFF-centre LGN cells, respectively, as in Figure 4.6B. Thus, a contour of the optimal type causes all the LGN neurons to fire, strongly activating the cortical simple cell. If the contour is of the wrong type (in particular, of the wrong orientation) only some of the LGN neurons are activated, and the cortical cell is not
sufficiently depolarized to fire at a high rate. There are several lines of evidence supporting this model. First, the widths of the RF subfields are similar to those of the RF centres of LGN cells at the same eccentricity (Hubel & Wiesel, 1962). Second, the excitatory input to simple cells is tuned for orientation (Ferster, 1986, 1987). Third, cross-correlation analysis of the activity evoked in LGN relay cells and simple cells by visual stimulation suggests that the RF centre of an ON-centre LGN cell with a monosynaptic connection to a simple cell always overlaps one of the ON-subfields of the simple cell, and similarly the RF centre of an OFF-centre LGN cell always overlaps one of the OFF-subfields of the simple cell, as predicted by the model of Figure 4.6B (Tanaka, 1983). In addition, the model is consistent with the fact that all types of neuron, including inhibitory cells, can have simple RFs (Gilbert & Wiesel, 1979; Martin, 1984).

Note that the Hubel and Wiesel model makes no use of inhibition. Most of the alternative proposals have emphasized the role of inhibitory processes in generating orientation selectivity. For example, in the cross orientation model, neurons tuned to different orientations inhibit one another, so that each cell is prevented from responding to its non-preferred orientations by other cells (Sillito, Kemp, Milson & Berardi, 1980). In this model, excitatory input from the LGN would be needed only to impose a weak orientation bias on the cortical cells, which would then be sharpened by inhibition. The main evidence suggesting that inhibition is important in the generation of orientation selectivity is from a series of experiments by Sillito (1975, 1984; Sillito et al., 1980). He found that iontophoresis of small amounts of bicuculline, an antagonist of GABA_A receptors, abolishes orientation selectivity in most simple cells. One difficulty in accepting these findings at face value is that the effects of the bicuculline are unlikely to be specific to one cell. Even if it diffuses over just a short distance, hundreds of neurons could be affected, and if there are any excitatory connections between the neurons in the affected region (as the Magpie theory assumes there to be) even relatively weak excitation from the LGN could drive any neuron into a highly-active state. Also, intracellular recordings do not reveal the presence of strong inhibition at non-preferred orientations, as the cross-inhibition model predicts (Ferster, 1986, 1987), although there are technical difficulties with these experiments too (Martin, 1988).

The present theory explains the RF properties of simple cells by the Hubel and Wiesel model, but supplemented with the effects of the local circuitry. The response properties of a neuron are basically determined by the feedforward excitatory connections that it receives, which in the case of simple cells in V1 are mainly the input fibres from the LGN. The function of the local circuitry is to amplify the neuron’s response to weak input (to low-contrast stimuli in the case of vision), sharpen the neuron’s tuning, and make the sharpness of the tuning less dependent on the absolute strength of the excitatory input. There is reason to believe that this third effect is operating in V1 neurons. The unelaborated Hubel and Wiesel model predicts that neurons should have much broader orientation tuning for high-contrast stimuli than for low-contrast stimuli, since in the high-contrast case LGN cells are firing more strongly, so that even a poorly-oriented pattern could strongly activate the cortical cell (Wehmeier, Dong, Koch & Van Essen, 1989). In fact, the sharpness of orientation tuning is fairly independent of stimulus contrast (Sclar & Freeman, 1982). Recall that the local circuitry is proposed to consist of short-range excitation and longer-range inhibition, such that each cell receives maximal excitation and inhibition from its immediate neighbours. Because of the smooth mapping of orientation selectivity across V1 (see below), the theory predicts that the inhibitory input to each simple cell should be tuned to roughly the same orientation as the simple cell itself, in contrast to the cross-orientation model. This is borne out by intracellular recordings of cat simple cells (Ferster, 1986; Douglas, Martin & Whitteridge, 1991).

Now consider complex cells. Unlike simple cells, complex cells have response properties that are highly nonlinear: the cell’s response to an elaborate pattern is not predictable in a straightforward way from an RF map plotted with small spots of light (Hubel & Wiesel, 1962, 1968). Complex cells respond poorly or not at all to spots of light, and those that do respond are found in most cases to give both ON- and OFF-responses throughout the RF. However, like simple cells, they respond vigorously to a stimulus pattern of a certain type and orientation, especially if it is moving. Unlike simple cells, they respond well to their preferred stimulus throughout their RFs, and so are not as sensitive as simple cells to stimulus location. Also, in comparison with simple cells, complex cells have larger RFs on average at each eccentricity, are less sharply tuned for orientation, have higher spontaneous

Hubel and Wiesel (1962) proposed that complex cells obtain their properties by convergent excitatory input from simple cells. All the simple cells are assumed to encode the same orientation and to have overlapping but slightly displaced RF positions. As a stimulus of the appropriate orientation is swept across the RFs of the simple cells, they all respond, activating the complex cell. This model is supported by detailed quantitative studies of complex cell RFs, which suggest that they are constructed from a number of overlapping subunits (Movshon, Thompson & Tolhurst, 1978b; Heggelund, 1981; Spitzer & Hochstein, 1985; Szulborski & Palmer, 1990). The subunits themselves have simple RFs (with segregated ON- and OFF-subregions that summate linearly), and the response of the complex cell is obtained by thresholding the outputs of each of the subunits and then summing their outputs linearly. This scheme is illustrated in Figure 4.7. The most straightforward way to interpret this model is that the subunits are indeed simple cells, and that the rectification of the subunit outputs results from the fact that the simple cell firing rates cannot be negative. Note that the simple cells might have different spatial arrangements of ON- and OFF subfields, provided that they are all tuned to the same orientation. This means that the complex cell may respond well to light/dark edges of either polarity of contrast. This is significant, because there is psychophysical evidence that the human visual system has contour-detectors with this property (Grossberg & Mingolla, 1985; Shapley & Gordon, 1985).

Hubel and Wiesel’s assertion that complex cells’ RFs are built from simple cells’ RFs in a hierarchical manner has come in for a good deal of criticism, since there are certain kinds of evidence that appear to indicate that complex cells obtain their properties directly from LGN input. The present theory assumes the basic principle of the hierarchical scheme, that complex cells receive their dominant excitatory input from simple cells (or from lower-order complex cells). This does not exclude the possibility that many complex cells receive a fraction of their excitatory input directly from the LGN, provided that this fraction has a comparatively minor role under normal physiological conditions. The hierarchical scheme has the virtue of explaining the properties of complex cells without requiring very elaborate biophysical properties, which is consistent with the model neuron of the Magpie theory. If the subunits of complex cells’ RFs were not other neurons then they would presumably have to be distinct processing units within the dendritic trees of the cells. It is important, therefore, to examine the evidence against the hierarchical model. First, many complex cells can be activated with shorter latencies than many simple cells, which led Hoffmann and Stone (1971) to propose that simple and complex cells receive inputs from long-latency and short-latency neurons of the LGN, respectively—that is, from the X and Y channels, in the cat. However, it is now clear that there is not a strict correlation between the X and Y channels in the LGN and simple and complex cells in V1, since many simple cells receive Y input (Henry et al., 1979; Tanaka, 1983; Martin, 1984). Second, there is evidence from cross-correlation analysis of the firing of LGN neurons and complex cells that many complex cells do indeed receive monosynaptic inputs from the LGN (Tanaka, 1983). Monosynaptic LGN inputs to complex cells have also been demonstrated by measuring the latency of

Figure 4.7: A complex receptive field model. The field consists of several overlapping but offset simple-like subfields, that sum their inputs linearly. The subunit outputs are thresholded and then summed.
complex cell activation following electrical stimulation of the subcortical visual pathway (Bullier & Henry, 1979). However, it has not been demonstrated that these connections provide the main part of the complex cells’ excitatory drive under normal physiological conditions. Perhaps the strongest evidence against the hierarchical model is that complex cells might be activated by kinds of visual stimuli that fail to activate simple cells. Hammond and MacKay (1977) reported that some complex cells, but no simple cells, could be strongly activated by two-dimensional random-dot textures. This claim has been challenged by Skottun, Grosos and DeValois (1988) who found that random-dot textures could activate a fraction of both types of cell. This has given rise to a dispute between the two groups focusing on procedural differences between the experiments (Hammond, Pomfrett & Ahmed, 1989; Skottun, Grosos & DeValois, 1991; Hammond, 1991). At present, however, the available evidence certainly does not rule out the basic Hubel-Wiesel model.

One of the most important ideas of the Magpie theory is the distinction between S-units and C-units. As their names suggest, the distinction between these two classes is inspired by the distinction between simple and complex cells, respectively (cf. Fukushima, 1980). Recall that S-units and C-units are assumed to differ only with regard to their plasticity, in particular in the duration of the synaptic trace that records the presynaptic firing rate (p. 66). Consider an S-unit that receives the bulk of its feedforward input directly from the LGN. Because of the complementary properties of ON-centre and OFF-centre cells in the retina and LGN, at any given point in the visual field it should not be possible for both types of cell to be strongly active at the same time. If an ON-centre cell is firing vigorously, an OFF-centre cell with its RF at the same location must be silent, and vice versa. Because S-units acquire patterns that are not greatly extended in time, it follows that an S-unit receiving inputs from the LGN will acquire an RF in which ON- and OFF-subfields are spatially non-overlapping; in other words, it will acquire a simple RF. C-units, on the other hand, have synaptic traces of longer duration. As a contour sweeps across the retinal image it will activate a succession of simple cells, all tuned to the same orientation but at different locations. A C-unit that receives its feedforward input from a number of such simple cells can therefore acquire a complex RF, on the Hubel-Wiesel model.

The previous paragraph has only addressed what might be called the prototypical cases, S-units that are activated by LGN inputs and C-units that are activated by simple cells. But, given that the laminar circuitry in the real area V1 is a good deal more complicated than that of the canonical circuit described in Chapter 3 (see below), we should expect to find other possibilities. These are summarized in Table 4.1. First, consider an S-unit that receives its feedforward input mainly from lower-order S-units with simple RFs. By the same reasoning as that given above for standard simple cells, such an S-unit would itself acquire an RF with segregated ON- and OFF-subfields, because the lower-order S-units could only be strongly active simultaneously if their RFs were ‘in register’, that is with ON-subfields overlapping ON-subfields and OFF-subfields overlapping OFF-subfields. In some other respects, however, the higher-order S-unit would be more like a complex cell; in particular it would have an unusually large RF for a simple cell, and it would probably not be activated by small spots of light. Neurons with properties like this have been described in the cat (e.g., Hubel & Wiesel, 1962, p. 114), although they seem to be rare; Henry (1977) calls them A cells. Second, consider a C-unit that received most of its feedforward input directly from the LGN. An edge of light sweeping across the cell’s RF could activate ON-centre LGN cells and then activate OFF-centre LGN cells at the same location, both within the time window of the synaptic trace mechanism. The C-unit could

<table>
<thead>
<tr>
<th>Neuron type</th>
<th>Main source of input to the neuron</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-unit</td>
<td>LGN cells</td>
</tr>
<tr>
<td></td>
<td>Small RF with segregated ON- and OFF-regions (standard simple RF)</td>
</tr>
<tr>
<td></td>
<td>C-unit</td>
</tr>
<tr>
<td></td>
<td>Small RF with ON- and OFF-regions overlapping substantially (B-type RF?)</td>
</tr>
</tbody>
</table>

Table 4.1: Classes of receptive field expected as a function of the neuron’s type (S or C) and the main source of its feedforward input.
therefore acquire an RF with overlapping ON- and OFF-subfields, although its RF dimensions would be comparable to those of simple cells. It is tempting to identify this kind of RF with the B cells of Henry, Harvey and Lund (1979), although, unfortunately, they failed to find evidence for monosynaptic LGN input to B cells (measuring latency of response after electrical stimulation of the visual pathway). Third, consider neurons that receive their feedforward input from complex cells, such as the neurons in layer 5 of the canonical circuit. These cells would have large complex RFs by the normal criteria (overlapping ON- and OFF-subfields and nonlinear summation) and it would not be easy to tell S-units and C-units apart.

4.3.2 Laminar and Columnar Organization

There are several indications that the intrinsic circuitry of V1, especially in primates, is particularly highly specialized compared with other areas of the cortex, and so the canonical circuit described in Chapter 3 would have to be elaborated quite a lot to account in detail for the processing in this area. The special complexity of V1 is most clear with regard to its laminar structure (reviewed by Lund, 1988). Layer 4, where the bulk of the LGN input arrives, has four distinct sub-layers in the macaque, denoted 4a, 4b, 4cα and 4cβ. The two channels of the subcortical visual pathway terminate in different sublayers of layer 4, the P channel in layers 4α and 4cβ and the M channel in layer 4cα. The two streams appear to remain segregated, to an extent, in different sub-layers of V1, although the fact that neurons have dendrites spanning several layers makes it unlikely that there is no mixing of information from the two streams. Nevertheless, the basic scheme of Figure 3.11 is discernible. The main input arrives in layer 4 and is analysed in layers 4, 3, and 2. The output of these layers are the feedforward connections to higher areas; connections to different areas arise from different levels within layers 2–4. The outputs of these layers are copied into layer 5, where output to subcortical motor centres (mainly the superior colliculus) is generated. This output is in turn copied into layer 6, where feedback connections to the LGN originate. This basic pattern is more clear in the cat, in which V1 appears less specialized than in primates (reviewed by Gilbert, 1983; Martin, 1984). On the other hand, the laminar distribution of RF types is more clear-cut in the macaque than in the cat: simple cells dominate layer 4, where the bulk of LGN fibres terminate, and are common in layer 6, where there is also a fair amount of LGN input; complex cells are common outside layer 4 (Hubel & Wiesel, 1968; Schiller et al., 1976a). In layer 4cβ most of the neurons have centre-surround RFs like those of LGN cells; the failure of simple RFs to develop in this layer might be explained if ON- and OFF-centre inputs are anatomically segregated onto different S-units under genetic control (cf. Miller, 1994). The same laminar distribution of simple and complex cells is present but less pronounced in the cat (Hubel & Wiesel, 1962; Gilbert, 1977; Henry et al., 1979; Martin, 1984; Orban, 1984). The proposed segregation of S-units in layer 4 and C-units in layers 2–3 is undoubtedly highly idealized.

There is an interesting columnar distribution of the property of orientation selectivity in area V1 (Hubel & Wiesel, 1962, 1968, 1974, 1977). An electrode penetrating tangentially through V1 will encounter successive neurons tuned to similar orientations. The preferred orientations of the recorded cells generally varies smoothly, such that a shift of 10° occurs over a distance of about 20–50 µm in the macaque (Hubel & Wiesel, 1974), although discontinuities in the distribution are not uncommon. Techniques have now been developed to map the distribution of this RF property over the cortical surface, confirming the existence of a systematic pattern (Blasdel & Salama, 1986; Blasdel, 1992). The Magpie theory, following several other theories (von der Malsburg, 1973; Kohonen, 1982; Swindale, 1982; Barrow & Bray, 1993), accounts for this distribution in terms of the influence of local interactions between neurons on plasticity: the local circuitry promotes a tendency for nearby neurons to acquire similar characteristic patterns and for neurons separated by a certain distance to acquire different characteristic patterns (p. 67).

The feedback connections from higher-order to lower-order cortical areas and the long lateral connections within layers 2–4 are suggested to be excitatory connections linking neurons that are often simultaneously active and exerting a contextual enhancement effect on the neurons that they contact. There is little experimental evidence on the function of the feedback connections, but it is well-established that the long lateral excitatory connections in layers 2–3 of area V1 interlink neurons with similar RF properties but different RF locations (Ts’o, Gilbert & Wiesel, 1986; Ts’o & Gilbert, 1988; Gilbert & Wiesel, 1989). It is reasonable to assume that a pair of neurons encoding similar
features (the same orientation, for example) would be co-active much more frequently than a pair of neurons encoding different features. The fact that, in area V1, these lateral connections link cells with non-overlapping RFs clearly suggests that they are incapable of strongly activating a neuron by themselves. In cat V1, Gilbert and Wiesel (1990) studied a neuron’s RF properties by presenting a line stimulus in its RF, either alone or surrounded by several lines of a given orientation outside its RF. They found that the neuron’s responsiveness and orientation-selectivity could be altered by the presence of the context. The effects were dependent on the orientation of the lines in the context, and could be either facilitatory or inhibitory, but no consistent pattern was observed in the results. In a similar study of V1 in the alert macaque, Knierim and Van Essen (1992) found that a context of lines generally had a suppressive effect on the studied neurons, but that the degree of suppression was dependent on the orientations in the context. Again, no consistent pattern of results was apparent. Further studies of this kind would be welcome.

4.3.3 The End-Stopping Property

A neuron with an end-stopped RF responds strongly to an oriented contour that just fills the RF, but the response declines if the contour is extended in length beyond the ends of the RF, in one or both directions depending on the cell. A cell that is not end-stopped responds just as well to an extended contour as to one that just covers its RF. The suppression appears to be associated specifically with a contour of the correct orientation beyond the end of the RF; it is not a general inhibitory effect like the surround of an LGN neuron (Orban, Kato & Bishop, 1979). End-stopped complex cells were first found by Hubel and Wiesel (1965a, 1968), who called these cells ‘hypercomplex’ in the belief that they were at a hierarchical stage beyond that of complex cells. However, it is now clear that there are end-stopped simple cells (Dreher, 1972; Schiller et al., 1976a), so end-stopping is better regarded as an optional RF property that can be possessed by both simple and complex cells.

Bolz and Gilbert (1986) discovered that inactivation of layer 6 of cat V1 by injection of GABA greatly reduces or abolishes end-stopping in the majority of cells in layers 2–4 above the injection, without affecting other RF properties. They suggested that end-stopping in layer 4 is mainly produced by the intrinsic connections from layer 6. Layer 6 neurons have RFs much larger than those in layer 4, responding to long contours but not very well to short contours (Gilbert, 1977; Gilbert & Wiesel, 1979), and because of the columnar organization of orientation selectivity the vertical connections between the two layers would connect cells with the same orientation preference. Although the projection is likely to be excitatory, consisting of collaterals of layer 6 pyramidal neurons, it could produce end-stopping in layer 4 if it terminated mainly on inhibitory neurons. In fact, most of the connections from layer 6 to layer 4 do terminate on cells with smooth dendrites (McGuire, Hornung, Gilbert and Wiesel, 1984). In the discussion of layer 6 in the previous chapter I suggested on the basis of these observations that the connections from layer 6 to layer 4 in the canonical circuit provide the main excitatory inputs to a subset of the inhibitory cells in layer 4 (Fig. 3.10). These inhibitory cells should have large RFs similar to those of layer 6, and their inhibitory effect on neighbouring neurons with comparatively small RFs produces the end-stopping property. Higher-order cells above layer 4 can inherit the property from these layer 4 neurons (Bolz & Gilbert, 1986). However, the connections from layer 6 to layer 4 are not included in the proposed canonical circuit merely to account for end-stopping in V1; they are proposed to be a general means by which neurons selective for particular patterns in their RFs can be inhibited in the presence of certain patterns in the surrounds of their RFs. Other examples are suggested later in the chapter (pp. 129, 133).

What is the purpose of end-stopping? Recall that an important goal of early processing is to locate discontinuities in the retinal image—places where some aspect of the image undergoes an abrupt change. In particular, it is useful to represent terminations of contours, as well as ‘corners’ or points of high contour curvature—in short, places where contours are not continuous and smooth. The usefulness of these features for shape perception will become apparent in Chapter 5.

4.3.4 Plasticity of Receptive Field Properties

One of the fundamental assumptions of the Magpie theory is that there is substantial plasticity in the cortex, and that neurons acquire characteristic patterns that correspond to stimulus patterns occurring frequently in the environment (although it is not claimed that visual experience is necessary to
establish adult-like properties in all neurons). Are the features encoded by cells in V1 the kind of patterns that occur frequently in natural images? Surprisingly, there has been little quantitative study of this question (although see Field, 1987), but intuitively it seems quite plausible. Both natural and artificial objects give rise to lots of luminance contours that are straight or only gently curved when examined on a small scale, and at the level of the retinal ganglion cells such contours would be represented by parallel lines of active ON-centre cells and active OFF-centre cells (on the bright and dim sides of the contours, respectively). S-units would be expected to acquire such patterns. A moving contour would activate a succession of simple cells tuned to the same orientation at different locations, enabling C-units to acquire complex RFs, as explained above.

There has been a long-standing controversy over the extent to which the orientation-selectivity of V1 neurons can be modified by experience (reviewed by Frégnac & Imbert, 1984). The Empiricist Principle predicts that neurons adopt RFs that reflect the structure of the environment, so if a kitten experiences only a limited range of orientations during the critical period of plasticity (as can be achieved, for example, using goggles with cylindrical lenses) most neurons should become tuned to those orientations. Inexperienced orientations should be poorly represented—although probably not entirely absent, because many neurons have oriented RFs even in kittens lacking visual experience, and some of these neurons might survive the period of deprivation unaltered. Early studies showed that rearing kittens in such an environment did result in virtually all of the orientation-selective neurons in V1 having preferred orientations matching that to which the animals had been exposed (Blakemore & Cooper, 1970; Hirsch & Spinelli, 1971). The source of the controversy was that it was unclear whether neurons originally tuned to non-experienced orientations had actually been modified, or whether they had simply become unresponsive. It now seems to be generally accepted that actual modification of orientation preference does occur, at least for those neurons that are not very sharply tuned for orientation prior to visual experience (Rauschecker, 1991). One particularly interesting study was performed by Rauschecker and Singer (1981), who raised kittens with various combinations of cylindrical lenses, monocular deprivation (restriction of vision to one eye), and normal vision. They concluded that their results were consistent with a Hebbian scheme of synaptic plasticity, in which connections were strengthened only by a conjunction of presynaptic and postsynaptic activity.

4.4 A Model of Simple and Complex Receptive Fields

In order to demonstrate some of the mechanisms proposed in the previous chapter I present here a computer simulation of a small part of the visual cortex. The model has two main points of emphasis. First, it demonstrates the Magpie theory’s account of synaptic plasticity by showing how neurons with initially unselective responses to visual stimulation can acquire orientation-selective receptive fields, of both the simple and complex types. Second, it illustrates the effects of the local circuitry by showing how localized excitatory and inhibitory interactions between neurons contribute to the generation of orientation selectivity.

This section has three subsections: the description of the model, the results, and comparison with other models. The first subsection is akin to the ‘methods’ section of an experimental paper: it aims to describe the model in sufficient detail for it to be replicated by a programmer familiar with the field. Most readers will probably wish to skim through this subsection.

4.4.1 Description of the Model

The structure of the model is illustrated in Figure 4.8. The model consists of three layers of cells. The first layer simulates a small piece of the lateral geniculate nucleus and is called the LGN patch. The units in this layer have their firing rates set by a procedure described below. These units send feedforward connections to the second layer, which corresponds to a part of layer 4 of area V1 and is called the middle layer. Units in this layer have potentials and firing rates governed by the LGN input and by the local circuitry within the layer. All of these units are S-units. These units in turn send feedforward connections to the third layer of the model, which corresponds to a part of cortical layer 3 and is called the upper layer. Units in the upper layer also have potentials and firing rates, governed by the input from the middle layer and by their own local circuitry. The upper-layer units are C-units. The three layers will be described in turn.
The LGN patch consists of 2500 units arranged in a 50 × 50 square grid. Half of the units have ON-centre RFs and half have OFF-centre RFs, arranged in a chessboard pattern. As explained in the review at the beginning of the chapter, the distribution of sensitivity in the RF of an LGN relay neuron (or retinal ganglion cell) is closely approximated by a difference-of-Gaussians (DOG) function (Fig. 4.2A):

\[
d(x, y) = k_c \exp\left(-\frac{x^2 + y^2}{r_c^2}\right) - k_s \exp\left(-\frac{x^2 + y^2}{r_s^2}\right)
\]  

(4.1)

where \(k_c/k_s\) is the ratio of the peak sensitivities of the centre and surround Gaussians, \(r_c\) and \(r_s\) are the radii of the centre and surround Gaussians (\(r_c < r_s\)), and \((x, y)\) is location relative to the centre of the RF.

The temporal properties of LGN cell responses are not simulated. Instead the firing rate of each LGN unit is taken to be a simple instantaneous function of the luminance distribution in its RF at that moment. If the distribution of luminance in the image is \(L(x, y)\) then the firing rates of the units are calculated from the convolution of the DOG function with the image:

\[
R(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d(x', y') L(x - x', y - y') \, dx' \, dy'
\]  

(4.2)

The firing rate of an ON-centre unit with RF at location \((x, y)\) is taken to be proportional to \(R(x, y)\) if \(R(x, y)\) is positive, or 0 if \(R(x, y)\) is not positive. Similarly the firing rate of an OFF-centre unit at \((x, y)\) is proportional to \(-R(x, y)\) if \(R(x, y)\) is negative, or 0 if \(R(x, y)\) is not negative. The spontaneous firing of LGN neurons is ignored. All the firing rates are scaled by a constant factor such that the greatest absolute value in the convolution yields a firing rate of \(f_{\text{max}}\), where \(f_{\text{max}}\) is a parameter of the model in the range [0, 1]. Normally \(f_{\text{max}} = 1\), but setting \(f_{\text{max}}\) to different values is a convenient if crude way of investigating the effect of varying image contrast on the behaviour of the model. A practical advantage of these simplifications is that the calculation of the convolution, which is quite time consuming, is done only once for each image; the LGN firing rates can then be set very quickly just by looking up the values in the convolution.

Figures 4.9 and 4.10 illustrate the convolutions of two images with the DOG function. The black spot at the bottom of each figure shows the size of the centre of the simulated RF, to scale.
Figure 4.9: One of the images used in the model and its convolution with the model LGN receptive field. Light and dark areas represent positive and negative values of the convolution. The black spot at the bottom of the figure shows the size of the RF centre.
Figure 4.10: Another of the images used in the model.
Light areas in the lower picture represent positive values of the convolution, where ON-centre cells are active, and dark areas represent negative values of the convolution, where OFF-centre cells are active. An intermediate shade of grey shows where the convolution is zero, where neither type of unit responds. Notice how the convolution emphasizes contours in the image.

It is desirable to make the maximum use of empirical data in constraining the parameters of the simulation. Following the model of Wehmeier, Dong, Koch and Van Essen (1989) I use data from the X channel of the cat’s visual pathway, at an eccentricity of about 4.5°. The radii of the Gaussian functions in Equation 4.1 are taken to be \( r_x = 12’ \) and \( r_y = 42’ \). These are typical values for X retinal ganglion cells at the chosen eccentricity (Linsenmeier, Frishman, Jakiela & Enroth-Cugell, 1982). Linsenmeier et al. found a large variation from cell to cell in the ratio \( k_c/k_s \), with a median of 16, but I have used 12. These values imply that the diameter of the RF centre is 39’. The ratio of the total sensitivity of the centre and surround mechanisms (that is, the ratio of the integrals of the two Gaussians) is \( k_c r_x^2 / k_s r_y^2 = 0.98 \). This implies that the response of a unit to a uniform luminance distribution is almost zero, since the centre and surround responses are evenly balanced. In fact the ratio is about 1.37 for retinal ganglion cells (Linsenmeier et al., 1982), but the principal difference between LGN and retinal ganglion RFs is that the former give hardly any response to uniform light (Hubel & Wiesel, 1961).

The LGN patch is taken to subtend a visual angle of 5°, so that the spacing between the RF locations of adjacent units is 6’). Several RF centres therefore cover any given point in the image, although the coverage factor is somewhat less than in the real system (see Wehmeier et al., 1989). The variation of RF size with eccentricity is ignored. As Figures 4.9 and 4.10 illustrate, the convolutions are about 53° wide and 33° high, much larger than the LGN patch. This is to allow the simulation of motion in the visual input. As the simulation proceeds, the LGN patch may glide across the image with a particular direction and speed (Fig. 4.8). If the velocity of the patch is \( v \) and the discrete step size of the simulation is \( \Delta t \), then in each step the LGN patch is advanced a distance of \( v\Delta t \) and the firing rates of all the units are updated accordingly.

We now consider the visual cortex simulation. An important simplification of the model is that excitatory and inhibitory neurons are not individually simulated. Instead the model is constructed from units that can be regarded as representing small clusters of excitatory and inhibitory neurons. This simplification is justified partly on the practical grounds that it reduces the computational burden in the simulation, partly on the empirical grounds that there are no obvious differences between the RF properties of the two classes of neuron in the cat’s visual cortex (Gilbert & Wiesel, 1979; Martin, 1984), and partly on the theoretical grounds that the Magpie theory assumes the two neuron types to have the same functional and connective properties.

The middle layer of the model represents a piece of layer 4B, where the bulk of the X-type LGN fibres terminate (Gilbert, 1983; Martin, 1984), of surface area about 2 \( \text{mm}^2 \). This layer contains 400 units on a two-dimensional square grid, with a spacing between the units of 72 \( \mu \text{m} \). The cat’s area V1 is retinotopically mapped such that a distance of 10 \( \mu \text{m} \) across the surface corresponds to about 1’ of visual angle at 4.5° eccentricity (Albus, 1975), so the modelled layer submends a visual angle of 2.4°. Under a surface area of 2 \( \text{mm}^2 \) there are about 28000 neurons in layer 4B (Beaulieu & Colonnier, 1983), so each unit in the model corresponds to 70 neurons, of which about 15 would be inhibitory cells.

Fibres from the LGN terminate in layer 4 with a localized cluster of collaterals. According to Humphrey, Sur, Uhlrich and Sherman (1985) these arborizations have an area of 0.6–0.9 \( \text{mm}^2 \) when projected onto the cortical surface. In the model the arborizations are much broader than this: each middle-layer unit receives connections from a \( 27 \times 27 \) square of LGN units (Fig. 4.8), which implies an arbour width of 1.62 mm. The justification for this is that the measurements of Humphrey et al. were made with adult cats, and arborizations of LGN fibres are known to be more exuberant in visually inexperienced kittens (LeVay, Stryker & Shatz, 1978). As well-defined RFs develop in the model many of the connection weights fall to zero around the RF periphery (see below), and it is plausible that such disused connections in the cortex physically retract, leaving more restricted arborizations in the adult. Each middle-layer unit receives connections from all 729 LGN units; in other words, the connection probability is 1. In reality each LGN fibre contacts only a fraction of the neurons within the scope of its terminal arborization (Wehmeier et al., 1989). Again, this may be
justified on the grounds that many of the connection weights drop to zero during development; remember also that each unit and connection in the model corresponds to several neurons and fibres in the real system. Each of the feedforward connections to the middle layer has both a hard and a soft weight. The unmodifiable hard weights are set according to a Gaussian distribution, $\exp(-r^2/40^2)$ where $r$ is the angular separation between the presynaptic LGN unit and the postsynaptic middle-layer unit in minutes of arc. This means that middle-layer units tend to be more strongly activated by stimulation at the centres of their RFs than at the periphery. The soft weights, which are adapted during training, are initialized to random values from the uniform range $[0, 1]$. The hard weights are all set to 1.

The upper layer of the model corresponds to a piece of layer 3 with a surface area of about 0.5 mm$^2$. This layer contains 100 units on a two-dimensional square grid, with a spacing between the units of 72 µm as in the middle layer. Each unit receives feedforward connections from all 400 units of the middle layer (Fig. 4.8). This corresponds to an arborization width of the layer 4 to layer 3 fibres of 720 µm. Again, during the development of the model many of the soft weights of these connections fall to near zero. The soft weights are initialized to random values from the uniform range $[0, 1]$. The hard weights are all set to 1.

Within both the middle and upper layers of the model, units are interconnected by the local circuitry. It is assumed in the Magpie theory that each neuron receives short excitatory connections from its immediate neighbours and inhibitory connections from a broader region (Fig. 3.3). Because the units of the model correspond to small clusters of excitatory and inhibitory neurons, each unit sends both excitatory and inhibitory connections to other units in the layer. These connections are not modifiable: they have no soft weights. The hard weights of the short excitatory and inhibitory
Connections are assumed to decline with distance from the unit according to Gaussian distributions. The hard weights of the short excitatory connections are set according to \( \exp(-r^2/100^2) \) and those of the inhibitory connections according to \( \exp(-r^2/400^2) \) where \( r \) is the radial distance from the parent unit in \( \mu\text{m} \). The short excitatory connections extend to a maximum distance from the unit of 216 \( \mu\text{m} \) (or 3 unit spacings) and the inhibitory connections extend to 864 \( \mu\text{m} \) (or 12 unit spacings). These distributions are the same in both layers; those of the inhibitory connections are broadly in line with the axonal distributions of inhibitory neurons. The connection probability of all these connections is 1. A complication is caused by the small size of the model: units near the edge of the grid receive fewer short excitatory and inhibitory connections than units near the centre. To compensate for this edge effect, the hard weights of these connections are ‘normalized’. The weights of the short excitatory connections input to each unit near the edge of the grid are scaled so that their sum is increased to 6.06, the value the sum would have if the unit received a full complement of connections. Similarly, the weights of the inhibitory connections to each unit are scaled so that their sum is 96.69.

The reader who has ploughed through all this tedious detail may be relieved to see a graphical illustration of the model’s connectivity. Figure 4.11 shows the connections received by a single unit in the middle layer, at grid location (15, 7). There is a black square representing each unit from which it receives input. Box C shows the units from which it receives short excitatory connections, box D the units from which it receives inhibitory connections, and boxes E and F show the ON-centre and OFF-centre LGN units from which it receives feedforward connections. The width of each square represents the weight of the connection (in the case of the feedforward connections this is the product of the hard and soft weights). Note that the squares in boxes E and F are displayed at a smaller scale because there are so many LGN units. Figure 4.12 shows the units sending connections to the unit at location (2, 2) in the upper layer. The soft weights of the feedforward connections from the middle layer are initially random.
location (2, 2) in the upper layer. In this case the feedforward connections are from the middle layer (box C).

So much for structure: we now consider function. The assumptions and empirical support for the dynamics of the theory were exhaustively described in Chapter 3; here I will just summarize the equations, parameters, and numerical methods used in the model. The model has seven sets of variables, which change value in time with a range of different time constants. These are (1) the firing rates of the LGN units, which are set from the convolution of the image as explained above; (2) the membrane potentials of the 500 cortical units; (3) the firing rates of the cortical units; (4) the soft weights of the 331,600 feedforward connections; (5) the traces of the cortical firing rates, used in adapting the weights of C-units; and (6) the long-term potentiation thresholds of the cortical units, which slowly vary as a function of (7) the average membrane potentials of the units.

The membrane potential \( V_j \) of the middle layer unit \( u_j \) is governed by the equation:

$$ C \frac{dV_j}{dt} = (E_r - V_j)g_r + (E_e - V_j)g_{em} \sum_{u_k \in F_j} \omega_{kj} w_{kj} f_k + (E_e - V_j)g_s \sum_{u_k \in S_j} \omega_{kj} f_k + (E_i - V_j)g_l \sum_{u_k \in I_j} \omega_{kj} f_k $$

(4.3)

where \( F_j, S_j \) and \( I_j \) are the sets of units from which \( u_j \) receives feedforward, short excitatory, and inhibitory connections, respectively; \( \omega_{kj} \) is the hard weight of the connection from \( u_k \) to \( u_j \); \( w_{kj} \) is the soft weight of the connection from \( u_k \) to \( u_j \); and \( f_k \) is the firing rate of \( u_k \). The equation governing the potential of units in the upper layer is identical except for a single parameter, the constant conductance \( g_{eu} \) associated with the feedforward connections from the middle layer:

$$ C \frac{dV_j}{dt} = (E_r - V_j)g_r + (E_e - V_j)g_{eu} \sum_{u_k \in F_j} \omega_{kj} w_{kj} f_k + (E_e - V_j)g_s \sum_{u_k \in S_j} \omega_{kj} f_k + (E_i - V_j)g_l \sum_{u_k \in I_j} \omega_{kj} f_k $$

(4.4)

The firing rate \( f_j \) of each cortical unit is a simple instantaneous function of the unit’s potential:

$$ f_j = \tanh^+ \left( \frac{V_j - \rho}{\gamma} \right) $$

(4.5)

where \( \tanh^+ x = \max(0, \tanh x) \). The heart of the model is therefore a system of 500 differential equations. The simulation program offers two alternative methods to solve these, which I call the slow and the fast methods. The slow method uses a standard numerical algorithm for solving the equations with high accuracy. The potentials are calculated for a series of discrete time intervals \( t_0, t_0 + \Delta t, \ldots t_0 + n\Delta t \) using an Adams fourth-order predictor-corrector method (Burden & Faires, 1985, Algorithm 5.4). The Milne estimate of the local truncation error is monitored and a warning given if it exceeds a user-specified criterion. If the LGN patch is moving relative to the image then the firing rates of the LGN units are computed afresh in each time step, as explained above. Although accurate and mathematically respectable, this algorithm was found to be impractically slow for the purposes of training the weights of the network. A much more crude but fast algorithm was therefore devised. If the LGN firing rates are constant then the potentials and firing rates of the cortical units will
eventually converge to steady values. By setting $dV_j/dt = 0$ for all $j$, Equations 4.3 and 4.4 define a nonlinear system for which the steady state can be computed quickly using a fixed-point iterative method (Burden & Faires, 1985, Chap. 9). If the LGN patch is moving relative to the image then a series of steady states are computed at time intervals $t_0, t_0 + \Delta t, \ldots t_0 + n\Delta t$. With this fast method the time step $\Delta t$ can be arbitrarily long, whereas with the slow method it must necessarily be quite short (e.g., 0.2 msec) to ensure accuracy. The principal inaccuracy involved in using the fast method is that it effectively makes the firing rate of each neuron an instantaneous function of the input from the LGN patch; when using the slow method there is a certain lag between a change in the LGN firing rates and the cortical response to the change, just as there is in the real system. The fast method does give a crude but tolerable approximation to the results of the slow method, provided the speed of movement of the LGN patch across the image is sufficiently slow for the cortical units to respond well to contours appearing in their RFs.

In Chapter 3 the rules governing the changes in the soft weights and related variables were stated in the form of differential equations (Eqns. 3.8–3.13). Because there are so many weights solving these equations numerically would be astronomically expensive; instead, they are converted to a number of discrete ‘update formulas’ which are applied in each time step of the simulation. The units in the middle layer are S-units, so the soft weights of their feedforward connections from the LGN are adapted by the rule:

$$w_{kj}^{(t)} = w_{kj}^{(t-\Delta t)} + (P(V_j^{(t)}) - D(V_j^{(t)}))f_k^{(t)} \Delta t$$

(4.6)

where $w_{kj}$ is the soft weight of the connection from $u_k$ to $u_j$, $V_j$ is the potential of $u_j$, $f_k$ is the firing rate of $u_k$, and the superscripts indicate whether the value of the variable is from the current or the previous time step. The units in the upper layer are C-units, so the soft weights of their feedforward connections from the middle layer are adapted by the rule:

$$w_{kj}^{(t)} = w_{kj}^{(t-\Delta t)} + (P(V_j^{(t)}) - D(V_j^{(t)}))f_k^{(t)} \Delta t$$

(4.7)

which is identical to Equation 4.6 except that the trace of the firing rate, $f_k^{(t)}$, is used instead of the firing rate itself. All soft weights are restricted to the range [0, 1]; no weight is allowed to move outside this range. The functions $P$ and $D$ give the magnitude of long-term potentiation (LTP) and long-term depression (LTD) and were defined by Equations 3.9 in Chapter 3, reproduced here for completeness:

$$P(V_j) = 2\lambda \tanh^+ \left( \frac{V_j - \theta_j^+}{5 \text{ mV}} \right)$$

(4.8)

$$D(V_j) = \lambda \tanh^+ \left( \frac{V_j - \theta^-}{5 \text{ mV}} \right)$$

Each cortical unit has a variable LTP threshold $\theta_j^+$, which is defined as:

$$\theta_j^+ = \theta_0 + (\bar{V}_j - V_0)\psi$$

(4.9)

where $\bar{V}_j$ is a long-term average of the unit’s membrane potential. This average is updated in each time step by the following rule (derived from Equation 3.11):
Finally, the firing rate traces $f_j$ of each unit are updated by the rule (from Equation 3.12):

$$f_j^{(t)} = \max(f_j^{(t)}, f_j^{(t-\Delta t)} e^{-\Delta t/\alpha})$$

(4.11)

The values of the parameters in Equations 4.3–4.11 are given in Table 4.2. The membrane capacitance, rest potential, leak conductance, reversal potentials and firing rate threshold were set to biologically plausible values at the outset. The other parameters were assigned values from experimentation with the model. Some of the parameters are weakly constrained by available empirical data; the various conductances, for example, have values that are not implausible for single synapses, even though each connection in the model can be taken to represent a large number of synapses. Some parameters are hardly constrained at all by empirical data; for example, it is not even known at present whether the LTP threshold really is variable in cortical neurons. The time constant of the average potentials, $\beta$, determines the rate at which the LTP threshold varies, and therefore influences the volatility of the weights: the longer $\beta$, the more stable are the weights. The value $\beta = 1$ sec was used in the simulation to limit the amount of training time needed, but a more realistic value would probably be much longer. The LTD threshold $\theta^-$ and the ‘target value’ of the LTP threshold $\theta_0$ seem rather low. Unfortunately the large amount of training needed by the model and the unavoidable use of a time-shared computer have made it impossible to systematically investigate the parameter space; improved values could no doubt be found.

For the purposes of training the soft weights, the model is run as a sequence of ‘cycles’. In each cycle the LGN patch drifts across the image with the starting location and direction chosen at random by the program and speed and duration specified by the user. Generally a speed of 2′/msec and cycle duration of 200 msec is used (simulated, not real time!). In most cases the orientation of the LGN patch relative to the image is also set to a random angle at the start of each cycle; this is done to increase the variety of input to the model and to eliminate the effect of any bias in the distribution of
contour orientations in the image. The membrane potentials are set to the rest potential at the start of each cycle, and the evolution of the potentials through the cycle is computed using the fast method. The time step $\Delta t$, the interval at which the weights and associated variables are updated, is 10 msec, so there are 20 updates in a 200 msec cycle. Note that the middle and upper layers are both trained at the same time. With the above parameters a few thousand cycles on a given image are required for the weights to converge to stable values; this takes several hours on a Sun 4 workstation.

The simulation program is written in the C programming language. The graphics reproduced in Figures 4.9–4.26 were generated by the program using the X Window System.

4.4.2 Results of the Model

We first consider the weight patterns that develop in the model from training on various different images. The first image is an artificial one composed entirely of straight, high-contrast contours. The model is trained on the image, as described above, for 5000 cycles. Figure 4.13 (A & B) shows the

Figure 4.13: Weights of the feedforward connections to the units at locations (2, 8), (15, 7) and (0, 8) in the middle layer after training on an artificial image. The left and right boxes show connections from ON-centre and OFF-centre LGN units respectively.
Figure 4.14: Receptive fields in the middle layer after training on the artificial image. Light and dark areas represent ON- and OFF-subfields.
resulting weights of two typical units in the middle layer, displayed using the same conventions as Figure 4.11. The weights of the feedforward connections from ON-centre and OFF-centre LGN units are shown in left-hand and right-hand boxes. Each unit has acquired an RF consisting of elongated and non-overlapping ON- and OFF-subfields, in accordance with the Hubel-Wiesel account of the construction of simple RFs (Fig. 4.6B). The great majority of the middle-layer units develop RFs with two subfields, although three subfields also occur. The subfields are generally about 40′ in width. This matches the diameter of the RF centres of the LGN units, a correlation also found experimentally (Hubel & Wiesel, 1962). A small number of units acquire more complicated RF structures, although these tend to develop into conventional simple RFs with extended training; an example is shown in Figure 4.13C.

An attempt to approximate the RF structure produced by these weights can be made as follows. The contribution of each LGN unit is taken as the DOG function of Equation 4.1, negated for an OFF-centre unit, with peak height scaled by the weight of the connection from that unit. These DOGs are summed for all the LGN units sending connections to the middle-layer unit in question. This procedure yields an ‘RF plot’ for the middle-layer unit in which ON- and OFF-subfields have positive and negative values respectively. (The scare quotes are necessary because the plot would probably not be exactly the same as that found experimentally by mapping the neuron’s RF structure with small spots of light, simple cells not being perfectly linear devices.) In practice the RF plot is very similar in appearance to the weight patterns, because of the clean spatial segregation of the ON- and OFF-subfields in these units. The main difference is that the surrounds of the DOG functions contribute additional weak ON- and OFF-subfields to the sides of the RFs. Figure 4.14 illustrates the RFs of all the units in the middle layer, using the convention that light and dark areas indicate ON- and OFF-regions respectively. Notice that neighbouring units tend to acquire RFs with similar orientations; this results from the influence of the local circuitry (p. 67). Figure 4.14 is a little misleading in that it suggests that the RFs of neighbouring units do not overlap. In fact the RFs of nearby units overlap substantially; they are just displayed at a small size so they can be seen separately.

Figure 4.15 displays the trained weights of the connections of a typical unit in the upper layer. A more intuitive illustration of this unit’s connectivity is given by Figure 4.16. This figure shows the RFs of all the middle-layer units, as in Figure 4.14, but with the brightness of each RF indicating the weight of the connection from that unit to the upper-layer unit. The illustration demonstrates that all the middle-layer units sending strong excitation to the upper-layer unit are tuned to similar orientations. Again, the figure is misleading because the middle-layer RFs are drawn at a small size, suggesting that the upper-layer unit has a relatively huge RF with a non-uniform spatial structure. In fact there is substantial overlap between the various middle-layer RFs; the upper-layer unit’s RF is no more than twice as large in area than those of the middle-layer units. Experimentally, complex cells in layer 3 have RFs only a little greater in size than those of simple cells in layer 4 of the cat’s visual cortex (Gilbert, 1977).

All the above results were obtained from an artificial image of straight contours, with the orientation of the LGN patch being randomized in each cycle so that the model would be exposed to a uniform distribution of contour orientations. If the model is instead trained on contours of a single
Figure 4.16: The same information as Figure 4.15 displayed in a more intuitive form.
Figure 4.17: Receptive fields in the middle layer after training on contours of a single orientation.
Figure 4.18: Receptive fields in the middle layer after training on the image in Figure 4.9.
Figure 4.19: Weights of the feedforward inputs to an upper-layer unit after training on the image in Figure 4.9.
Figure 4.20: Receptive fields in the middle layer after training on the image of Figure 4.10.
orientation, all the cortical units develop RFs tuned to that orientation. Figure 4.17 shows the RFs of the middle layer. This result is also obtained if the fully-trained model of Figure 4.14 is used as the initial state instead of the random-weight model. This complete transformation is reminiscent of the dramatic results of the experiments with kittens by Blakemore and Cooper (1970) and Hirsch and Spinelli (1971) mentioned earlier in the chapter, although it is debatable whether the kitten’s visual cortex is really quite this plastic (Frégnac & Imbert, 1984; Rauschecker, 1991).

Perhaps more interesting are the results of training the model on natural visual images. Figure 4.18 shows the middle-layer RFs that develop with 5000 training cycles on the image of Figure 4.9, again with the LGN patch orientation being randomized to increase the variety of the input. Although generally similar to the results with the artificial image, the RFs have a more rounded appearance. Some of the RFs have conspicuously curved subfields, a phenomenon not reported experimentally although it might be hard to detect. Anomalous RF structures are more common than in Figure 4.14. It is likely that more extended training would eliminate some of these, although many might correspond to genuine features in the image. The upper-layer units also develop weight patterns similar to those trained by the artificial image, although the weights of connections from units tuned to non-optimal orientations tend to be somewhat higher. Figure 4.19 shows a typical example.

The results of the same amount of training on the beech tree image (Fig. 4.10) are dramatically different. As shown in Figure 4.20, the middle-layer units acquire RFs with segregated ON- and OFF-subfields, but the subfields are much less elongated and more rounded than before. Some units even develop LGN-like RFs with a centre-surround organization; Figure 4.21 gives a more detailed view of the weights of one of these. These results apparently reflect the sparsity of extended contours in this image; it consists largely of rather unstructured texture that is poorly resolved by the RFs of the LGN units. It is the ‘blobs’ in the convolution of Figure 4.10 to which the cortical units become sensitive. Although there is some similar texture in Figure 4.9, such as the hedge on the right, it is of lower contrast than the extended contours, and so is less likely to activate the cortical units sufficiently to reach the Hebbian threshold. It is a plausible conjecture that the features of highest contrast in natural images are usually extended contours, and that this is why most cortical neurons become tuned to such contours.

We now consider the orientation selectivity of units in the fully trained model. The model trained on the artificial image (Fig. 4.14) is used for this assessment; very similar results are obtained for the model trained on the first natural image (Fig. 4.18). The orientation tuning of a unit is measured by effectively simulating a neurophysiological experiment: the LGN patch is repeatedly swept across an image of a sine wave grating while the orientation of the patch relative to the grating is systematically varied, and the firing rate of the unit during each sweep is recorded. The speed of movement is 1/40msec and the spatial frequency of the grating is 0.75 cycles/degree. The effect of varying the contrast of the grating is investigated by changing the parameter $f_{max}$, the maximum firing rate of the LGN units. The slow simulation method is used to ensure accuracy. The usual methodology in actual experiments is to record the number of action potentials produced by the
Figure 4.22: Orientation tuning curves of the middle-layer units whose weights are displayed in Figure 4.13 (A & B), assessed with a high-contrast sine grating. The abscissa is the angle of the grating in degrees clockwise from vertical, and the ordinate is the average firing rate.

Figure 4.23: Orientation tuning curves of two units in the upper layer assessed with a high-contrast grating.
neuron during the sweep, and to divide this by the duration of the sweep to estimate the mean firing rate. In the model the firing rate \( f \) is a continuous variable, so formally the mean firing rate in the interval \([t_0, t_1]\) is:

\[
\frac{1}{t_1 - t_0} \int_{t_0}^{t_1} f \, dt
\]

In the discrete simulation this mean is approximated by summing the firing rate in each time step and dividing the total by the number of time steps in the sweep.

Figure 4.22 shows orientation tuning curves for the two middle-layer units whose weight patterns were illustrated in Figure 4.13 (A & B). The orientation selectivities of these units, measured as the width of the tuning curve at half the peak height, are both about 60°. This is a typical value for the model. Real neurons are generally more sharply tuned than this: a curve width of 30–40° is typical for adult cat simple cells (Orban, 1984, p. 153), although the average selectivity of the model units is well within the range of values found experimentally. The discrepancy is probably related to the low density of units in the model compared with the density of neurons in the cortex. The selectivity of the units is strongly influenced by the lateral spread of the short excitatory connections relative to that of the inhibitory connections: the more localized are the short excitatory connections, the more sharply tuned are the units. The spread of the short excitatory connections in the model could not be made much narrower because of the wide spacing of the units. If a more realistic number of units were included in the model then the scope of the short excitatory connections could be more restricted.

Figure 4.23 shows the orientation tuning curves of two units in the upper layer. These units have about the same selectivity as the middle-layer units, even though not all the subfields are tuned to exactly the same orientation (Fig. 4.16). Experimentally, simple and complex cells have a similar degree of orientation tuning, although complex cells are a little more broadly tuned on average (Orban, 1984).

The graphs in Figure 4.24 show the tuning curves of the same two middle-layer units as Figure 4.22, but with all the local circuitry disconnected. In other words the activity of the units is determined solely by their feedforward input from the LGN, as in the most basic version of the Hubel-Wiesel scheme for simple cells. Although the units are still clearly orientation selective, the selectivity is somewhat degraded. More importantly, the purely feedforward model is unable to cope well with different levels of image contrast (Wehmeier, Dong, Koch & Van Essen, 1989). Whereas the tuning curves of Figure 4.22 were obtained with a high contrast grating \( f_{\text{max}} = 1 \), those of Figure 4.24...
Figure 4.25: Orientation tuning curves of the same units as Figure 4.22 with a medium-contrast grating.

Figure 4.26: Orientation tuning curves of the same units as Figure 4.22 with a low-contrast grating. Comparison with Figures 4.22 and 4.25 shows that the local circuitry maintains orientation selectivity as contrast is varied.
were obtained with a rather low level of contrast \( f_{\text{max}} = 0.3 \), because the feedforward model only gives reasonable performance when the LGN firing rates are moderate in value. In the absence of the local circuitry, a higher-contrast grating causes the middle-layer units to give a strong response at all orientations, and because there is an upper limit on the firing rate it follows that the tuning of the cortical units becomes more broad as the LGN firing rates increase. The performance of the middle layer is much better when the local circuitry is intact: Figures 4.25 and 4.26 show the tuning of the units with gratings of medium and low contrast \( f_{\text{max}} = 0.5, 0.3 \). Comparison of Figures 4.22, 4.25, and 4.26 shows that although the mean firing rates increase with contrast, the degree of orientation tuning varies little in comparison with the purely feedforward model. This is also observed in experiments (Sclar & Freeman, 1982). Notice that with the unit at (2, 8) there is a small shift in the optimal orientation as contrast is varied. This is caused by increasing excitatory input from neighbouring units and is probably another symptom of the short excitatory connections having too broad a spread in the model.

### 4.4.3 Comparison With Other Models

Several models of orientation-selective receptive fields have been proposed over the years. I will limit the discussion in this subsection to four models that have been implemented as simulations incorporating comparatively detailed experimental data: the models of Wörgötter and Koch (1991), Clothiaux, Bear and Cooper (1991), Barrow and Bray (1992, 1993), and Miller (1994).

Wörgötter and Koch (1991) used a detailed model of the cat’s X channel, including portions of the retina, LGN, and layer 4 of V1, to investigate the generation of orientation selectivity in simple cells. The model was somewhat more biologically detailed than the present one—for example, it included action potentials and some of the temporal properties of the retinocortical pathway—and included a much larger number of simulated cells in the LGN and cortex (Wehmeier, Dong, Koch & Van Essen, 1989). It was not an adaptive model; all the connections were ‘hard wired’. The focus of interest of the model was the connectivity in layer 4 required to generate sharp orientation tuning in simple cells. Wörgötter and Koch found that a Hubel-Wiesel wiring scheme, in which an orientation preference was imposed on cortical neurons by the spatial pattern of excitatory LGN inputs (Fig. 4.6), with no interactions between the cortical neurons, was sufficient to account for much of the empirical data on orientation tuning. It was not, however, able to account for the sharp tuning observed empirically, unless the RF was made unrealistically long and thin. The model also lacked gain control with varying contrast (Wehmeier et al., 1989). These deficiencies could be rectified by the addition of inhibitory interactions between the cortical neurons. Wörgötter and Koch specifically proposed a hybrid scheme of inhibitory connectivity, combining ‘circular’ inhibition acting at a radius of 500 \( \mu \text{m} \) and more localized inhibition between neighbouring cells. In contrast with the present model, their proposal did not consider excitatory interactions within the cortex; in fact only inhibitory neurons were simulated in the modelled piece of V1. There is evidence from intracellular recordings that excitatory interaction does make an important contribution to neuronal responses in the cat’s visual cortex (Douglas & Martin, 1991a).

The model investigated by Clothiaux, Bear and Cooper (1991) was vastly more simple than that of Wörgötter and Koch. Its focus of interest was the development of orientation selectivity and ocular dominance in the kitten’s visual cortex under various experimental rearing conditions (these experiments are reviewed in the discussion of ocular dominance later in the chapter). They considered a single cortical neuron receiving inputs from the LGN. Interactions between cortical neurons were not considered, except by means of a formal statistical approximation. The output of the simulated neuron in each discrete time step was simply the linear sum of the inputs at that time, weighted by the adaptable connection strengths. The inputs to the neuron were a set of vectors, representing image contours, that were presented in random order interspersed with vectors of random numbers. Various different rearing conditions were simulated by different combinations of noise vectors and pattern vectors. From the perspective of the present model, the main point of interest of the Clothiaux, Bear and Cooper model is its weight modification scheme: it used the BCM rule of Bienenstock, Cooper and Munro (1982), upon which the Magpie theory’s account of plasticity is based (p. 63). The present model can be seen as an attempt to apply a BCM-like mechanism in a much more biologically-
detailed context, by using a more realistic model neuron, explicitly simulated cortical interactions, LGN RFs with a centre-surround organization, and natural visual input.1

Barrow and Bray’s (1993) model is a biologically-detailed simulation of simple cell development that inspired the present model in many respects. It models the spatial RF properties of the retina and LGN, incorporates both excitatory and inhibitory interactions in the modelled cortex, and demonstrates that neurons can acquire simple RFs, organized in a smooth map of orientation selectivity, by training on natural visual input. It differs from the present model in detail, especially in that it has separate populations of excitatory and inhibitory cortical cells. The most important difference between the two models is the form of the weight modification rule: whereas the present model is descended from the BCM theory of Bienenstock et al. (1982), the Barrow and Bray model is descended from the work of von der Malsburg (1973). Both schemes are Hebbian, but they use very different means to regulate the plasticity of individual cells. In the present model the threshold $\theta^+$ varies as a function of the average activity of the neuron, so that low average activity makes it easier for an input pattern to excite the cell enough to reach the threshold, and high average activity makes it more difficult. The approach of von der Malsburg (1973) and Barrow and Bray (1993), in common with a great many other connectionist architectures, is to normalize the weights of the input connections to the neuron, so that the sum of the weights remains fixed for each cell at all times. Bienenstock et al. (1982) describe the difference between the two approaches as a competition between synapses versus a competition between input patterns. Normalization causes a competition between synapses because if some weights are increased then other weights must be decreased to compensate, and vice versa. BCM-like theories, on the other hand, postulate a competition between patterns arriving at different times: if a pattern excites the neuron above the Hebbian threshold, all the weights are increased in proportion to the presynaptic activity, and if the pattern fails to excite the neuron sufficiently, all the weights are decreased in proportion to the presynaptic activities. Unfortunately, there is no experimental evidence at present that directly supports or refutes either weight normalization or a sliding threshold, although it is easy to invent hypothetical mechanisms by which either scheme could be implemented in real neurons. Normalization, for example, could be achieved by conservation of the total number of membrane channels distributed among the various synapses, although this proposal might be hard to reconcile with the fact that learning can have a large effect on the number of synapses per cortical neuron in rats (Greenough, Withers & Anderson, 1992).

Barrow and Bray (1992, 1993) also extended their weight modification rule with a synaptic trace mechanism to demonstrate the development of complex RFs in a higher layer of the model, although the layer contained only 8 or 16 units organized as a ‘winner takes all’ network (only the most active unit had its weights adapted for each stimulus). Apart from the different modification rule the present model builds upon Barrow and Bray’s by simulating an upper layer with a larger number of units and with the same local circuitry as the middle layer.

Whereas Barrow and Bray’s model and the present model were intended to demonstrate how neurons with initially random weights could acquire simple RFs and become organized into a smooth orientation map from exposure to natural visual input, Miller (1994) presented a model showing how these results could occur without visual stimulation. This is an important issue because the experimental evidence indicates that kittens already have many orientation-selective cells in the visual cortex at the time of eye opening, although normal visual experience appears to be necessary for full adult-like selectivity to develop in all neurons (Frégnac & Imbert, 1984). Miller’s was a comparatively simplified model, with units like those of Clothiaux et al. (1991) and a weight modification rule that used normalization to model competition between the input synapses of each unit. The model demonstrated how simple RFs could be acquired from correlated activity generated prenatally in the retina or LGN, under certain assumptions. Specifically, Miller assumed that ON-centre cells have activity most strongly correlated with that of other ON-centre cells, and OFF-centre cells with other OFF-centre cells, at each retinal location, but that ON-centre and OFF-centre cells are more strongly correlated with each other at retinal locations separated by a short distance. Under this assumption the cortical neurons of the model acquired RFs with spatially segregated ON- and OFF-subfields. Miller’s assumption that simple RFs are acquired in the absence of visual experience and

1 Since the completion of this project a study of the original BCM model using natural visual input has been published (Law & Cooper, 1994).
the viewpoint of the present theory that they can be acquired from natural visual images are not incompatible. Any theory of plasticity that allows normal RF properties to be acquired from visual experience can also account for the prenatal acquisition of such properties, provided the input to the visual cortex generated by the retina or LGN before birth is similar to normal visual input in the appropriate respects. This does not mean, of course, that the subcortical centres need to generate realistic visual images! Because V1 neurons have very small RFs it is only required that the subcortical centres generate activity that is structured like normal visual input at a very local scale. Specifically, one would expect local clusters of ON-centre and OFF-centre cells to generate activity similar to that produced visually by short segments of contour.

4.4.4 Summary

This section has described a detailed simulation of the early stages of the cat’s visual pathway. The model demonstrates that the laws governing plasticity in the Magpie theory can account for the acquisition of simple and complex receptive fields from exposure to natural visual input. The model also demonstrates the contributions of feedforward excitation and of the local circuitry in producing the property of orientation selectivity.

In considering all of these neurophysiological details it is important not to lose sight of the big picture: simple and complex cells are proposed to be signalling the presence of short contour segments in the luminance distribution of the retinal image, a necessary prelude to the more interesting business of shape perception and object recognition. But there are several other types of information in visual images that must also be made explicit in early processing. The remainder of this chapter gives a much less detailed account of some of these other attributes.

4.5 Motion

The image falling on the retina is never static under natural conditions. Motion of the eyes, of the body, and of objects in the environment all cause continuous temporal change in the spatial structure of the image, and we are able to perceive and use this change for a variety of purposes (see Nakayama, 1985, for general review). Even when fixating an unmoving object, the retinal image is still not static because of tiny eye movements (Ditchburn, 1973). In fact, if an image is artificially stabilized on the retina it rapidly fades and disappears from consciousness, presumably because of neural adaptation. Thus, motion is one of the basic parameters of the visual image.

Hubel and Wiesel (1959, 1962, 1968) discovered that most V1 neurons, especially those with complex RFs, respond more vigorously to a moving stimulus than to a static one. A fraction of cells, however, give a markedly stronger response to movement in one direction, called the cell’s preferred direction, than the other, nonpreferred direction. Some of these neurons give no response at all in the nonpreferred direction. (These measurements are generally performed with stimuli that are optimal for the cell in other respects—orientation and so on—and moving in the two directions orthogonal to the axis of orientation.) Neurons are described as direction selective if the normalized difference in response between the preferred and nonpreferred direction exceeds some criterion (Orban, 1984). Direction selectivity is not present in the LGN, and so, like orientation selectivity, it is generated in the cortex.

A number of studies have indicated that direction selectivity in simple cells involves an inhibitory influence operating in the non-preferred direction (e.g., Goodwin, Henry & Bishop, 1975; Ganz & Felder, 1984). For example, Ganz and Felder (1984) flashed pairs of thin light bars within the RFs of direction-selective simple cells in cat V1, testing a range of spatial separations and temporal delays between the first and second bars. They found that the response to one bar is suppressed when it is preceded by the other bar, relative to the response to one bar alone. However, this inhibitory effect only occurs in one direction across the RF, namely the cell’s non-preferred direction. In the preferred direction, the response to the second bar is not reduced by the first bar—in fact there is a weak facilitation in some cells. The inhibitory effect occurs for very small displacements between the two bars, and occurs within single ON-subfields and OFF-subfields of the RF. The fact that individual subfields of the RF can be direction selective rules out interactions between subfields as an explanation of the effect. Intracellular recordings confirm that direction-selective cells receive greater
inhibition when stimulated in the non-preferred than the preferred direction (Creutzfeldt, Kuhnt & Benevento, 1974), although this inhibition is not strong (Douglas, Martin & Whitteridge, 1991).

The present theory accounts for direction selectivity by the assumption that the feedforward input to simple cells from the LGN can impose a strong directional bias on some cells, and that the local circuitry enhances this selectivity. A moving stimulus will activate simple cells with matching RF properties, and cells in which the directional bias matches the stimulus motion will, ceteris paribus, be activated more strongly than those with the opposite bias. Those cells with the matching bias will tend to reinforce each other’s activity via the short excitatory connections, and inhibit cells with the non-matching bias. Complex cells inherit direction selectivity from these simple cells, although they have their own local circuitry that has the same effects. There is evidence that the excitatory input to at least some complex cells is itself direction selective, consistent with their inheriting the property from simple cells (Goodwin & Henry, 1975).

This explanation assumes that a direction-selective cell should receive more inhibition from cells with the opposite preferred direction than from cells with the same preferred direction, all else being equal. This is to be expected from the Hebbian nature of synaptic plasticity. Because it is assumed that directional bias is produced in simple cells by their excitatory inputs from the LGN, this bias is subject to modification by experience. (Recall that the connections making up the local circuitry are assumed to be non-modifiable.) Neurons with relatively strong inhibitory interactions could not be strongly active simultaneously, and because plasticity depends on membrane depolarization above a threshold, such pairs of cells would tend to acquire significantly different characteristic patterns, such as opposite direction preferences. If the interaction between the cells was excitatory, they would be likely to adopt similar characteristic patterns, including identical direction preferences. There is evidence that direction selectivity in kitten V1 develops in accordance with the Empiricist Principle. Cynader, Berman and Hein (1975) raised kittens in an environment in which contours moved in only one direction (although eye and head movements were not controlled, so it is unlikely that only one direction occurred in the retinal image). They found that about three quarters of direction-selective neurons in V1 preferred the experienced direction.

There are several conceivable ways in which the convergence of LGN input could impose a directional bias on a simple cell, including: direction-dependent temporal interactions between ON- and OFF-subfields in the cell’s RF (Hubel & Wiesel, 1959; Wörgötter & Holt, 1991); convergence of LGN inputs with different temporal properties, such as neurons in the X and Y channels in cat, onto the cell (Marr & Ullman, 1981); and different latencies of the LGN inputs to the cell (McLean & Palmer, 1989; Reid, Soodak & Shapley, 1991; Tolhurst & Dean, 1991; DeAngelis, Ohzawa & Freeman, 1993; Jagadeesh, Wheat & Ferster, 1993). The latter possibility is particularly attractive in view of the discovery of a subset of neurons in the cat LGN that have responses lagged by several tens or even hundreds of milliseconds behind those of other neurons (Mastronarde, 1987; Saul & Humphrey, 1990, 1992).

Some neurons are also tuned to a particular stimulus velocity. In the monkey, such cells are quite rare in V1, but common in area MT (Maunsell & Van Essen, 1983). Velocity selectivity has not been studied as thoroughly as direction selectivity. Each of the methods mentioned above for imposing a directional bias on a cell could also impose a preference for a particular velocity (Wörgötter & Holt, 1991; Harris, 1986; McLean & Palmer, 1989), and this bias could be refined to full velocity selectivity by the local circuitry.

In primates, motion selectivity is more prevalent in those parts of the cortex receiving input predominantly from the M channel (e.g., layers 4ca and 4b of V1, and area MT) than from the P channel (reviewed by Livingstone & Hubel, 1987). This suggests that some property of the M cells in the LGN, perhaps their relatively transient responses and greater sensitivity to moving stimuli, is particularly instrumental in establishing direction and velocity selectivity in the cortex. It would be interesting to know whether lagged cells are present in the primate LGN as they are in the cat and, if so, whether they are concentrated in the magnocellular layers.
4.6 Spatial Frequency

Natural visual images are generally structured at several different spatial scales. For example, consider the tree in Figure 4.10. At a small scale one can distinguish the individual twigs and leaves, at a larger scale the major branches are apparent, and the overall shape of the tree is on a still larger scale. This observation, together with the Empiricist Principle, suggests that neurons are tuned to features over a range of different sizes. It is desirable to have neurons tuned to very fine features so that one can discriminate small details like the tree’s leaves. It might be difficult, however, to recognize the gross structure of objects using only the responses of these small-feature neurons. The outline of a tree can be hard to define when examined closely, but it is clear enough when the details of the individual twigs and leaves are ignored (unless the tree is in a wood!). Thus, neurons tuned to relatively coarse luminance variations, insensitive to fine details, are also useful. The intuitive concept of spatial scale can be quantified as spatial frequency. For an introduction to spatial frequency and its application to vision see Shapley and Lennie (1985).

Just as most neurons in the low-level visual cortical areas are selective for orientation, they are also selective for spatial frequency. When stimulated with a sine wave grating of optimal orientation and fixed contrast, the neuron responds only if the spatial frequency of the grating lies within a certain range, with a peak response at one particular frequency (Maffei & Fiorentini, 1973; Schiller, Finlay & Volman, 1976c; Movshon, Thompson & Tolhurst, 1978c). Thus, the grating must be of optimal orientation and optimal spatial frequency to maximally activate the cell. DeValois, Albrecht and Thorell (1982) studied the contrast sensitivity function (CSF) of neurons in foveal and parafoveal area V1 of the macaque monkey. This function is obtained by stimulating the cell with sine wave gratings over a range of spatial frequencies, measuring for each frequency the minimum contrast necessary to obtain a criterion level of response from the cell. The reciprocal of this minimum contrast is the contrast sensitivity of the neuron for that particular spatial frequency, and plotting this value for a range of frequencies yields the cell’s CSF. In practice, the CSF is very similar to the neuron’s spatial frequency tuning curve—the neuron’s response to sine wave gratings of fixed contrast over a range of spatial frequencies (Movshon et al., 1978c). The curve always has a single peak, indicating the frequency for which the contrast sensitivity is maximal. A useful measure of the sharpness of a neuron’s spatial frequency tuning, or bandwidth, is the width of the CSF at half-height—that is, the ratio between the two frequencies, one on either side of the peak, for which the contrast sensitivity is half that of the peak value, expressed in octaves. DeValois et al. found that most of the neurons in their sample had bandwidths between 1.0 and 1.5 octaves, although many were very sharply tuned with bandwidths below 1 octave, and others were very broadly tuned with bandwidths over 2 octaves. The optimal spatial frequencies of different neurons were distributed through a range of over 4 octaves, although most neurons had peak frequencies between 1 and 8 cycles/degree. Thus, each point in the visual field is analysed at a range of different spatial frequencies.

The existence of neurons tuned to different spatial frequencies has led some people to argue that the visual system is performing a spatial Fourier analysis of the image (e.g., Maffei & Fiorentini, 1973). However, the properties of cortical neurons deviate from those that one would expect were the cortex performing a true Fourier transform (see Shapley & Lennie, 1985). Furthermore, it is not clear that a Fourier transform of the image would serve any useful purpose. Consider object recognition: why would it be easier to recognize objects from a Fourier transform of the image than from the image itself? In common with Marr (1982), I would argue instead that the purpose of early vision is to extract features of the image—features that are likely to correspond to actual features of the viewed scene, and are therefore useful as a basis for higher-level processes like object recognition. Spatial frequency selectivity in visual neurons reflects the fact that features (specifically, luminance contours) are extracted at a variety of spatial scales.

If this view is correct, then the spatial frequency tuning of a simple cell should be a consequence of its RF organization. Specifically, the peak frequency for a sine wave grating should be the frequency that maximally covers the ON- and OFF-subfields of the RF with peaks and troughs of the grating, respectively. The bandwidth should be related to the number of subfields in the RF: an RF with only one subfield should be broadly tuned for spatial frequency, whereas an RF with three subfields of alternating sign should be more sharply tuned. The results of studies relating spatial
frequency selectivity with RF organization in cat simple cells are in fairly good agreement with these predictions (Movshon et al., 1978a; Jones & Palmer, 1987; DeAngelis, Ohzawa & Freeman, 1993). As always, selectivity should be sharpened by the local circuitry; it has been shown that spatial frequency tuning becomes broader under disinhibition by application of bicuculline (Vidyasagar & Mueller, 1994). The fact that the sharpness of spatial frequency tuning is fairly invariant with stimulus contrast (Skottun, Bradley & Ramoa, 1986) is suggestive of the gain-control properties of the local circuitry.

4.7 Ocular Dominance

The ocular dominance (OD) of a neuron is a measure of the extent to which it is activated by one eye or the other. OD does not itself have great perceptual significance—when both eyes are open we are not aware of which eye is delivering what information. However, the property is worth a brief discussion because its development in area V1 of immature animals, usually kittens, is by far the most thoroughly investigated aspect of plasticity in the visual cortex.

Neurons in V1 dominated by the same eye tend to be grouped together into columns (Hubel & Wiesel, 1962, 1968, 1977). Thus, a recording electrode advanced tangentially through the cortex encounters alternating sequences of cells dominated by one eye or the other. This organization is much more well defined in the macaque monkey than the cat. Figure 4.27 is a map of the OD columns in one side of V1 of a macaque (LeVay, Hubel & Wiesel, 1975). Note that the columns are more accurately described as slabs, about 200–500 µm wide. In layer 4c of macaque V1, where most of the LGN fibres terminate, most neurons are strongly dominated by one eye or the other, and so the OD columns are very well defined. Above and below layer 4c, and in all layers of cat V1, the columns are not discrete groups. Rather, there is a roughly continuous distribution of OD, such that neurons in the centre of each column are completely dominated by one eye and neurons at the borders between columns are driven about equally by both eyes. The anatomical substrate of OD in layer 4 is thought to be differential convergence of input fibres from the LGN, which are themselves almost purely monocular. That is, a cortical neuron dominated by one eye receives greater excitatory drive from LGN fibres carrying information from that eye than from fibres carrying information from the other eye. This is supported by the finding that LGN fibres corresponding to the left and right eyes

Figure 4.27: Ocular dominance columns in layer 4c of the part of area V1 on the surface of the right cerebral hemisphere of a macaque monkey. Black slabs contain neurons dominated by one eye, white slabs are dominated by the other eye. The representation of the fovea is on the right. (From D. H. Hubel and D. C. Freeman, 1977, Projection into the visual field of ocular dominance columns in macaque monkey, Brain Research 122: 336–343, Fig. 1A. Copyright 1977 Elsevier/North Holland Biomedical Press, Amsterdam. Reproduced with permission of Elsevier Ltd.)
terminate in alternating patches in layer 4c of macaque V1, in register with the physiologically-measured OD distribution (Hubel & Wiesel, 1977).

Various manipulations of the kitten’s visual input cause large modifications of the OD distribution in V1, provided the manipulation is carried out during a certain critical period of development, lasting from about 2 weeks to about 4 months of age with a peak sensitivity at about 5 weeks (reviewed by Frégnac & Imbert, 1984; Rauschecker, 1991). At the time of birth, the visual cortex is rather immature in kittens: although many neurons have RF properties qualitatively similar to those of adult cats, they tend to be poorly responsive, and many other neurons are either unresponsive to visual stimulation or have rather unselective responses (Hubel & Wiesel, 1963; Blakemore & Van Sluyters, 1975). If a kitten is deprived of visual input to both eyes from birth (binocular deprivation), by rearing in the dark or by eyelid suture, the cortex fails to mature to the adult state (Wiesel & Hubel, 1965; Blakemore & Van Sluyters, 1975). If a kitten is raised with one eye occluded so that it receives normal visual input to one eye only (monocular deprivation) the effects are quite different. Normal, adult-like RF properties develop in V1, but the great majority of neurons become monocular, driven only by the unoccluded eye (Wiesel & Hubel, 1963, 1965; Blakemore & Van Sluyters, 1974). The OD columns of cells dominated by the unoccluded eye expand in width, whereas the columns dominated by the other eye shrink. This OD shift is equally severe if one eye receives only unpatterned light (Wiesel & Hubel, 1963); recall that uncontoured light fails to activate retinal ganglion cells strongly. Finally, if a kitten’s eyes are misaligned by sectioning one of the eye muscles (artificial strabismus), most neurons in V1 become monocular, with about half of them dominated by each eye (Hubel & Wiesel, 1965b).

Qualitatively, the results of these different rearing conditions can be explained on the assumption that neurons become tuned to particular patterns occurring in the visual input, in accordance with the Empiricist Principle. In normal rearing, there is a correlation between the activity in the two eyes: if a particular pattern occurs at one location in the visual field seen by one eye, it is likely that a similar pattern will be seen in a similar location in the other eye, and so most neurons develop binocular RFs. In a kitten reared with artificial strabismus, this correlation does not occur (Hubel & Wiesel, 1965b), and so most neurons develop monocular RFs. In a monocularly deprived kitten, only weak, unstructured activity will arrive at the cortex from the occluded eye, and so only the strong, patterned input from the open eye will depolarize cortical neurons above the Hebbian threshold. Finally, in binocular deprivation, activity from both eyes will be weak and unstructured and so incapable of driving cortical cells; neurons therefore do not develop mature RFs. Quantitatively, the BCM theory of weight modification provides a good account of the effects of these rearing conditions on the OD of individual neurons (Bienenstock, Cooper and Munro, 1982; Bear, Cooper & Ebner, 1987; Bear & Cooper, 1990; Clothiaux, Bear & Cooper, 1991). The present theory explains OD plasticity in much the same way as the BCM theory on a qualitative level.

Another rearing condition with interesting results is reverse occlusion. The kitten is monocularly deprived for a time sufficient to cause a substantial shift in OD distribution in favour of the open eye. The deprived eye is then opened and the other eye is deprived for a time. Provided the reversal is done while still in the critical period, the result is another strong shift in OD distribution, this time in favour of the newly-opened eye (Blakemore & Van Sluyters, 1974). This result presents a problem for many Hebbian schemes of synaptic plasticity. At the time of the reversal, most neurons are monocular, and so presumably receive only weak connections from the initially-deprived eye. How, then, can the newly-opened eye activate neurons sufficiently to exceed the threshold required for increasing synaptic strength? Bienenstock, Cooper and Munro (1982) proposed that the modification threshold is variable, and that a fairly prolonged period of low activity (as must presumably follow the reversal) causes the threshold to drop, so that even the weak connections from the newly-opened eye can eventually drive the neuron’s membrane potential high enough to exceed it. The sliding long-term potentiation (LTP) threshold of the Magpie theory gives the same explanation.

Pharmacological manipulations (reviewed by Rauschecker, 1991) support the contention that the plasticity mechanism involved in OD development is the same as that underlying LTP and long-term depression. Infusion of 2-amino-5-phosphonovalerate (APV) into the visual cortex of a kitten while it is monocularly deprived prevents the normal OD shift (Kleinschmidt, Bear & Singer, 1987; Bear, Kleinschmidt, Gu & Singer, 1990). APV is an antagonist of NMDA receptors, which are involved in the induction of LTP, as explained in Chapter 3. This result does not conclusively
demonstrate that NMDA receptors are directly involved in OD plasticity, however, because APV also reduces neural responsiveness (Rauschecker, 1991). Reiter and Stryker (1988), on the other hand, found that infusion of muscimol into a small part of the visual cortex of kittens during monocular deprivation resulted in an OD shift in favour of the deprived eye in the affected region—the opposite of the normal monocular deprivation effect. Muscimol is an agonist of GABA_A receptors, and presumably has the effect of holding down the membrane potentials of affected neurons without blocking synaptic activity. Reiter and Stryker argued that synaptic plasticity is dependent on the postsynaptic membrane potential, the sign of change being positive if the cell is depolarized above a threshold and negative if its potential is below that threshold. This is consistent with the weight modification rules of the BCM theory and the Magpie theory.

It is possible that the level of plasticity in the visual cortex, which rises to a peak and then declines during the critical period, is regulated by the acetylcholine (ACh) and norepinephrine (NE) fibre systems. Kasamatsu and Pettigrew (1979) found that infusion of 6-hydroxydopamine (6-OHDA) into kitten visual cortex during monocular deprivation prevents the OD shift. Fibres that contain NE are killed by 6-OHDA, and Kasamatsu and Pettigrew interpreted their results as showing that the level of NE in the cortex regulates OD plasticity. However, subsequent studies placed doubt on this conclusion, showing that other methods of depleting cortical NE do not disrupt plasticity so severely (e.g., Daw, Robertson, Rader, Videen & Coscia, 1984). Bear and Singer (1986) appear to have resolved the controversy by showing that 6-OHDA interferes with the effects of both NE and ACh in the visual cortex. Combined destruction of both fibre systems greatly reduces OD plasticity, but destruction of either system alone has little effect. In the present theory the ACh system is proposed to be the anatomical substrate of the focusing mechanism, which is consistent with the widely-held view that selective attention is one of the factors influencing plasticity.

## 4.8 Binocular Disparity

Because our eyes look at the world from two different locations, the image formed on the retina of one eye is different from that in the other eye, and the brain is capable of using the differences between the two images to derive information on the three-dimensional structure of the scene, a process called **stereopsis** (reviewed by Poggio & Poggio, 1984). The difference between the location of a viewed point on one retina and the corresponding location of the point on the other retina is called the **disparity** of the point. The disparity of a fixated point is zero, and points nearer and further than the point of fixation have **crossed** and **uncrossed** disparities, respectively. This section considers the early processing of disparity for segments of contour. Disparity processing of textured regions is considered in the following section, and the problem of ‘global stereopsis’ (the integration of these local disparity measurements) is postponed to the next chapter.

Early studies in the cat revealed that many cortical neurons respond selectively to certain ranges of disparities (Barlow, Blakemore & Pettigrew, 1967; Nikara, Bishop & Pettigrew, 1968). More recent studies have employed awake monkeys that are actively fixating a stimulus (Poggio & Fischer, 1977; Poggio & Talbot, 1981; Poggio, 1984). Poggio and Fischer (1977) tested neurons in foveal areas V1 and V2 of the macaque with moving light or dark bars of optimal orientation and varying horizontal disparity. They classified the cells that were disparity selective into four groups. **Tuned-excitatory** cells give a large response only if the stimulus is within a narrow range of disparities, with a peak response at or close to zero disparity. Thus, these neurons are selective for a limited range of depths, close to the depth of the fixation point. The much more rare **tuned-inhibitory** cells have responses opposite to tuned-excitatory cells, responding well to stimuli with large crossed or uncrossed disparities but poorly to stimuli with disparities close to zero. **Near cells** respond well to stimuli over a broad range of crossed disparities, but poorly to stimuli with uncrossed disparities and especially poorly to disparities near zero. **Far cells** are the inverse of near cells, responding best over a broad range of uncrossed disparities. Near and far cells, therefore, are selective for depths nearer and further than the fixation point, respectively. Similar classes have been reported in V1 and V2 of the cat (Ferster, 1981).

It seems likely that the tuned-excitatory cells are the most important neurons for measuring disparities. Disparities are typically small: a disparity of 1’ corresponds to a 1 cm depth difference about 150 cm in front of the eyes. Different tuned-excitatory cells in the macaque have preferred
disparities (corresponding to the peaks of the tuning curves) distributed through a range of at least ±6′ (Poggio, 1984). The smallest change in disparity from zero to non-zero that can be reliably discriminated by macaques (and humans) is less than 10″ (Sarmiento, 1975). This disparity is much smaller than the widths of the tuning curves of tuned-excitatory neurons in V1 and V2, which are rarely more narrow than 5′ (curve width measured at peak response /√2) (Poggio, 1984). This presumably indicates that stereo acuity makes use of population coding. A change in disparity of 10″ cannot reliably cause a change in the activity of any single neuron, because such a change would be swamped by the variability of neuron firing rates, but it could cause a reliable, distinguishable change in the activity of a population of neurons. Lehky and Sejnowski (1990) have presented a model of stereo acuity based on this idea.

The Magpie theory assumes that the tuning properties of visual neurons are established mainly by the feedforward excitatory inputs to the cells, although they are subject to modification (especially sharpening of tuning) by the local circuitry. In the case of disparity selectivity, the optimal disparity for a simple cell should be given by the relative positions of its monocular RFs in the two eyes and by the relative strengths of the connections from the two eyes (Barlow et al., 1967). Thus, a tuned-excitatory simple cell that is maximally responsive to a certain disparity near zero should be well driven from both eyes and have ON-subfields matching ON-subfields, and OFF-subfields matching OFF-subfields, at positions in the two retinas displaced by that disparity. Detailed comparisons of the monocular RFs of tuned-excitatory neurons in the cat are consistent with this explanation (Ferster, 1981).

The Empiricist Principle predicts that strongly binocular neurons should acquire characteristic patterns tuned for disparities close to zero. When an animal fixates an object, each feature in the left retinal image, the projection of some point on the object, is very likely to be correlated with a similar feature in the right retinal image, the projection of the same point. However, the probability of matching features being present at any given pair of locations on the two retinas is a function of the disparity between the locations. The probability is greatest at zero disparity (corresponding to the distance of the fixation point) and falls off rapidly as disparity increases, dropping to the level of chance at large disparities. This probability distribution should be reflected in the distribution of preferred disparities among binocular neurons. Within the tuned-excitatory group there is in fact a clear clustering of peak values around zero disparity (Poggio, 1984).

The other types of disparity-selective neuron (near, far and tuned-inhibitory) are not as easy to explain as the tuned-excitatory cells. First, it is not clear why there should be separate categories of disparity-selective neurons. In fact, it is quite possible that there are not discrete classes at all, but that all disparity-selective cells are points on a continuum. LeVay and Voigt (1988) performed a detailed quantitative study of disparity selectivity in V1 and V2 of the cat, and argued against the existence of separate categories of cells, stressing the variability of tuning curve shapes and the many cells with intermediate properties. They also failed to find any clear examples of the tuned-inhibitory type.

Second, disparity-selective neurons other than the tuned-excitatory type appear to violate the Empiricist Principle. As explained above, the principle leads us to expect that neurons should acquire preferred disparities close to zero. However, this is only the case for strongly binocular neurons. It is consistently found that tuned-excitatory neurons are strongly binocular, but disparity-selective neurons not in the tuned-excitatory group generally have unbalanced ocular dominance (OD): they are strongly excited by visual stimuli in one eye and are driven weakly or not at all by stimulation of the other eye (Poggio & Fischer, 1977; Ferster, 1981; LeVay & Voigt, 1988). Moreover there appears to be an inhibitory influence from the non-dominant eye that makes a large contribution to the disparity selectivity of near and far cells (Ferster, 1981). This inhibition could only come from other cortical neurons. Actually, it might be possible to explain this inhibition by the following argument. There is a strong correlation between tuned-excitatory selectivity for disparity and balanced OD, as noted above. There is a systematic columnar distribution of OD, as described in the previous section. Therefore, disparity-selective neurons should not be uniformly distributed across the cortex: tuned-excitatory cells should be clustered near the borders of the OD columns, whereas near and far cells should be more common in the centres of the columns. Evidence for this was found by LeVay and Voigt (1988). But the distance from the borders to the centres of the OD columns (100–300 µm) is about the distance over which the inhibitory effects of the local circuitry are operating. It should therefore not be surprising to find that neurons which are poorly selective for disparity suffer inhibition from the
tuned-excitatory cells when the disparity of the visual stimulus is close to zero. This, I suggest, is the origin of the near, far and tuned-inhibitory types of receptive field.

Disparity of locations in the two retinas is not the only kind of disparity caused by stimuli in depth. If an edge or mark on an object is slanted towards the viewer its projections onto the two retinas have slightly different orientations, and there is a relation between the degree of slant and the disparity of the two retinal orientations. This means that a binocular neuron with different preferred orientations in the two eyes could encode orientation-in-depth. Blakemore, Fiorentini and Maffei (1972) and Nelson, Kato and Bishop (1977) described cells in cat V1 with different preferred orientations in the two eyes, different cells having orientation disparities distributed over a range of ±15°. Nelson et al. argued that the cat could not make much use of this mechanism on the grounds that the cells are not sharply tuned for orientation-in-depth; however, we have seen that population coding allows precise measurements to be made with broad tuning curves. Another kind of disparity is disparity of motion: neurons with different preferred directions or velocities in the two eyes could encode motion-in-depth. Poggio and Talbot (1981) have described such cells in the macaque.

4.9 Texture

Most surfaces in natural environments are textured, which results in the luminance distribution across an image of the surface having a ‘granular’ quality. Familiar examples include a field of grass, foliage seen from a distance, a shingled beach, and the fur of an animal. The importance of texture in visual perception was stressed by Gibson (1950, 1966). However, the great majority of neurophysiological studies of the visual system have used simple luminance contours as stimuli; texture has been somewhat neglected, despite its ubiquity in natural scenes. In this section, therefore, I will mostly rely on psychophysical research, and try to connect this with what is known of the neurophysiology.

It should be noted that two adjacent surfaces differing in texture may be distinguished because they differ in other qualities besides texture, such as colour or luminance. In particular, if the surfaces have different mean luminance, the border between them can be detected by simple cells with sufficiently large RFs, or more precisely by simple cells tuned to spatial frequencies lower than that of the ‘grain size’ of the textures. Such cells fail to discriminate the details of the texture, but can respond to the difference in mean luminance between the two regions (Nothdurft & Li, 1984, 1985). Thus, the visual system may discriminate two textured regions without needing to form any explicit representation of the textures themselves. On the other hand, many textures can be distinguished even if their mean luminance is equated, indicating that representations of textures are explicitly computed. It is not the case that arbitrary pairs of textures can be discriminated, however. Two adjacent textured regions, equated for mean luminance, may or may not be perceived as different, depending on the characteristics of the textures. The bulk of the research on texture perception has focused on determining what these crucial characteristics are.

There are two fundamentally different ways in which textures can be discriminated. This is illustrated by the textures in Figure 4.28. Figure 4.28A is immediately perceived as consisting of two regions of different texture. The border between the two regions may be called a texture contour. Figure 4.28B appears at first to be uniformly textured; in fact, it also contains a subregion of different texture, as the reader may confirm by examining the elements of the texture closely. The difference between these two cases indicates clearly that two different processes are involved. In the first case, the two regions segregate immediately and effortlessly. In the second case, the border between the two regions is not immediately apparent, and its location can only be found by actively attending to the details of the texture, a process called scrutiny by Julesz (1984). Scrutiny is not a purely perceptual process—it involves a deliberate strategy on the part of the observer and therefore engages central cognitive processes—and is not considered further in this chapter. (However, the Magpie theory does provide central cognition with the means for examining the details of the texture, namely, the focusing mechanism.) The immediate and effortless perception of texture contours in Figure 4.28A, however, suggests a process nearly as low-level as the recognition of luminance contours.

A theory developed by Julesz (1981, 1984, 1986) has been very successful in accounting for texture discrimination in displays like that of Figure 4.28. According to this theory, a textured region can be regarded as an aggregation of discrete texture elements called textons, and two adjacent
Textures can be discriminated without scrutiny only if they are made up of different textons or have different texton densities. Textons are simply line segments, each characterized by a collection of parameters, including size, orientation, velocity, and binocular disparity. Isolated ends of lines, or terminators, are also textons. (Julesz originally argued that crosses formed by intersecting line segments are textons, since a region formed from crosses is immediately discriminable from a region containing the same kinds of line segments without crossings. However, Bergen and Adelson (1988) demonstrated that if the crosses were made equal in size to the textons in the other region, and the two regions equated for mean luminance, the border was no longer easily perceived, indicating that crosses need not be regarded as separate textons. Nothdurft (1990) has argued on the basis of similar demonstrations that terminators are also not textons, or at least are not very salient ones.) According to the theory, the visual system contains processes that measure the type and density of textons in small patches of the image, and a texture contour is indicated if a significant difference between two adjacent patches is detected. These processes are not sensitive to the precise locations of the textons, nor to their spatial arrangements relative to one another. For example, the two regions in Figure 4.28A differ in the orientations of the component textons, but in Figure 4.28B the texton composition of the two regions is identical—the same kinds of texton are present with the same density. A very similar theory was proposed by Beck (1982), who explicitly related the texture elements to the features detected by neurons in the visual cortex.

Although texture discrimination is usually studied with drawings like those of Figure 4.28, it should be noted that textons are assumed to possess velocity and binocular disparity among their parameters. Thus, the border between two textured regions in relative motion, in which the textons are identical except for their velocity, is easily perceived (Kaplan, 1969). The saliency of these contours is illustrated by the fact that the shape of a camouflaged animal immediately becomes apparent when the animal moves. Also, two textures that are identical when viewed monocularly can be easily discriminated in binocular vision purely on the basis of their differing binocular disparities, as convincingly shown by Julesz’s (1971) random dot stereograms (Fig. 5.19).

Although the texton theory is based on psychophysical experiments, it is not hard to provide a neurophysiological version of it. There are three stages to the computation: first, the individual textons must be detected; second, an explicit representation of the textures must be computed; third, contours must be located between regions of significantly different texture. It is plausible to assume that the first stage is realized by simple and complex cells in area V1. We have seen in previous sections that these cells have just the properties required of texton detectors: they respond to line segments of specific orientation, size, motion and disparity. End-stopped cells respond selectively to terminators. The second stage can be performed by neurons that pool the outputs of a number of these cells with RFs having similar properties and scattered over a small patch of the visual field. Such neurons respond to textures that activate several of the lower-order cells. The existence of such texture-sensitive cells is predictable from the Empiricist Principle. Most natural textures do not
consist of a random collection of textons; rather, all the component textons (at least within a localized patch) tend to have similar properties. In particular, they generally have similar motion and disparity, because they lie on the same surface, and they usually have similar size. Many natural textures also consist of textons with roughly-aligned orientations—for example, the fur of an animal, or a field of wheat. Thus, a texture-responsive neuron would tend to pool the outputs of simple or complex cells which all have similar RF properties. In neurophysiological experiments, the RF of this texture cell would be found to contain a number of relatively small subunits with properties like those of simple or complex cells, as shown in Figure 4.29. The cell itself would appear to have a large, complex RF; this might be part of the reason why some complex cells, but no simple cells, have been reported to respond to texture (Hammond & MacKay, 1977). Many of the simple or complex cells contributing to the texture cell’s RF might be end-stopped, in which case the texture cell would not be strongly activated by a single extended contour.

In area V1 of the cat, neurons that appear likely to have RF organizations like that of Figure 4.29 are a type of complex cell, discovered by Palmer and Rosenquist (1974) and called special complex by Gilbert (1977), which respond just as well or better to a short contour as to one that extends across the full length of the RF. If this is correct, these cells should be maximally activated by a textured field in which the texture elements match many of the small subfields. According to Edelstyn and Hammond (1988), complex cells that respond strongly to random-dot texture are mainly found in layer 5 and to a lesser extent in lower layer 3 in cat V1. This is precisely the distribution of special complex cells (Gilbert, 1977). Edelstyn and Hammond found that nearly all special complex cells respond strongly to texture. However, they also found that a proportion of standard complex cells respond strongly to texture. It is possible that these cells also had RFs constructed in the manner of Figure 4.29, but because some of the subfields happened to be roughly aligned, and were not end-stopped, the cells responded more to an extended contour than to a short one, and so were classified as standard complex. The prediction that texture-responsive cells have RF organizations like that of Figure 4.29 should be investigated experimentally. Note that special complex cells in V1 probably contribute little to texture perception per se because most of them are located in layer 5 and project to the superior colliculus (Palmer & Rosenquist, 1974), but there might well be similar cells in layers 2–4 of other visual areas.

As well as neurons sensitive to texture, the saliency of texture contours like those in Figure 4.28A leads us to expect there to be neurons sensitive to such contours. Such a neuron should respond when a border between regions of different texton composition lies in its RF. It might be thought that this could easily be achieved by the convergence of the outputs of texture-sensitive cells, encoding different textures and with adjacent but non-overlapping RFs, onto a single neuron. However, the
Empiricist Principle suggests that neurons of this type would be very unlikely to occur in great numbers, because at any given location in an image a border between differently-textured regions is statistically much less likely to occur than a uniform texture. Fortunately, there is a less obvious way in which neurons sensitive to texture contours can be wired-up which is consistent with the Empiricist Principle.

Allman, Miezin and McGuinness (1985) studied neurons in area MT of the owl monkey that responded to moving random-dot textures within their RFs in a direction-selective manner. They were interested in how the response of each neuron to stimulation within its RF was affected by stimulation in the surround of its RF. The majority of neurons, stimulated by a texture moving in the optimal direction in the RF, gave a considerably reduced response when the surround also contained a moving texture. Some cells were antagonized by a texture in the surround moving coherently in any direction. Others were antagonized in a direction-selective manner, with the greatest reduction in response occurring when the surround texture moved in the same direction as the texture within the RF. In short, these neurons responded well only when the texture in the RF moved differently from the texture in the surround; that is, they responded to a discontinuity in the movement of texture. I propose that the circuit responsible for these effects is the projection from layer 6 to inhibitory neurons of layer 4, introduced in the discussion of layer 6 of the canonical circuit (Fig. 3.10). This explanation requires that there are neurons in layer 6 of owl monkey MT with very large RFs that are activated by textures moving in particular directions. A fraction of the inhibitory neurons in layer 4 are assumed to receive their main excitatory input from these layer 6 cells and thus to have similar RF properties, and they inhibit neighbouring neurons in layer 4 that have relatively small RFs. Hence, for some neurons in layer 4, a moving texture that just fills the RF produces a strong response, but a larger textured field of certain kinds activates inhibitory cells that reduce or abolish the response. Neurons in layers 2–3 can inherit similar RF properties from their precursor neurons in layer 4. This explanation leads to several predictions about MT: anatomically, the projection from layer 6 to smooth neurons in layer 4, and physiologically, the properties of layer 6 and some inhibitory layer 4 cells. Born and Tootell (1992) have found that cells with antagonistic surrounds of the type studied by Allman, Miezin and McGuinness (1985) are aggregated into columns in owl monkey MT, and that within those columns there are many cells lacking antagonistic surrounds in layer 4 and upper layer 6, which is strikingly consistent with the present proposal. Note that the explanation is exactly the same as that given for the property of end-stopping in area V1. Only the pattern selectivity of the neurons involved is different; the mechanisms are proposed to be the same in both areas.

This explanation for neurons sensitive to contours between textures with different directions of motion is easily extended to neurons sensitive to other kinds of texture contour. In general, a neuron responding to a texture with certain properties in its main RF should be antagonized by texture with different properties in the surround of the RF. Other neurons of this kind have been described in macaque area V4 (Desimone & Schein, 1987). Note that these neurons are not selective for the orientation of the texture contour, but higher-order cells receiving feedforward connections from these neurons could be.

4.10 Colour and Lightness

Newton (1704) originally proposed the theory that colour is determined directly by the spectral composition of the light entering the eye. According to this theory, light containing a higher proportion of long wavelengths than short or middle wavelengths appears red, an excess of short wavelengths appears blue, and an equal amount of all wavelengths appears white. This theory gives a reasonable explanation of the colour of a uniform field, as when looking up into a clear blue sky, and of isolated lights viewed in darkness. However, in a typical scene containing many differently-coloured surfaces the theory breaks down badly. In such a scene a surface of a given colour always looks that colour, although the spectral composition of the light reflected to the retina from the surface can vary drastically with the illumination (Land & McCann, 1971). For example, sunlight at noon has a very different spectral power distribution from sunlight at dusk, and artificial lighting is different again, yet these differences have only minor effects on the colours of objects, provided the illumination contains at least a small amount of all wavelengths. This is the phenomenon of colour constancy (reviewed by Pokorny, Shevell & Smith, 1991).
Colour constancy can be understood by considering the closely-related phenomenon of lightness constancy. An achromatic surface is perceived as white if it is very light and black if it is very dark, with intermediate degrees of lightness corresponding to shades of grey. A lump of chalk under artificial lighting appears white, and a lump of coal under sunlight appears black, although the coal might actually be reflecting more light than the chalk. This indicates that the lightness of an object in a scene is not determined simply by its luminance—the amount of light reaching the eye. Furthermore, lightness constancy does not result merely from adaptation of the retina to the prevailing level of illumination; two lumps of chalk are perceived to be equally white even if one is in direct sunlight and the other in shade. (Incidentally, this example illustrates that lightness is a distinct percept from brightness. The brightness of the two lumps of chalk clearly differs, and yet their lightness is the same—the chalk in the shade is white, not grey. Unfortunately, ‘brightness’ is often used as a synonym of lightness in the literature.) The luminance of a surface is the product of the surface’s reflectance and the illumination, the reflectance being the proportion of the incident light that the surface reflects. In a complex scene, perceived lightness is much more strongly correlated with reflectance than with luminance (Land & McCann, 1971). Lightness constancy results because reflectance is an intrinsic property of a surface which does not vary with changing illumination. Thus, the problem of computing lightness is that of recovering an approximation to the reflectance of surfaces from the luminance distribution of the retinal image.

Just as lightness is related to reflectance, colour is related to the spectral reflectance of surfaces—the proportion of incident light of a given wavelength that the surface reflects, specified for each wavelength. Thus, a surface with a higher reflectance for long wavelengths appears red, one with a higher reflectance for short wavelengths appears blue, and a surface with uniform reflectance for all wavelengths appears achromatic (white or grey). The problem of colour constancy is that of computing an approximation to the spectral reflectance characteristics of surfaces from the responses of the three types of cone in the retina, factoring out the spectral power distribution of the illumination. Unfortunately, the characteristics of the illumination are not known to the observer; the only source of information for this computation is the distribution of luminance and wavelengths in the retinal image.

Several methods have been proposed for doing this. By far the most famous is the retinex algorithm, developed over many years by Edwin Land (1964, 1986; Land & McCann, 1971). The algorithm is quite successful at determining lightness and colour in artificial and natural scenes (Land,
Although it does not exactly correspond to human performance (Pokorny, Shevell & Smith, 1991). For example, it does not suffer from various lightness illusions to which humans are subject (Reid & Shapley, 1988). I will not describe the retinex algorithm in detail here, but I do want to note three important insights that the algorithm employs.

The first insight is that the problem of computing colour can be reduced to that of computing ‘lightness’ in three independent channels. Land conceived these channels as corresponding to the three classes of retinal cone. Having obtained the lightness of a point in the image for each channel independently, these three numbers can be used as coordinates in a three-dimensional colour space, yielding the colour of the point. Physiologically, the three cone types are unlikely to constitute independent channels because their outputs are pooled at a very early stage in the retina. However, as described in the review at the beginning of the chapter, the P channel in the primate’s subcortical visual pathway does indeed contain three separate sub-channels, one in which R cone outputs are opposed to G cone outputs, one in which B cone outputs are opposed to a combination of R and G outputs, and an achromatic channel in which the outputs of all three cone types are pooled. If lightness is computed separately in each of these channels, these three values can yield colour in the manner of the retinex algorithm (Livingstone & Hubel, 1984; Land, 1986). That is, the axes of the three-dimensional colour space, illustrated in Figure 4.30, are red-green, blue-yellow, and light-dark, as in the classical opponent-process theory of colour vision proposed by Hering (1920; Hurvich and Jameson, 1957). Note that the lightness value of the achromatic channel corresponds to the percept of lightness.

The second insight is that contours are important for computing lightness. Spatial variations in luminance that are caused by differences in surface reflectance tend to take the form of sharp contours—for example, where a surface joins or occludes another surface of different reflectance. Luminance variations that are caused by the illumination, on the other hand, tend to have much more shallow gradients. This suggests the heuristic that sharp contours in the retinal image can generally be assumed to mark reflectance boundaries in the scene, whereas shallow gradients are probably due to the illumination and can be ignored. The importance of sharp contours in lightness perception is demonstrated by the Craik-Cornsweet illusion, explained in Figure 4.31.

The third insight is that the ratio of luminances on either side of a contour will generally be approximately equal to the ratio of the reflectances of the two surfaces. Luminance is the product of reflectance and illumination, so if a ratio of luminances is taken, the illumination will cancel out, because it is likely to have roughly the same value on both sides of the contour.
These three observations suggest that the initial step in computing lightness and colour should be to measure the ratios of luminances across contours, independently in each of the three colour channels. In the case of the achromatic channel, it is quite plausible that this measurement is performed subcortically (Reid & Shapley, 1988). As described earlier, the cells in this channel at the level of the LGN have RFs called Type III by Wiesel and Hubel (1966). As illustrated in Figure 4.32C, these are centre-surround RFs with an ON-centre opposed by an OFF-surround, or vice versa. They therefore respond to a luminance contour lying across the RF in such a way that the ON-subfield is stimulated more than the OFF-subfield. The firing rate of the cell is determined by the contrast of the stimulus, which can be defined as \( \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{mean}}} \), where \( L_{\text{max}} \), \( L_{\text{min}} \) and \( L_{\text{mean}} \) are the maximum, minimum and mean luminances of the pattern in the RF. Since contrast is defined as a ratio of luminances, it is independent of illumination, as explained above. The relation of the firing rates of LGN cells in the P channel to contrast is approximately linear, except for high contrasts at which the firing rate saturates (Derrington & Lennie, 1984). Thus, except at high contrasts, the firing rates of Type III cells carry the information which is the prerequisite for computing lightness (Reid & Shapley, 1988). It is noteworthy that lightness constancy breaks down at high contrasts (Heinemann, 1955).

Our consideration of the retinex algorithm leads us to expect that a computation similar to that in the achromatic channel should also be performed in the two chromatic (R versus G and B versus Y) channels. By analogy with the Type III cells, we should find in these two channels cells with double-opponent RFs (Daw, 1984). For example, in the RG channel, there should be centre-surround RFs in which the centre is excited by R cones and inhibited by G cones (denoted R+G-), and the antagonistic surround is inhibited by R cones and excited by G cones (denoted R-G+), as illustrated in Figure 4.32E. Such a cell will be activated by a ‘colour contour’ lying across its RF, such that an excess of long-wavelength light falls in the centre and an excess of middle-wavelength light stimulates the surround. There should also be cells with a R-G+ centre and a R+G- surround. Similarly, in the BY channel there should be cells with a B+Y- centre and B-Y+ surround, and others with a B-Y+ centre and B+Y- surround. These cells should respond linearly with contrast, so as to be independent of illumination.

Unfortunately, no double-opponent cells have been found in the primate retina or LGN, despite many attempts to find them. Instead, the cells in the two chromatic streams of the P channel have RFs denoted Type I and Type II by Wiesel and Hubel (1966), as described in the earlier review and shown here in Figure 4.32 (A & B). If double-opponent cells do exist, they must be in the cortex. Wavelength-sensitive cells in area V1 are concentrated in discrete columns, called blobs because of their appearance in tangential sections of V1 that have been stained for the enzyme cytochrome oxidase (Livingstone & Hubel, 1984). The higher concentration of this enzyme in the blobs than in the regions between the blobs is a result of higher metabolic activity, which may in turn result from

![Figure 4.32: Various types of wavelength-selective receptive field.](image-url)
lower stimulus-selectivity of blob neurons—that is, they are active more often. Most neurons in the blobs are not orientation selective, in contrast to the simple and complex cells that lie between the blobs (Livingstone & Hubel, 1984). According to Ts’o and Gilbert (1988), most blob neurons in the macaque have RFs like those in the P layers of the LGN—that is, Types I, II and III. About a quarter of their sample, however, had a novel RF organization that they called modified Type II (Fig. 4.32D). These cells have chromatic opponency in the RF centre (R+G−, R−G+, B+Y− or B−Y+), just like Type II cells in the LGN, but also have inhibitory surrounds that appear not to be wavelength selective. That is, light of any spectral composition in the surround inhibits the response from the centre. Ts’o and Gilbert found only 5 cells out of a sample of 474 with true double-opponent RFs. Earlier studies of V1 had reported a large number of double-opponent cells (Michael, 1978; Livingstone & Hubel, 1984). However, the stimuli used in these experiments would not have been able to distinguish between double-opponent and modified Type II RFs. If double-opponent cells exist in large numbers, they must be in areas higher than V1.

Because the blob neurons have RFs mostly like those of LGN cells, they can be explained simply by excitatory input from the LGN. The surrounds of modified Type II cells may result from inhibition between blob cells. The blobs appear to receive input from both the M and the P channels, and more particularly from some sparse groups of small cells that lie between the six main layers of the LGN (Livingstone & Hubel, 1982; Fitzpatrick, Itoh & Diamond, 1983; Lachica, Beck & Casagrande, 1993). RF properties in these cell groups have not been studied in detail, but there are hints that Type II cells are concentrated there in the macaque (Livingstone & Hubel, 1984).

It might be possible to construct a double-opponent RF using the projection from layer 6 to inhibitory neurons of layer 4 (Fig. 3.10). This is the same mechanism as was used to explain the end-stopping property and the wiring of neurons sensitive to texture contours. Suppose that some inhibitory neurons in layer 4 receive their main excitatory input from layer 6 cells with large Type II RFs, and that they inhibit other layer 4 cells which have smaller Type II RFs at the same location. A neuron receiving this inhibition would effectively have a double-opponent RF, provided that it was in the same chromatic channel as the inhibition (that is, both RG or both BY) and that it had the same opponency as the inhibition (that is, both R+G−, both R−G+, both B+Y− or both B−Y+). The first of these requirements will be guaranteed if the two chromatic channels remain anatomically segregated at the level at which double-opponent RFs are constructed. This does seem to be the case at least in area V1, where each blob contains either RG cells only or BY cells only according to Ts’o and Gilbert (1988). A neuron with an RF constructed in this way would not be strictly double-opponent, because it could not be excited by stimulation in the surround, but it would give responses similar to those of a double-opponent cell to stimulation in the centre. Neurons with properties similar to this have been described in macaque area V4 (Schein & Desimone, 1990). One problem with this proposal is that the firing rate of the double-opponent cell should be related approximately linearly to contrast, so as to be independent of illumination, as explained above. It is not at all clear that this will be the case if double-opponent RFs really are constructed by the means described here.

If this is roughly correct, the most important neurons for colour perception at the level of the LGN are those with Type II RFs. What, then, is the purpose of the LGN cells with Type I RFs, which are by far the most common type in the P channel? There is no straightforward way to construct double-opponent RFs from these cells (see Fig. 4.32A & E). On the other hand, most of the neurons in area V1 have oriented RFs which respond to luminance contours, and a large proportion of these cells receive input predominantly or exclusively from the P channel. Presumably, LGN cells with Type I RFs must provide most of the input to these neurons. Earlier in the chapter the Hubel and Wiesel model for the construction of simple cell RFs was described (Fig. 4.6), in which only the luminance sensitivity of the precursor LGN cells was considered. Figure 4.33 shows the same model, but using Type I cells in the RG channel. All the cells contributing to the ON-subfield of the simple cell are assumed to be R+ centre, G− surround, and those contributing to the OFF-subfield are R− centre, G+ surround. When light that contains about equal proportions of middle and long wavelengths is used, this model reduces to that of Figure 4.6B. But consider a contour between two regions that have equal luminance but different spectral composition, such that one region activates R cones more than G cones and the other activates G cones more than R cones. The reader may confirm that such a contour will maximally activate a simple cell constructed in the manner of Figure 4.33 when it is oriented as
shown in the figure. Thus, the simple cell can respond to an isoluminant ‘colour contour’ as well as to a luminance contour. In support of this proposal, Gouras and Krüger (1979) have found that many macaque V1 neurons respond well to an oriented contour between red and green regions at all relative luminances of the two regions. Note that cells of this kind provide a poor substrate for the perception of colour itself, especially if RG Type I cells, BY Type I cells and Type III cells in the LGN all converge onto single simple cells.

The measurement of contrast at contours is only the first step in lightness and colour constancy. The lightness and colour of a region in an image is influenced not only by the luminance and spectral composition of immediately adjacent regions but also by that of more distant parts of the image (Land & McCann, 1971; Reid & Shapley, 1988). This indicates that contrast information is integrated in some way over fairly large regions of the image, although the process is certainly not global as it is in the retinex algorithm (Reid & Shapley, 1988). There is rather little detailed psychophysical evidence on this process, and no neurophysiological information. Zeki (1983) has reported neurons in area V4 which are genuinely colour-sensitive as opposed to wavelength-sensitive (that is, they respond to a particular colour independently of illumination) but there is no information on how their properties are constructed.

In conclusion, the computation of colour and lightness is one of the most under-developed parts of the present theory. As well as the incomplete account of constancy, there are various lighting effects that have not been considered at all, such as shading, transparency, and luminosity. This is partly due to the paucity of neurophysiological data. However, the theory does seem consistent with what little is known of colour and lightness processing in area V1.

### 4.11 Summary

This concludes the account of early visual processing. The low-level areas of the visual cortex are proposed to extract a vast array of elementary features from the retinal image. Most features are short segments of contour with properties including position, orientation, size, motion, and binocular disparity. Texture is analysed, and contour segments are explicitly represented between adjacent patches of the image that differ in the form, motion, or depth of texture. Small patches of colour and lightness are also represented. All of these different kinds of feature are extracted by a uniform set of mechanisms. In particular, the response properties of neurons are mainly determined by the feedforward excitatory input they receive, selectivity is sharpened by the local circuitry, C-units generalize responses of S-units over position on the retina, and all features may be acquired in accordance with the Empiricist Principle. We now move on to consider how these local features are integrated to yield the perception of shape.
Chapter 5

Visual Shape Perception

It is common in perceptual psychology to regard the visual system as consisting of two processing stages, the first performing a fast, automatic, highly parallel, localized analysis of the visual image, the second carrying out slow, complex, attention-demanding, limited-capacity operations that are necessary for organizing the information delivered by the first stage (e.g., Neisser, 1967; Ullman, 1984; Treisman, 1988). On the contrary, this chapter will argue that the mechanisms are basically the same at all levels, although there are perhaps some differences in emphasis—some mechanisms assume a greater importance at later levels of analysis than at earlier stages. The chapter begins with an overview of how the theory introduced in Chapter 3 can be applied to shape perception and object recognition. The remainder of the chapter considers a number of special topics: the Gestalt grouping effects, illusory contours, perception of three-dimensional shape in pictures, multistability of pictures like the Necker cube, hysteresis in the organization of visual patterns, stereopsis, the figure and ground phenomenon, and selective attention as studied by visual search.

Unfortunately, there is very little neurophysiological information available in this field compared with early processing. This deficit is partially offset by the fact that shape perception is relatively more accessible to conscious introspection, with the result that there is a large psychological literature, dating back to the classical studies of the Gestalt school. Because I will be making some use of introspective observations, it is necessary to make a brief statement regarding the relation between neural activity and the phenomenology of conscious awareness, without going into the philosophical complexities of this issue (see Churchland, 1986; Dennett, 1991). I assume that visual consciousness is produced by certain computations in various parts of the visual cortex. For each quality of the conscious visual image there must be some corresponding representation in active storage, and the occurrence of that representation in active storage is both necessary and sufficient for that quality to be experienced. It seems that not all of the information represented in the visual cortex is available to consciousness—for example, we are aware of the colours of objects but not of the spectral composition of light, even though the wavelength-selective neurons in area V1 are more closely related to the latter than the former—but the question of why this is the case will not be addressed here.

The theory that consciousness is produced by computations performed by the brain is called functionalism. Functionalism is opposed to dualism because it is assumed that the computations are implemented in the physical material of the brain; indeed, the Magpie theory is based on the assumption that the relation between computation and the physical substrate is a very intimate one in the case of the cortex. But functionalism is also not quite the same as materialism, because in principle the same computations could be carried out by a completely different physical substrate, such as a silicon-based machine. It is quite possible that, in practice, a thinking machine would have to be organized rather like the brain, with artificial neurons and the like, but that is beside the point. It is beside the point because the artificial neurons could be materially completely different from our neurons; they might, for example, be analogue electric circuits à la Figure 3.1B. I remain agnostic on the question of whether a von Neumann computer, a Turing machine, or Searle’s (1980) infamous Chinese Room could ever be conscious. My general position on philosophical thought experiments of this kind is that the scenarios they invite us to imagine are so wildly unrealistic that to draw conclusions from them purely on the basis of intuition (as Searle does) is futile.
Introspective evidence can be misleading if it is not treated cautiously. For example, many people are struck by the unitary nature of the conscious visual image and by the paradox that the corresponding computations are widely distributed across many different areas of the cortex; this is probably part of the motivation behind the current theoretical interest in the binding problem (p. 72). But unitary consciousness and distributed brain activity are not really inconsistent. The reader is referred to the excellent analysis by Dennett (1991).

5.1 Perception of Shape in General

The shape of an object is the three-dimensional organization of its edges and surfaces relative to one another. Under normal conditions, shape is perceived effortlessly, even if the observer is not familiar with the objects in question. This is strikingly illustrated by the ease with which shape can be perceived in scenes created by special effects in science fiction films. As a more homely example, many people have had the experience of not recognizing a familiar object when it is seen from an unfamiliar angle, but with no accompanying difficulty in perceiving the visible part of its shape. These observations indicate that shape perception is a distinct processing stage in vision, following early processing but preceding the recognition of familiar objects. Having said this, the basic premise of the present theory is that the computational mechanisms responsible for shape perception are very similar to those applied in early processing and object recognition. That is, these separate processing stages may well be carried out in different regions of the visual cortex, but the neural circuitry within these regions is approximately the same.

In outline, the theory of shape perception is as follows. Neurons in certain areas of the visual cortex are proposed to encode as their characteristic patterns an enormous vocabulary of shape fragments. A shape fragment is simply a piece of shape—that is, a particular three-dimensional configuration of edge segments and pieces of surface. Some hypothetical examples of shape fragments are suggested in Figure 5.1. Neurons encoding relatively simple shape fragments form feedforward excitatory connections with neurons encoding more elaborate shape fragments, which themselves form feedforward connections with neurons encoding still more elaborate shape fragments. Shape perception basically consists of the hierarchical, feedforward activation of shape fragments by the outputs of the early-processing stage of analysis. The firing rate of a neuron encoding a shape fragment is a graded measure of how well that shape fragment matches the retinal image.
of a given object’s shape (the representation of its shape in active storage) is the distributed pattern of activity that it arouses in this region of the visual cortex.

Neurons encoding shape fragments do not merely represent templates to be matched against the retinal image, responding only if there is a precise match. There are two reasons for this. First, neurons are quite broadly tuned, and can therefore tolerate a certain amount of distortion in the stimulus pattern. Second, and more importantly, the generalizing properties of C-units allow neurons to achieve a limited degree of stimulus invariance. Recall that, in the idealized canonical circuit, S-units are assumed to be concentrated in layer 4, where the feedforward input arrives, whereas C-units are concentrated in layers 2–3, which receive the output of layer 4. The function of S-units in shape perception is to build more elaborate shape fragments from the less elaborate fragments or features they receive input from. The function of the C-units is to recognize similar fragments, but to be relatively less specific for the stimulus position and size on the retina as well as for the three-dimensional orientation of the stimulus shape fragment in space. The progression up through the hierarchy of cortical areas is accompanied by a step-by-step increase in shape fragment complexity and average receptive field (RF) size, and by a corresponding decrease in specificity for position, size and orientation within the RF.

It is important to realize that the perception of an object is not proposed in general to require the activation of a single ‘pontifical’ neuron. Consider the perception of an unfamiliar object. The function of shape perception is to compute an explicit representation in active storage of the shape information that the observer may need, for example, to manipulate the object with his or her hands. There is no reason why the activation of an ensemble of neurons, each encoding a fragment of the shape of the whole object, would not be adequate for this purpose. In any case, it is clearly out of the question to have one neuron encoding each possible object, because there are so many possible objects. Furthermore, all such neurons would have to be wired-up genetically since, by definition, the appearance of an unfamiliar object could not have been learned. It is perfectly conceivable, however, to have a (admittedly very large) repertoire of shape fragments, each of which can participate in the representation of many different objects. Because the representation of a given object in active storage is population coded by a collection of broadly tuned neurons, only a finite number of shape fragments needs to be held in latent storage to cover a continuum of possible shapes in the environment.

The simplest shape fragments are those built directly from the short segments of contour extracted in early processing. (As noted earlier, the decision of where to draw the boundary between early and late processing is a rather arbitrary one in this theory, so these neurons could just as well be viewed as the culmination of early processing as the beginning of shape perception.) These fragments are likely to be also single contours, but extending over large angles of the visual field. Convergence of short, roughly collinear contour segments can construct long curved contours (with large angle of curvature) as well as straight ones. Contours containing corners (that is, curves with small angle of curvature) can be constructed using precursor cells with end-stopped RFs, as shown in Figure 5.1. Another important aspect of the construction of shape fragments is the representation by single cells of different types of contour, caused by convergence of inputs from lower-order cells that represent those contours in isolation. The existence of such neurons is consistent with the Empiricist Principle, since in natural images the boundary between two regions is quite likely to be a discontinuity in colour and texture as well as mean luminance. For example, a single neuron may represent both a luminance contour (of low spatial frequency) and the boundary between two textures. Such a cell responds most strongly to a conjunction of both the luminance and the texture contour, but it also responds to either the luminance or the texture contour alone. Neurons that respond to different types of contour go a long way towards explaining how we can perceive shapes independently of the type of contour with which they are defined, as in Figure 5.2. It was explained in Chapter 4 that many neurons in area V1 respond to luminance contours regardless of their contrast polarity, and that many also respond to isoluminant colour contours. The convergence of different types of contour in single shape fragments can be seen as extending this generalization process.

More elaborate shape fragments are constructed hierarchically. Patches of texture, lightness and colour may also be incorporated in shape fragments—that is, fragments represent characteristics of the surfaces as well as the edges of shapes. Note that shape fragments are three-dimensional rather
than two-dimensional patterns. This is because the lower-order neurons encoding contours and texture patches are often selective for binocular disparity, and so a higher-order neuron representing a shape fragment may require a particular three-dimensional stimulus for maximal activation (Fig. 5.1).

Because the neurons encoding shape fragments receive inputs from many lower-order neurons, and therefore integrate information from broad regions of the retinal image, they reliably signal the presence of shapes in the viewed scene. This is not generally the case for neurons at the early processing stage of analysis; because they perform relatively local measurements of the image, they can be fooled by noise or local luminance variations into signalling features of the image that do not correspond to anything real in the outside world. Simple cells in area V1 are not literally ‘edge detectors’. Neurons encoding shape fragments, on the other hand, can legitimately be regarded as shape detectors. For example, a neuron encoding a long contour is very unlikely to be strongly activated by chance luminance variations alone, since they would have to affect most of the precursor neurons simultaneously. On the other hand, if there really was an edge of an object of the appropriate kind, the neuron would respond well even if a few of the precursor cells were not activated.

Although the activation of neurons encoding shape fragments is assumed to be basically a feedforward process, the local circuitry and the feedback and lateral connections are also important. The feedback and lateral connections are responsible for the contextual enhancement effects described in Chapter 3 (Fig. 3.9). In particular, the feedback connections which return from neurons encoding shape fragments to lower-level neurons encoding short contour segments are suggested to be at least partly responsible for the so-called object-superiority effect. In the original demonstration of this effect, Weisstein and Harris (1974) measured the accuracy with which subjects could identify line segments. Four different line segments were used, differing in orientation and location relative to the point of fixation, and in each trial the subject had to identify which of the four segments was displayed. The lines were flashed on a screen sufficiently briefly (about 40 msec) that the subjects misidentified the segments on a large proportion of the trials. Weisstein and Harris were interested in how the subjects’ accuracy was affected by various different context patterns in which the line segments were embedded. They found that the fewest errors were made when the line to be identified was part of a pattern that could be interpreted as a coherent three-dimensional shape. Such a context can even improve identification accuracy relative to a condition in which the line segment is displayed with no context at all (Williams & Weisstein, 1978). Suppose that neurons encoding shape fragments are activated more by coherent, three-dimensional context patterns than by less structured, two-dimensional patterns, and that feedback from these neurons enhances the activity of neurons signalling the line segments themselves, hence improving identification accuracy. This explanation of the object-superiority effect is similar to that given by McClelland and Rumelhart’s (1981) interactive-activation model of the closely-related word-superiority effect. Feedback may not be the

Figure 5.2: Shapes can be defined by several different types of contour: (A, B) luminance, (C) colour, (D) texture form, (E) texture disparity, and (F) texture motion.
complete explanation of the effect, however. A single three-dimensional context pattern may cause the activation of different shape fragments when different line segments are added to it. These differences presumably make the displays more distinguishable from one another, which may contribute to the observer’s ability to identify the line segments (Enns & Gilani, 1988).

The focusing mechanism is relatively more important in shape perception than it is in early processing, because of the greater attenuation suffered by unfocused stimuli at higher levels of the visual hierarchy. The main purpose of this mechanism is to overcome the problem of interference between representations in active storage. Because neurons encoding shape fragments have fairly large RFs, the collections of neurons activated by the various objects in the scene are likely to overlap considerably. Focusing is required to restrict the analysis of shape to one part of the scene at a time.

It is assumed that shape fragments are acquired from experience. Although it is not inconceivable that neurons could be genetically wired to encode shape fragments, this would seem to place an enormous burden on the genetic code, since shape fragments are more complicated and have a much greater variety than the features used in early processing. The Empiricist Principle leads us to expect that the kinds of shape fragment stored should reflect the common shape structures present in the environment. For example, shape fragments having the form of three intersecting flat surfaces, like the corner of a box, can be predicted to be common in humans living in modern buildings (and laboratory animals, for that matter), whereas they would be uncommon in wild animals or humans living in less artificial environments. C-units could acquire their limited stimulus-invariance properties by means of the trace mechanism, as Földiák (1991) suggests. Consider an object undergoing a three-dimensional translation or rotation relative to the observer. The changing aspect of the object in the retinal image activates a succession of shape fragments encoded by different S-units, and by integrating these inputs over a period of 0.5 sec or so a C-unit can acquire a characteristic pattern that enables it to respond well to any of the experienced views.

We now consider some evidence for the existence of shape fragments. Palmer (1977) found psychophysical evidence for hierarchical structure in the representation of visual line patterns. More direct evidence comes from the perceived fragmentation of images of patterns stabilized on the retina. As noted in the previous chapter (p. 119), stabilized images rapidly fade and disappear from consciousness. Parts of the image can also reappear intermittently, although this might be because of slippage of the apparatus used to project the image onto the retina, resulting in imperfect stabilization. The images do not disappear or reappear all at once, but do so in a fragmentary fashion, with some pieces of the pattern vanishing while others remain visible. Furthermore, it seems that the fragmentation is not random, but that coherent structures within the pattern tend to disappear and reappear as units (Pritchard, Heron & Hebb, 1960; Evans, 1965; Ditchburn, 1973). An example is illustrated in Figure 5.3. Although this evidence should be interpreted with caution, because it depends entirely on introspection and is difficult to quantify, it does lend support to the idea that there are explicit representations in the human visual system of pattern fragments that are more complex than short contour segments.

It is well-established that bilateral excision of the inferotemporal (IT) cortex in monkeys causes a severe impairment of object recognition (Mishkin, 1982). This is a large region in the temporal lobe that contains several distinct areas (all the areas with initials including IT in Figure 2.9). Early neurophysiological studies of IT had shown that a fraction of the neurons respond preferentially to certain kinds of complex visual pattern, including faces (Gross, Rocha-Miranda & Bender, 1972; Schwartz, Desimone, Albright & Gross, 1983; Desimone, Albright, Gross & Bruce, 1984; reviewed by Gross, 1992). However, the majority of cells appeared to respond in a general way to a variety of simple edge- and bar-like stimuli, albeit with very large RFs. More recently, Tanaka, Saito, Fukada and Moriya (1991) studied the visual responses of over 700 cells in macaque IT using a large variety of three-dimensional shapes, colours and textures, and attempted to isolate the optimal stimulus pattern for each cell. They found that 69% of the neurons that gave clear responses in the anterior two-thirds of IT were maximally responsive to a particular shape, or a combination of shape with a colour or texture. Some cells preferred simple geometric shapes like crosses, circles, triangles, or rectangles; others required combinations of such shapes for a strong response. Many cells preferred particular shapes that were textured in a certain way, and some also had a colour preference. (With neurons responding to such complex stimuli it is of course difficult to know whether one has isolated...
the optimal stimulus; it is always possible that an untested pattern would elicit a greater response. So far, few neurons have been reported to be tuned to patterns like those of Figure 5.1, but such patterns have generally not been tested.) All the cells studied by Tanaka et al. were quite broadly tuned in that they gave good responses to patterns that resembled the optimal one, and were not very selective for the size or position of the pattern in the large RF, although the majority were quite selective for orientation. Tanaka et al. proposed that the neurons in anterior IT encode patterns that commonly occur in natural images, and that particular objects are generally represented by the activity of several such neurons (see also Tanaka, 1993). Interestingly, these complex patterns are organized in a columnar fashion in IT: nearby neurons tend to encode similar patterns (Fujita, Tanaka, Ito & Cheng, 1992).

It is appropriate at this point to consider the localization of shape perception in the visual cortex. The evidence reviewed above suggests that the anterior IT region is the most important site for elaborate shape fragments in primates. However, lower-order shape fragments can be expected to be found in other areas that lie in the pathway from V1 to anterior IT, such as V4 and posterior IT (Fig. 2.9). Kobatake and Tanaka (1994) have recently found neurons in these two areas that are selective for complex stimulus patterns, although with RFs smaller than those of anterior IT. This is strongly suggestive of a hierarchical construction of these RF properties (Tanaka, 1993). There is substantial neurophysiological and neuropsychological evidence for segregation of different kinds of visual information in separate cortical areas (reviewed by Van Essen, 1985; Livingstone & Hubel, 1987; Maunsell & Newsome, 1987; DeYoe & Van Essen, 1988). For example, V4 is known to be important for the processing of colour, and MT for the processing of motion. Perhaps the best way to view segregation of function in the visual cortex is to relate it to the division between the M and P channels in the retinocortical pathway (p. 84). The rich intrinsic circuitry of each cortical area and the connectivity between areas make it seem very unlikely that the M and P streams remain segregated in the cortex. Nevertheless, it is certainly plausible that different areas beyond V1 receive their input predominantly from one channel or the other. For example, the direct connections from V1 to MT originate from layer 4b, which is associated with the M channel, whereas the pathway from V1 via V2 to V4 arises from the upper layers, which are associated with the P channel (Livingstone & Hubel, 1987). Because neurons in the two channels have different physiological properties at the level of the LGN, it is possible to differentially stimulate the two channels by devising appropriate visual stimuli. For example, cells in the M channel can only poorly discriminate differences in wavelength, and so an image in which luminance is uniform, but a shape is defined purely by colour variations, presumably activates the P channel well but the M channel weakly. Several psychophysical studies have employed this technique to explore segregation of function in the human visual system. Livingstone and Hubel (1987) review many of these studies as well as reporting a series of experiments of their own, and argue that there is indeed a segregation of functions between the M and P channels in shape perception. Processing of motion and depth is mainly performed by the M channel, whereas the P channel provides colour perception as well as a high-resolution analysis of patterns. These psychophysical results reinforce what the physiological studies of monkeys suggests, that shape
fragments having different properties are not all jumbled up together but are segregated in different areas, at least to some extent. Note that shape perception per se is not confined to either channel, since several kinds of cue (including motion, depth and colour) can provide information about shape. And, again, the anatomy makes it very unlikely that there is no mixture of M and P information.

It is worth stressing once again that the segregation of different kinds of shape fragment in different regions of the visual cortex is not inconsistent with the basic claim of the present theory, that a single system of general computational mechanisms is applied throughout the cortex. The proposal is that different areas are responsible for different functions because they apply the same mechanisms to different kinds of input information. This was discussed in detail in Chapter 2. For example, colour processing occurs in areas that mainly receive P input because only P cells in the retina are wavelength selective, whereas areas that mainly receive M input are more important for motion processing because M cells in the retina have greater sensitivity to moving stimuli.

5.2 Object Recognition

Here and throughout this chapter I am using the term object in a very general way. I include objects with fairly invariant shapes (e.g., furniture, buildings, cars, trees, the moon) as well as objects with more flexible shapes (e.g., people, animals, clothing). It is also convenient to treat whole scenes as single objects (e.g., landscapes) as well as parts of things (e.g., hands, fingers). In short, an object is any particular thing that an observer can remember and recognize as a unit by its visual characteristics, especially its shape. The recognition of a familiar object is generally assumed to require two stages: first, observation of the object causes the access of a stored representation of its visual characteristics; second, this representation can be used as a key to access further representations that encode knowledge of the object, such as its name, its use, past experiences involving it, and the like. The second stage is probably best regarded as a case of language or memory processing rather than visual perception per se, and is not considered further. The first stage is the topic of this section.

The general theory of the first stage in object recognition is as follows. For each familiar object there is a collection of neurons in the cortex, each of which encodes as its characteristic pattern the shape of the object and probably also aspects of its colour and texture. The proposal is therefore a simple continuation of the process of shape perception that was described in the previous section. Here, instead of neurons encoding general fragments of shape, they encode particular objects. As with shape fragments, broad tuning and the generalizing capability of C-units allow each neuron to respond well over a range of positions, sizes, and orientations in its RF. However, the theory of how C-units acquire their characteristic patterns does not make it likely that a single neuron could generalize its responses over all possible three-dimensional orientations of the object. It is therefore assumed that each neuron encodes a limited range of views of the object, and that several neurons are necessary to cover several significantly-different views of each object. Also, in the case of objects with flexible shapes, there may be different neurons encoding significantly-different common views—for example, there may be different cells for a cat walking, sitting, and curled-up asleep. Each neuron receives feedforward excitatory connections from lower-level neurons, most of which encode shape fragments, although some of the connections might come directly from early processing. Because objects can be hierarchically organized, neurons encoding whole objects can receive inputs from neurons encoding component objects. The percept of an object (its representation in active storage) is the pattern of activity that it arouses in this population of neurons. Theories of object recognition broadly similar to this one have been suggested previously by Barlow (1972), Gross (1992) and Rolls (1992); see also Poggio and Edelman (1990) and Edelman and Weinshall (1991) for related computational models.

Theories of this kind are often dismissed as implying the existence of the notorious grandmother cell—a neuron that fires if, and only if, one looks at one’s grandmother. This implication is generally regarded as a reductio ad absurdum of the theory. However, careful consideration of the issues shows that the idea is not as implausible as it at first appears. First, it is suggested not that there is a single grandmother cell, but many of them, each encoding a limited range of views of one’s grandmother! Crucially, these neurons are broadly tuned and show a degree of stimulus-invariance. This means that there does not need to be one neuron for every possible view. Another result of broad tuning is that the representation of the current view of grandmother in active storage is population coded, with the firing rate of each grandmother cell encoding the degree of
match between the current view and the view represented by the cell’s characteristic pattern. It is the pattern of activity over this population of cells that represents the current view in active storage. One way to interpret this is that a continuum of possible views of grandmother can be represented by interpolation between the finite number of views encoded by the cells (cf. Poggio, 1990; Poggio & Edelman, 1990). Population coding also overcomes the objection that the death of one neuron could cause the loss of the ability to recognize grandmother. Second, it is not suggested that the perception of grandmother consists only of the activation of these few cells. If all of one’s grandmother cells were to be destroyed by brain damage one would indeed lose the ability to recognize a woman as one’s grandmother—although perhaps only temporarily, because one could re-learn her appearance using other neurons if the lesion was not too extensive. One would not, however, lose the ability to perceive that woman, because the processes of central cognition would still have access to lower-level neurons that signal her shape, colour, texture, motion, and so on. It is assumed that object-specific neurons are needed only as keys to access associated knowledge, not as the sole substrate of the perception of those objects. Third, it is not claimed that there is a collection of cells for every conceivable object. Only a finite (albeit large) number of familiar objects are assumed to be encoded by single neurons. There is unlikely to be a ‘yellow Volkswagen cell’, for example, unless one happens to be familiar with a yellow Volkswagen.

This theory claims that several representations (encoded by different neurons) are required to cover all possible viewing angles of each object. It also claims that there are a number of separate representations for different shapes that can be attained by a single flexible object. This perspective follows from the limited generalizing capability of any single C-unit, and it can be contrasted with some other theories of object recognition which posit a single view- and shape-invariant representation for each object (e.g., Marr & Nishihara, 1978). From a purely computational perspective (that is, ignoring empirical evidence) the scheme proposed here has two advantages and two disadvantages compared with single-representation models. The first advantage is that it is fast. Each neuron is activated in a simple, feedforward manner (although the local circuitry and feedback and lateral connections complicate matters somewhat). Single-representation models, by contrast, face the problem of gaining access to the object’s representation from the view-dependent information in the retinal image. This access problem is even worse in the case of flexible objects like cats. The computations needed to solve the problem are likely to be elaborate, and will therefore take much more time than the feedforward activation of a hierarchy of neurons. The second advantage of the present theory is that it is very general, placing very few constraints on what objects can be represented. Single-representation theories, in order to make the access problem more tractable, generally include limiting assumptions about the nature of the objects, such as that they are rigid, or that they are seen at an angle at which their ‘main axis’ is not too foreshortened. The first disadvantage is that the scheme does not make economical use of latent storage, since several representations need to be stored for each object instead of just one. However, it is likely that this disadvantage is strongly outweighed, in a Darwinian context, by the advantage in processing speed. An advantage of a fraction of a second in recognizing an object as a hungry leopard may make the difference between escaping and being eaten; storage space, on the other hand, is biologically quite cheap, there being some tens of thousands of neurons beneath each square millimetre of cortical surface.

The second and more important disadvantage of the present theory is the problem of recognizing a familiar object when its image does not match well with any of the stored views. This may happen when an object is seen from an unfamiliar angle, or when a flexible object takes on an unfamiliar shape. In fact, there is evidence that objects are difficult to recognize when viewed from an unfamiliar angle. Rock, DiVita and Barbeito (1981) and Rock and DiVita (1987) studied the recognition of novel wire shapes as a function of the angle of view. Subjects were shown one such object from a single viewing angle, and were then tested for their ability to recognize whether an object was the same as the one studied. Subjects were poor at recognizing the rotated shape. A similar experiment was performed by Bülthoff and Edelman (1992) using pictured wire-frame objects. Subjects were first given a training period in which they studied one object from two particular viewing angles 75° apart, and were then tested for their ability to recognize the object from a variety of angles. Error rates were lowest for the two practised viewing angles, higher for angles in between the practised views, and much higher for other unpractised viewing angles. The fact that error rates
were relatively low for views between the two studied ones may be explicable by population coding: if neurons encoding the studied views are broadly tuned, an intermediate view would activate neurons from both pools (cf. Poggio & Edelman, 1990). Subsequent experiments confirmed that subjects were poor at recognizing these wire-frame objects at unfamiliar angles even when they were viewed stereoscopically (Edelman & Bülthoff, 1992). These experiments appear to contradict the proposal that object representations are view-invariant, and provide support for a multiple-view scheme. It should be noted, however, that the wire-frame objects used in these experiments are exceptionally hard to recognize; more natural objects, with surfaces and distinctive parts, are much more easily recognized at novel viewing angles (e.g., Biederman & Gerhardstein, 1993; Farah, Rochlin & Klein, 1994).

It is feasible that the recognition of familiar objects at views that do not match any of the stored representations can be achieved by the use of visual imagery. Shepard and Metzler (1971) were the first to describe a process that is particularly relevant here. They presented subjects with pairs of pictures of simple three-dimensional objects built from cubes, as in Figure 5.4, and required them to judge whether the objects in each pair were identical or mirror images. The angle of view of each object was randomized, and the subjects’ reaction time was measured as a function of the three-dimensional angle between the objects of each pair. The result was that reaction time increased linearly with angle. The explanation offered by Shepard and Metzler was that the subjects were ‘mentally rotating’ one of the objects to the same orientation as the other, so that their shapes could be compared. They argued that this transformation was analogous to a physical rotation of the object in that it was necessary to pass through intermediate states corresponding to intermediate orientations, and that the greater the angle between the objects the longer the rotation took. Further studies provided additional support for this explanation (Metzler & Shepard, 1974; Shepard & Cooper, 1982). It was natural to assume that the process discovered by Shepard and Metzler had some role in the viewpoint-invariant recognition of objects (e.g., Rock, 1973). Tests of this hypothesis generally yielded negative results, showing that the time taken to recognize familiar two-dimensional shapes like letters, for example, is largely independent of their orientation. Corballis (1988) concluded that mental rotation is not usually required to recognize objects, except when it is important to determine their ‘handedness’ (as in Shepard and Metzler’s experiment), because, he proposed, the stored representations of objects do not discriminate between mirror images. However, Tarr and Pinker (1989) pointed out that the subjects in these studies had generally been given considerable practice at recognizing shapes at the studied orientations, and so they did not rule out the possibility that mental rotation could be required for recognizing shapes at unpractised orientations. Tarr and Pinker gave their subjects practice at recognizing novel two-dimensional patterns at certain orientations, and then measured the time the subjects took to discriminate these shapes from their mirror images at various orientations. They found that the time taken to make this judgement was roughly constant for the practised orientations, but that for orientations at which the subjects had not been given practice the time taken increased monotonically with the angle between the stimulus orientation and one of the practised orientations. Tarr and Pinker explained this result by suggesting that the subjects stored multiple, orientation-specific representations of each shape, one for each of the practised orientations, but that mental rotation was necessary to match a shape at a novel orientation against one of the stored representations.

I make no attempt in this chapter to explain the neural mechanisms underlying visual imagery. However, it is necessary to address one paradoxical aspect of the above explanation. In order to mentally transform an image to a stored representation of an object, it is first necessary to access that representation. Apparently, therefore, it is necessary to recognize the object before performing the transformation! One possible way out of this dilemma is that central cognition uses partial information in the image to activate a candidate object representation, and then uses imagery to test whether it adequately matches the image; if not, another candidate representation is selected, and so on. Metaphorically, the observer forms a hypothesis as to what the object is, then evaluates the hypothesis using imagery. The selection of candidate object representations might be made using information that is relatively view-independent and that is extracted at a lower level in the visual system, such as size, colour and texture; knowledge of what the object is likely to be might also be used. The fact that object representations can be accessed using only impoverished information from
Figure 5.4: Pictures used in the Shepard and Metzler study of mental rotation. (A) The objects can be made congruent by an 80° rotation in the picture plane. (B) The objects can be made congruent by an 80° rotation in depth. (C) The objects cannot be made congruent by rotation. The time taken by subjects to decide whether a pair is congruent is a linear function of the rotation angle. (Adapted from R. N. Shepard and J. Metzler, 1971, Mental rotation of three-dimensional objects, *Science* 171: 701–703, Fig. 1. Copyright 1971 American Association for the Advancement of Science. Reproduced with permission.)
the image is illustrated by the ease with which we can recognize objects that are partly occluded from view, as well as our ability to recognize objects in cartoons, fragmented images like the well-known Dalmatian dog picture (Gregory, 1970, p. 14), and even silhouettes. These stimuli presumably only weakly activate the neurons encoding objects, which suggests the involvement of a central cognitive process actively helping to interpret the image.

An unfortunate drawback of the theory of object recognition described here, from the point of view of evaluating it empirically, is that it does not unequivocally predict when mental rotation will need to be used. Because neurons are assumed to show significant but not complete viewpoint-invariance, a given view of a familiar object may or may not activate neurons representing that object, and so mental rotation may or may not be required. Repeated exposure to a particular view will certainly eliminate any need for mental rotation, however. Even in the case of an unfamiliar object it is possible that two significantly different views will activate similar populations of shape fragments, because the neurons encoding shape fragments also show limited viewpoint-invariance. On the other hand, two completely different views of an unfamiliar object should certainly require mental rotation for them to be compared. This is basically the result obtained by Biederman and Gerhardstein (1993, Experiment 3), but detailed modelling will be required before it can be claimed that the present theory accounts for the empirical facts as well as Biederman’s (1987) theory.

The claim that single objects have multiple representations in latent storage might appear to be at variance with one of the most compelling of visual phenomena, that of object constancy. There may be no single neuron that responds to every possible retinal position, size and orientation of an object, yet it is always perceived as the same object. As noted above, there is a second stage in object recognition involving the retrieval of associated knowledge, and so object constancy might be partly explained by the multiple representations all having similar semantic associations. A related issue is the representation of object categories. So far I have only considered the recognition of particular objects, but we also have the ability to recognize a novel object as an instance of a familiar category. This has generally been taken to imply that humans store abstract representations of categories, but this is not necessarily the case (cf. Medin & Schaffer, 1978; Jacoby & Brooks, 1984; Hintzman, 1986). A novel object generally has a roughly similar size and shape to at least some other exemplars of the same category, and so it presumably activates neurons encoding particular exemplars that have been encountered previously. These familiar exemplars will all have some semantic associations in common, and so central cognition could assume these common associations to apply also to the novel object.

So much for the psychological evidence. The most important neurophysiological evidence for the present account of object recognition is provided by the neurons in the macaque temporal cortex that respond selectively to faces, both of monkeys and humans (reviewed by Perrett, Mistlin & Chitty, 1987; Desimone, 1991; Perrett, Hietanen, Oram & Benson, 1992). Faces are visual patterns of great social significance, and face perception is one of the most important uses of vision (see Bruce, 1988, for general review). The importance of faces for humans and for primates generally has led some researchers to propose that faces are processed by specially-evolved mechanisms in the cortex. Of course, the assumption of the present theory is that faces are not special—the mechanisms of face recognition are basically the same as the mechanisms of recognition of any other kind of visual pattern. This claim is perfectly compatible, however, with faces being analysed in a certain specific region of the cortex. Localization of face representation in the human cortex is suggested by the neuropsychological disorder of prosopagnosia, the loss of the ability to recognize faces following brain damage, although it remains controversial whether this deficit is really specific for faces (Damasio, Tranel & Damasio, 1990).

Face-selective neurons are found in several parts of the macaque temporal cortex, and are especially common in a part of IT cortex in the superior temporal sulcus, and in the superior temporal polysensory area (Desimone, 1991; Gross, 1992; Perrett et al., 1992; Rolls, 1992). It is quite likely that the face cells in these two regions serve different functions, because their connections with other parts of the cortex are very different (Desimone, 1991). These neurons do not respond to simple stimuli like line segments, nor to complex objects other than faces, and their activity does not appear to be caused merely by arousal or by emotional responses that are associated with faces (Perrett, Rolls & Caan, 1982; Desimone, Albright, Gross & Bruce, 1984). They respond well to faces over a range of positions within large RFs, and are not very selective for face size, distance, colour, contrast, or
lighting conditions (Perrett et al., 1982; Perrett, Smith, Potter, Mistlin, Head, Milner & Jeeves, 1984; Desimone et al., 1984; Rolls & Baylis, 1986; Hietanen, Perrett, Oram, Benson & Dittrich, 1992; Tovee, Rolls & Azzopardi, 1994). At least some of them respond more to certain faces than to others (Perrett et al., 1984; Baylis, Rolls & Leonard, 1985; Young & Yamane, 1992). However, a particular face presumably activates many different face cells to different degrees, which suggests (although it does not necessarily imply) that particular faces are population coded (Baylis et al., 1985; Gross, 1992; Rolls, 1992; Young & Yamane, 1992). Some cells appear to respond only to particular facial features, such as the mouth or the eyes; others respond to a whole face, since their activity is reduced by masking various parts of the face or by presenting the parts in the wrong spatial configuration (Perrett et al., 1982; Desimone et al., 1984). This suggests that there is a hierarchical organization among these neurons (Perrett et al., 1987). These observations are consistent with the general outline of object recognition proposed above.

To what extent do face cells show viewpoint invariance? Most face cells in the macaque are not, in fact, very specific for face orientation in the frontoparallel plane, responding nearly as well to horizontal and inverted faces as to upright ones (Perrett et al., 1982, 1984, 1985; Desimone et al., 1984). Incidentally, this lack of selectivity for face orientation might seem surprising given that humans are extremely bad at recognizing inverted faces (Yin, 1969; Bruce, 1988). It makes more sense, however, when the monkey’s arboreal lifestyle is considered—they are presumably used to seeing one another upside-down. Interestingly, face-sensitive neurons in the sheep (!) only respond to faces in near-upright orientations (Kendrick & Baldwin, 1987), consistent with the fact that sheep, like humans, normally see one another upright. With regard to orientation about other axes, the majority of face cells in the macaque are quite selective for head view: some respond well only to the frontal view, others to a profile (Perrett et al., 1985; Perrett, Oram, Harries, Bevan, Hietanen, Benson & Thomas, 1991). However, some cells are activated by many different views of the head (Perrett et al., 1984, 1985, 1991; Hasselmo, Rolls, Baylis & Nalwa, 1989). Perrett et al. (1987) suggest that these neurons are pooling the outputs of several view-selective cells. This is supported by the fact that single face cells often respond to different views with different latencies, the latency difference typically being 10–60 msec (Perrett et al., 1985). Also, face cells have been reported in IT that are strongly orientation-selective (Tanaka, Saito, Fukada & Moriya, 1991).

Following Rolls (1992), the present theory assumes that some neurons (the C-units) can integrate their inputs over an interval of 0.5 sec or so, and that the head may be seen changing its orientation during this interval, thus allowing the neurons to learn several different views. It is not clear whether this is a sufficient explanation for the broad generalization shown by some of the face cells. One indication that something more complicated may be happening is the fact that many shape-selective neurons in the IT cortex of an awake monkey can be activated in the absence of their preferred visual stimuli while the animal is performing a short-term visual memory task (Miyashita & Chang, 1988). It is conceivable that if some cells are activated when the observer sees one view of an object, and a short time later some other cells are activated when the observer imagines the same object from a familiar viewpoint, then a C-unit could associate these two inputs, thereby acquiring a viewpoint-invariant response.

5.3 The Gestalt Grouping Effects

In a classic paper, Wertheimer (1923) argued that perception of a pattern involves the organization of its components into larger units, and he attempted to discover the underlying principles of this organizing process by studying the perceived grouping of simple patterns like those in Figure 5.5. This effort was the inspiration for the Gestalt school, with which the phrase ‘The whole is greater than the sum of its parts’ became identified (e.g., Koffka, 1935; Köhler, 1947). Although the Gestalt psychologists concerned themselves with many branches of psychology, and even attempted to relate their ideas to a hypothetical neurophysiology, they are now remembered mainly for the many visual phenomena that they discovered (reviewed by Pomerantz & Kubovy, 1986). Gestalt psychologists formulated many laws in order to explain the perceived groupings of simple patterns like those studied by Wertheimer, but in my view just three laws are adequate to explain most of the grouping phenomena. The law of proximity states that, ceteris paribus, elements that are close to one another
will be grouped together, as in Figure 5.5B. The law of similarity states that, ceteris paribus, elements that are similar to one another will be grouped together, as in Figure 5.5C. This is actually a class of laws rather than a single law, because elements can be similar on several different dimensions, such as size, colour, binocular disparity, and velocity. (Similarity of velocity is what Wertheimer called ‘common fate’.) The law of good continuation states that, ceteris paribus, smooth curves are perceived in preference to angular ones. In the examples of Figure 5.5D, each meeting of lines is interpreted as a crossing of two smooth curves rather than two angled lines that just happen to touch at their angles. The ceteris paribus condition means that these laws only make definite predictions about perceived grouping in situations sufficiently simple that there is no conflict between them; there is no good predictive theory of grouping in complex displays. Other Gestalt grouping laws have been proposed, such as that convex forms are perceived in preference to concave ones and symmetrical forms in preference to irregular ones, but these effects are found to be weak at best when they are not confounded with good continuation (Kanizsa, 1979; Pomerantz & Kubovy, 1986).

Two aspects of the present theory are particularly important in explaining grouping. First, a neuron encoding a luminance contour at a particular scale will respond not only to a straight ‘edge’ or ‘line’ contour but also to a line of ‘blobs’ lying in its RF at appropriate locations. This is assumed to be the basic cause of the grouping of dots into line segments in dot patterns like those of Figure 5.5. Brookes and Stevens (1991) discuss the plausibility of this proposal. Second, image features are represented at several different sizes, as explained in the discussion of spatial frequency (p. 121). A small pattern with an elaborate structure is no more than a blob when analysed at a low spatial frequency. Thus, a grouping process that operates on the outputs of neurons tuned to low spatial frequencies will group small patterns together without regard to the internal structure of those patterns, as in Figure 5.5A.

Using these two principles, together with some reasonable assumptions about the patterns represented by the neurons involved, a qualitative explanation of the grouping of elements into lines in simple patterns like those of Figure 5.5 can be given. The basic claim is that the Gestalt grouping laws are not implemented in the cortical architecture in any straightforward way, but that they are emergent properties of the interactions of large numbers of cells. The Gestalt laws can be regarded as

Figure 5.5: (A) Different types of element can be grouped to form a larger-scale structure. (B) Elements that are close together are grouped. (C) Similar elements are grouped. (D) The perceived grouping is one in which smooth contours are preferred. (B–D adapted from Wertheimer, 1923.)
excellent examples of abstract and approximate descriptions of an underlying physiological process which is much more complicated. Consider first grouping on the basis of proximity, as in Figure 5.5B. At the early processing stage, neurons encoding short contour segments are activated by aligned elements. The more closely spaced the elements are, the more elements there are in a given length of contour; and the more elements lie within a neuron’s RF, the more strongly the neuron is activated. Hence, in configurations like those of Figure 5.5B, the dots activate neurons encoding many different orientations, but the neurons activated most strongly are those encoding the orientation at which the dots are most closely-spaced. These neurons activate higher-level neurons encoding longer contour segments, and these neurons in turn send feedback connections to the lower-level ones, enhancing their activity. Inhibitory interaction reduces the activity of neurons encoding other orientations.

The explanation of good continuation, the tendency to group elements into smooth lines rather than angular ones, is a bit more involved. Consider the simplest case of good continuation, the X shape in Figure 5.6A. This is interpreted as two straight contour segments, although it could, for example, be interpreted as two V shapes that happen to touch at their points (Fig. 5.6B). Suppose that the four patterns in Figure 5.6B are encoded as shape fragments, so that the problem is to explain why the X pattern activates the two straight lines rather than the two V shapes. These fragments are built up from short contour segments in the early processing stage in the manner shown in Figure 5.6C, where the small dots indicate end-stopped cells. Note that the V-shaped fragments receive input from end-stopped cells at the corner of the V as well as at the line ends, whereas the only end-stopped cells activating the straight shape fragments are at the line ends. The X-shaped pattern activates end-stopped neurons only at the four line ends, as shown in Figure 5.6D. The end-stopped cells that feed into the V-shaped fragments at the angle of the V are not activated by the X-shaped pattern, and so the straight shape fragments match this pattern better than do the V-shaped fragments. This, I suggest, is the basic explanation for the law of good continuation. As always, feedback from the shape fragments to the lower-level contour segments and inhibitory competition at each level enhances the effect.

The third major grouping phenomenon is grouping by similarity. First, it is important to distinguish two cases, that of grouping elements into lines (as in Fig. 5.5C) and that of grouping two-dimensional collections of elements into regions. A region filled with similar elements is a texture, and the formation of contours between regions of different texture has already been considered (p. 126). The other case, that of grouping similar elements into lines, is rather difficult to account for. My own introspection is that the groupings in Figure 5.5C are less salient than those caused by proximity in Figure 5.5B, and it is even less salient in displays analogous to Figure 5.5C in which similarity is defined by the shape of the elements. A line of elements is, of course, a thin two-dimensional region which could be considered a texture, and so perhaps the formation of weak texture boundaries between lines of similar elements contributes to this grouping effect. Another factor that might contribute is a tendency to see a group of similar elements as the ‘figure’ and the other elements as the ‘ground’; a full discussion of this effect is given later in the chapter.
The reader might complain that the account of grouping given here is no more successful as a predictive theory than the classical Gestalt laws, and is far more complicated to boot. This is quite true, because the proposal is that grouping phenomena emerge from the excitatory and inhibitory interactions of very large numbers of neurons; moreover, because the tuning properties of the neurons are not ‘hard-wired’ but are subject to modification by experience, it is not generally possible to predict with certainty what organization will emerge for a complex pattern. However, the present theory does have the virtue of generality: the mechanisms underlying grouping are proposed to be the same as the mechanisms underlying other visual phenomena. Furthermore, there is no reason why there should necessarily be an elegant, self-contained theory of grouping effects in simple artificial patterns like those studied by the Gestalt psychologists. We interpret the structure of such patterns by applying a shape-perception system that is meant to deal with the much more rich and complex patterns in natural scenes.

### 5.4 Illusory Contours and Interpolation

We have already seen in the discussion of texture (p. 126) that it is not necessary for two regions to differ in mean luminance or colour for a contour to be perceived between them. A difference in the form of their textures is also sufficient, provided the textures differ in certain ways, as is a difference in the direction of motion or the binocular disparity of their textures. Contours that are clearly perceived in the absence of a luminance or colour change have been called illusory, subjective, anomalous, or cognitive contours. None of these names is entirely satisfactory, but I will use the term *illusory contour* here. There are other situations besides texture discontinuities in which illusory contours are perceived, such as a discontinuity in a grating pattern and, in some circumstances, a gap between two luminance contours (Schumann, 1904; Kanizsa, 1979). Some examples are illustrated in Figure 5.7. Note that the Kanizsa triangle (Fig. 5.7B) appears lighter than the background, although its luminance is the same. If the contrast of the picture is reversed, the black triangle appears darker than the black background. This lightness difference is probably caused by the mean luminance contrast across the outline of the triangle (Frisby & Clatworthy, 1975); the relation between lightness and contrast was discussed in the previous chapter (p. 131). This is not the cause of the illusory contours, as pictures can be devised in which the mean contrast across the outline of the triangle is zero, and no lightness difference is perceived, but the illusory contours are still clearly present (Prazdny, 1983; Kellman & Shipley, 1991).

Our understanding of these types of illusory contour has been advanced considerably by the neurophysiological studies of von der Heydt, Peterhans and Baumgartner (1984; von der Heydt & Peterhans, 1989; Peterhans & von der Heydt, 1989). Using the stimulus patterns shown in Figure 5.8

---

Figure 5.7: Illusory contours. (From G. Kanizsa, 1979, *Organization in Vision: Essays on Gestalt Perception*, New York: Praeger Publishers, Figs. 12.1a, 12.12.)

---
(A & B), they found that about one third of their sample of neurons in macaque area V2 that responded to an oriented luminance contour also responded to some degree in situations that cause humans to perceive an illusory contour. For each cell, the preferred orientation of the illusory contour was about the same as that of the luminance contour, but could be a little different, suggesting that the two kinds of response are produced in different ways. Neurons that responded to grating discontinuities (Fig. 5.8A) also usually responded to contours bridging gaps (Fig. 5.8B). A number of control experiments ruled out various simple explanations for the effect. In particular, it was shown that the responses to contours bridging gaps (Fig. 5.8B) could not be due merely to partial activation of the neuron by the two luminance contours on either side. This is because addition of thin lines to the pattern as shown in Figure 5.8C, which abolishes the illusory contours in human perception, also greatly reduced the neuron response, although it causes only a small difference in the luminance of the pattern. Neurons responding to illusory contours are also present in cat visual cortex (Redies, Crook & Creutzfeldt, 1986). They have complex RFs when studied with ordinary luminance contours (Redies et al., 1986).

Peterhans, von der Heydt and Baumgartner (1986) presented a theory which, they argued, could account for their observations. The neuron in their model receives inputs from two groups of precursor neurons, one which causes the response to an ordinary luminance contour, the other responsible for producing the illusory contour response. The cells in the former group have RFs oriented along the preferred orientation of the neuron. The cells in the latter group have end-stopped RFs which are oriented roughly orthogonally to the neuron’s preferred orientation, as shown in Figure 5.8D. A grating with a discontinuity at the correct orientation activates some of the cells with end-stopped RFs (each responding to one of the lines of the grating) and the summed outputs of these cells is sufficient to activate the neuron. The response to the other kind of illusory contour is explained by the pattern on either side of the gap activating some of the end-stopped RFs in the manner of Figure 5.8F. The addition of the thin lines to the pattern in Figure 5.8C is assumed to cause a large reduction in the activation of the end-stopped cells, which have comparatively small RFs.

Figure 5.8: (A–C) Stimuli used by von der Heydt, Peterhans and Baumgartner to study illusory contour responses of neurons in macaque V2. (D) Model of the receptive field construction of cells responsive to illusory contours. (E) The optimal stimulus for a cell with this receptive field. (F) Stimulation of the cell by a pattern like that in B. (Parts A–C adapted from R. von der Heydt, E. Peterhans and G. Baumgartner, 1984, Illusory contours and cortical neuron responses, Science 224: 1260–1262, Fig. 2. Copyright 1984 American Association for the Advancement of Science. Reproduced with permission.)
This model can be directly translated into the terms of the present theory by the assumption that the neuron in question receives its feedforward excitatory connections from lower-order cells with RFs arranged as in Figure 5.8D. Need this connectivity be genetically programmed, or could it be acquired from experience in accordance with the Empiricist Principle? This depends on the frequency with which patterns of the type that is optimal for this RF organization, shown in Figure 5.8E, occur in natural images. T-shaped contour junctions are, in fact, very common, occurring wherever an edge of one object passes behind an occluding edge of another. Presumably, therefore, it is also fairly common for a number of T-junctions to occur along one edge, as in Figure 5.8E. There are two difficulties with this explanation. First, if the main luminance contour is caused by an occluding edge of an object, one would expect the other contours to occur mainly on the side of the main contour corresponding to the background, and not on the side corresponding to the object’s surface. That is, the T-junctions should be mainly to one side, not on both sides as they are in Figure 5.8E. Actually, the presence of end-stopped RFs on both sides of the main contour does not seem to be crucial for explaining the results of the above experiments; perhaps more detailed investigations of these neurons will reveal that most of the end-stopped RFs are in fact on just one side of the main contour in most of these cells. The second difficulty is that one would not expect the T-junctions to be right-angled in general, since an occluding edge can intersect a background edge at any angle. One would expect to find cells in which the orientations of the end-stopped fields was oblique to the orientation of the main contour. In fact, von der Heydt and Peterhans (1989) found some evidence for such fields. It would also explain why several cells in their study had preferred orientations for illusory contours significantly different from that for luminance contours.

An interesting prediction of this account is that the end-stopped cells should encode different binocular disparities, corresponding to greater depth, from the cells signalling the main contour. This is because the main contour is assumed to usually result from an occluding edge, which is a discontinuity in depth. This would explain why illusory contours are often perceived as depth discontinuities (Coren, 1972). The Kanizsa triangle, for example, is seen in a nearer depth plane than the black discs.

The perception of contours straddling gaps in Figures 5.7B and 5.8B is an example of contour interpolation. There is good reason to believe that there is more to interpolation than the generation of illusory contours, however. Consider the modified Kanizsa triangle in Figure 5.9A. The account given above explains why the addition of the thin lines prevents the perception of illusory contours. However, it does not explain why a triangle is still clearly perceived, as if it is now being seen through three holes rather than in front of three discs. More simply, a triangle is seen in Figure 5.9B, again with no illusory contours. This demonstrates an interpolation process in shape perception that is more general than the process underlying illusory contours. Such a process is essential, because an edge of an object may not cause a luminance contour in the image along its whole length. The illumination and reflectance conditions may result in equal luminance on both sides of the edge in some places,
Figure 5.10: We have no difficulty in interpreting these pictures as three-dimensional forms, although there are few depth cues.

Figure 5.11: Impossible objects. The strong three-dimensional impression given by these pictures demonstrates the visual system’s ability to form representations of localized fragments of shape, even when these representations are globally inconsistent. (Part A adapted from L. S. Penrose and R. Penrose, 1958, Impossible objects: a special type of visual illusion, *British Journal of Psychology*, 49: 31–33, Fig. 1. Part B from D. H. Schuster, 1964, A new ambiguous figure: a three-stick clevis, *American Journal of Psychology*, 77: 673.)
and some parts of the edge may be obscured by other objects in front of it. I propose that the general explanation of interpolation is that a neuron encoding a shape fragment can be activated reasonably well by just some of its feedforward excitatory inputs. That is, not all of the lower-level cells feeding into the neuron need be firing strongly to activate it. This explanation of interpolation is just a generalization of the explanation given above for the generation of illusory contours; Kellman and Shipley (1991) also argue for the identity of these processes on the grounds that they have a number of common properties. The difference between the perception of illusory contours and interpolation without illusory contours is merely the level in the visual hierarchy at which the relevant neurons are located: illusory contours are seen only if relatively low-level neurons (encoding contour segments) are activated.

5.5 Perception of Three-Dimensional Shape in Pictures

Humans are good at perceiving three-dimensional shape in two-dimensional pictures. It is important to provide an account of this ability, not only because of the importance of pictures in our culture but also because a great deal of research on shape perception has used pictorial stimuli. The recovery of 3-D structure from pictures is not trivial because, unlike the 2-D images on the retina, binocular disparity does not provide any information on the 3-D orientations of the depicted edges and surfaces. Some pictures do, of course, contain other kinds of information which could be used to estimate 3-D surface orientation, including perspective, shading, and gradients in the size of texture elements (Gibson, 1950), but even for the impoverished outline drawings in Figure 5.10 the visual system can recover 3-D shape remarkably well.

The general explanation proposed for this effect is that shape fragments that incorporate edge and surface fragments with particular 3-D orientations can be activated by 2-D pictures. Note that, because most of our visual experience is with real objects rather than pictures, the Empiricist Principle leads us to expect that the great majority of shape fragments are 3-D rather than 2-D. Shape fragments are 3-D because many of the features of which they are composed (both contour segments and texture patches) are selective for binocular disparity (Fig. 5.1). If a picture is viewed in the frontoparallel plane, each line segment will maximally activate neurons tuned to zero disparity (approximately). However, because neurons are broadly tuned, neurons with non-zero preferred disparities will also be activated to various degrees. Suppose that several such partially-activated cells, at different locations in the visual field, feed into a neuron at a higher level which encodes a 3-D shape fragment. Because, by assumption, most shape fragments are 3-D rather than 2-D, this particular shape fragment could well be the one that best matches the retinal image. Hence, the picture is interpreted as 3-D.

The strength of the visual system’s tendency to impose a 3-D interpretation on 2-D pictures is demonstrated by drawings of so-called impossible objects. Some examples due to Penrose and Penrose (1958) and Schuster (1964) are shown in Figure 5.11. More elaborate examples can be found in the work of the artist M. C. Escher. Local regions of these pictures can be given a 3-D interpretation, but the local interpretations cannot be made consistent with one another. The strong impression of three-dimensional structure given by these pictures, in spite of global inconsistency, is explained by the existence of shape fragments which integrate shape information within relatively localized parts of the image.

Although the 3-D effect of pictures is definitely real, it is obviously less vivid than that of looking at real objects. We seem to be simultaneously aware that the picture is 2-D in one sense and 3-D in another sense. This may be because of a conflict between the information delivered by early and late stages of analysis: the activation of 3-D shape fragments indicates depth, but this is contradicted by the lower-level analysis of binocular disparity. This reinforces the plausible assumption that central cognition has access to information represented at many levels of the visual system, not just the highest levels.
5.6 Multistability and Hysteresis

One of the basic principles of the present theory is that visual perception is mainly a feedforward, ‘bottom-up’ process, and the account of shape perception is in accordance with this principle. The excitatory connections that transmit information from lower-level to higher-level neurons provide the main driving force of those neurons. These connections are the most important factor in determining the pattern of activity in the visual cortex that represents the analysis of the viewed scene. However, the theory also has room for feedback or ‘top-down’ activation from higher-level to lower-level neurons, as well as mutual interactions between neurons at about the same level via both the inhibitory and excitatory connections of the local circuitry and the longer lateral excitatory connections. Although the function of these mechanisms is subordinate to that of the feedforward connections, they nevertheless have an important influence on shape perception. In this section I consider two phenomena that, I propose, require these mechanisms for their explanation.

First consider the influence of inhibition. Recall that inhibitory neurons have strictly localized axonal distributions within each cortical area. In early processing and the early stages of shape perception, neurons have rather small RFs, and the areas involved are retinotopically mapped, and so inhibitory interactions will occur between neurons with RFs at the same location in the visual field. This means that there is a kind of competition between the cells encoding features or low-level shape fragments at each location, such that if one feature or shape fragment is strongly active, very different features or shape fragments at the same location tend not to be active. This will only be the case, however, within each of the cortical areas in question—there are no inhibitory connections between areas. Insofar as different types of shape fragment are segregated in different areas, this means that competition does not occur between different types. Unfortunately, the neurophysiological evidence is not sufficient at present to make more precise claims about which types of shape fragment inhibit one another.

Competition between features and low-level shape fragments at each location is proposed as the reason why we only perceive one organization of the image and are not aware of other possible organizations. Usually, the rich information available in a natural image enables the visual system to arrive at a single interpretation of it. Furthermore, this interpretation is normally the correct one—although it is possible to fool the visual system by contriving scenes for which the obvious interpretation does not correspond to the real situation, the Ames distorted room being a well-known example. However, if the information in the image is impoverished, it is sometimes possible for it to be organized in two or more different ways that are equally plausible. In this case, all the possible organizations can be perceived, but only one at a time. This phenomenon is called multistability (Attneave, 1971). Multistability is usually demonstrated by ambiguous pictures such as those in Figure 5.12. The most thoroughly-studied ambiguous picture is the Necker cube (Fig. 5.12A). Most observers see this picture as a wire-frame cube. There are two different ways in which this interpretation can be made, however, and under prolonged viewing the organization is seen to flip spontaneously from one interpretation to the other every few seconds.

Figure 5.12: The Necker cube and the Schröder staircase.
Multistability is explained as follows. A two-dimensional picture is proposed to activate neurons that encode three-dimensional shape fragments, as explained in the previous section. In the case of an ambiguous picture, the basic feedforward process of neural activation caused by the presentation of the picture can potentially result in the activation of two or more significantly different sets of shape fragments, with about equal probability. The reason that only one of these sets is active at any one time is that the different sets place different 3-D interpretations on at least some of the 2-D contours of the image. For example, the two interpretations of the Necker cube picture impose different 3-D orientations on the four edges that are diagonal in the picture. The different 3-D orientations are encoded by different neurons. As stated above, inhibitory interactions are assumed to occur between neurons encoding different features and low-level shape fragments at the same location in the visual field. Hence, there is a competition between the different 3-D interpretations of at least some parts of the 2-D picture, and so only one interpretation can be active at any time.

The reason for the spontaneous alternation between the different organizations may be adaptation of the firing rates of the inhibitory neurons involved (Attneave, 1971). By assumption, the feedforward excitatory input to the competing neurons is about equal, so the activity of the neurons involved in the non-dominant interpretation of the picture must be suppressed by inhibitory input from some of the neurons involved in the dominant interpretation. If these inhibitory neurons suffer a reduction in firing rate over a period of seconds, the neurons encoding the non-dominant interpretation become progressively more active. Eventually, a point is reached at which the suppressed inhibitory neurons become sufficiently active to win the competition, and a reversal occurs. Adaptation has not been incorporated into the model neuron of the Magpie theory, but it would not be hard to do so.

Inhibition and adaptation are not the whole story behind multistability. Feedback and lateral excitatory connections are also important. If neurons encoding shape fragments were activated in a purely feedforward manner, there would be no way of ensuring that the fragments activated by different parts of an ambiguous picture were mutually consistent. Perception of an ambiguous picture would then be like perception of the impossible objects in Figure 5.11—different regions of the image would have inconsistent 3-D interpretations. (This would actually be the case not just for ambiguous pictures but for any image that is locally ambiguous.) In fact, each interpretation of an ambiguous picture is a valid 3-D shape. I suggest that a cooperative effect realized by the feedback and lateral connections between neurons encoding shape fragments is responsible for this. These connections are assumed to interlink neurons that are frequently simultaneously active. Activated neurons that have often been co-active in the past cooperate to enhance each other’s activity levels. If two competing neurons are activated to about the same degree by their feedforward inputs, as can happen in perception of an ambiguous picture, the neuron that receives contextual enhancement from other strongly-active cells will win the competition over the neuron that does not receive such enhancement. This, I suggest, is why ambiguous pictures are interpreted in a globally-consistent manner, and also why a local ambiguity in an image is not recognized as such but is interpreted consistently with the surrounding context. Recall that intrinsic lateral connections, like inhibitory connections, have a limited radial extent within each cortical area, although they do extend further than the inhibitory connections. Therefore, in early processing and the early stages of shape perception, at which cortical areas are retinotopically mapped and RF sizes are small, lateral connections tend to link neurons whose RFs are at near but not identical locations in the visual field.

Cooperation between neurons encoding shape fragments has other important effects in shape perception. Wertheimer (1923) noted that if the dot configuration in Figure 5.13A was slowly transformed into that in Figure 5.13B, the original grouping perceived on the basis of proximity persisted up to the point at which all the dots were equally-spaced, and even persisted beyond that point, so that the distance between dots within the perceived groups was actually greater than the distance between the groups. Eventually, the configuration suddenly flipped into a new grouping, based on the proximity of dots in Figure 5.13B. In the reverse procedure, when the constellation in Figure 5.13B was transformed into that in Figure 5.13A, the original grouping again persisted beyond the halfway point. Thus, when the dots were about equally-spaced, the perceived grouping depended on the history of dot movement leading up to that configuration. This is an example of hysteresis, the persistence of an effect after changes in the conditions that caused it. Much more recently, Chang and
Julesz (1984) and Williams, Phillips and Sekuler (1986) have demonstrated analogous effects in the perceived direction of motion of fields of moving random dots. There is also an interesting hysteresis effect in stereopsis which we will return to in the next section. Several authors have taken hysteresis as evidence of cooperative interactions between neurons (e.g., Wilson & Cowan, 1973), since there are cases in physics of hysteresis effects resulting from local cooperative interactions, the magnetization of iron being one example. Specifically, it may be the feedback and lateral connections between neurons encoding shape fragments that are responsible for hysteresis in grouping. An ensemble of strongly-active neurons that are interlinked by these connections can form a stable coalition (Feldman & Ballard, 1982), enhancing each others’ activity. Such a coalition of neurons can remain active, and continue to inhibit competing cells, even after the feedforward input to the component neurons that originally caused the coalition to form is reduced in strength.

5.6.1 A Model of Necker Cube Multistability

Because the ideas in this section may be less intuitively clear than the basically feedforward processes we have been considering earlier in the chapter, I will present a simple model of the neural interactions proposed to underlie multistability of the Necker cube (Fig. 5.12A). The Necker cube is named after Louis Necker who first described it in 1832. It has been the subject of a number of experimental studies (reviewed by Rock, 1975; Pomerantz & Kubovy, 1986), and has also been used to demonstrate connectionist models of cooperative and competitive processes in vision (Feldman, 1981; Rumelhart, Smolensky, McClelland & Hinton, 1986). The model presented here is based on the model neuron and general connectivity constraints assumed in the Magpie theory. The inputs to the model are the firing rates of 144 feature units, which may be thought of as depth-tuned complex cells, arranged at 48 locations in the pattern of Figure 5.14A. There are three units at each location, called the front, middle and back units, all tuned to the same orientation but with different sensitivities to binocular disparity. In accordance with the behaviour of tuned-excitatory neurons described in Chapter 4 (p. 124), the three units all respond to an edge appearing at that location, but to different degrees depending on the distance of the edge. Specifically, an edge at about the fixation distance is assumed to activate the middle unit with a firing rate of 1, and the front and back units with firing rates of 0.5. A ‘near’ edge activates the front unit with firing rate 1, the middle unit with firing rate 0.5, and the back unit with firing rate 0.2, and a ‘far’ edge activates the back unit with firing rate 1, the middle unit with firing rate 0.5, and the front unit with firing rate 0.2. I ignore the complication that neurons tuned to horizontally oriented contours cannot very well be depth-tuned.

The feature units send excitatory connections into a network of sixteen shape-fragment units. There are two such units at each of eight locations, corresponding to the eight corners of the cube. Each shape-fragment unit receives connections from six of the feature units. At each of the eight locations, one shape-fragment unit receives connections from five back units and one middle unit, such that it is maximally excited by a ‘far’ corner of a 3-D wireframe cube, and the other shape-fragment unit receives connections from five front units and one middle unit, such that it is maximally excited by a ‘near’ corner of a wireframe cube. The patterns represented by the shape-fragment units are illustrated in Figure 5.15.

The shape-fragment units are interconnected in a manner consistent with the assumptions of the Magpie theory, albeit in a highly simplified form. Each unit contributes just three connections to the
local circuitry: it sends a short excitatory connection to itself, and inhibitory connections both to itself and to the other unit at that location. As in the model described in Chapter 4, the units should be considered to represent clusters of neurons with similar characteristic patterns, so that the self-connections of the units represent connections within the cluster. In addition, each unit sends long-range lateral excitatory connections to three units at other locations, corresponding to the three adjacent corners of the wireframe cube. This connectivity is illustrated in Figure 5.14B.

In accordance with the model neuron described in Chapter 3, each shape-fragment unit has a membrane potential $V_j$ and firing rate $f_j$ governed by the equations:

$$\frac{C}{\partial t} V_j = (E_r - V_j) g_r + (E_e - V_j) g_e \sum_{u_k \in F_j} f_k + (E_e - V_j) g_s f_j$$
$$+ (E_i - V_j) g_i \sum_{l_k \in I_j} f_k + (E_e - V_j) g_l \sum_{l_k \in L_j} f_k \tag{5.1}$$

$$f_j = \tanh^+ \left( \frac{V_j - \rho}{\gamma} \right) \tag{5.2}$$

where $F_j$, $I_j$ and $L_j$ are the sets of units sending feedforward, inhibitory and lateral connections, respectively, to $u_j$. The free parameters of these equations are the conductances $g_e$, $g_s$, $g_i$ and $g_l$, associated respectively with the feedforward, short excitatory, inhibitory, and lateral connections; these were set by trial and error with the model to $g_e = 0.01 \mu S$, $g_s = 0.05 \mu S$, $g_i = 0.1 \mu S$, and $g_l = 0.005 \mu S$. Biologically these are high values, except for $g_l$, but this is a consequence of the tiny number of connections in this highly simplified model. All the other parameters were set to the same values as in the model of the previous chapter. For simplicity all the connections have weight 1; the results of the simulations are hardly affected if the lateral connections are given random weights from the range $[0, 1]$.

The dynamics of the network are not simulated. Instead the feature units are assigned constant firing rates, as described above, and a fixed-point iterative algorithm is used to compute the steady state of the shape-fragment network ($dV_j/dt = 0$ for all $j$). Figure 5.15 shows the steady states that result when the input to the model is a 3-D wireframe cube. Figure 5.15A shows the result when the

Figure 5.14: Structure of the Necker cube model. (A) Feature units are arranged at 48 locations. (B) Shape-fragment units send short excitatory connections to themselves (not shown), inhibitory connections to themselves and to the other unit at the same location (dashed lines), and lateral excitatory connections to units at three other locations (solid lines). Lines without arrowheads indicate reciprocal connections.
Figure 5.15: Steady states resulting from two views of a three-dimensional wireframe cube. Feature units are shown on the left. The white and black boxes under each shape fragment indicate the feedforward input and the firing rate of each shape-fragment unit. See text.

Figure 5.16: Two stable states resulting from a 2-D Necker cube stimulus.
lower-right face of the cube is nearest, and Figure 5.15B shows the result when the upper-left face is nearest. The pattern on the left of the figure illustrates the activity of the feature units: thick lines show locations where the front unit is most active, thin lines show locations where the back unit is most active, and lines of medium thickness show locations where the middle unit is most active. In the middle and right of the figure sixteen icons represent the shape fragments. The white box under each icon shows the magnitude of the feedforward input to each unit, and the black box shows the steady-state firing rate of each unit. The white boxes show that with these 3-D stimuli the input is unambiguous: at each location one of the two shape-fragment units is receiving much greater feedforward input than the other, and not surprisingly that unit is the one firing strongly.

A more interesting result is shown in Figure 5.16. This is a simulation of the classic Necker cube stimulus: the input pattern is two-dimensional, which means that the middle unit is the most active feature unit at all locations. The white boxes show that all sixteen shape-fragment units are receiving the same amount of feedforward excitation. Nevertheless, in Figure 5.16A only the eight shape fragments representing one 3-D interpretation of the pattern are active in the steady state, the other eight units being inhibited. In Figure 5.16B the other interpretation of the pattern is dominant. This is the outcome of the competitive and cooperative interactions between shape fragments described earlier. A state in which two units at the same location are both strongly active is highly unstable because of the strong inhibition between them. Each pair of units adopts a state in which one is firing strongly and the other is inhibited. But the lateral connections cause each dominant unit to help other units with which it is consistent to win their competitions as well. The network as a whole converges to a state in which one coalition of eight shape fragments is dominant, and there is a high probability that these eight fragments are mutually consistent. In the iterative algorithm used in the simulation the unit activities are updated asynchronously and in random order, so the model sometimes converges to one interpretation and sometimes the other; with the parameters listed above stable states that do not represent valid 3-D interpretations of the pattern occur on less than 5% of trials.

The importance of the lateral connections in obtaining a consistent interpretation of the Necker cube is illustrated in Figure 5.17A. Here the lateral connections have been disconnected (i.e., \( g_L = 0 \)). Although the local circuitry still ensures that only one shape-fragment unit at each location is active, there is no means to ensure that the network as a whole arrives at a globally consistent interpretation of the pattern. The lateral connections do not, however, need to be very powerful. Figure 5.17B shows the results with a Necker cube missing one corner. At six locations no feature units are active, so two of the shape-fragment units are receiving no feedforward input. Neither unit is firing in the steady state, in spite of the fact that one of them is being excited by three other units via the lateral connections. This demonstrates that the lateral connections can fulfil their purpose while being much weaker than the feedforward connections.

Figure 5.18 gives a further demonstration of the contextual enhancement exerted by the lateral connections. The stimulus is a Necker cube in which the feature units at a single location are providing unambiguous depth information: in Figure 5.18A one front unit is firing strongly, and in Figure 5.18B one back unit is firing strongly. This gives one shape-fragment unit a small advantage over its competitor, and this bias is sufficient to cause the favoured depth-organization to become dominant in the vast majority of trials. It would be interesting to perform an analogous experiment with human observers.

It is difficult to demonstrate hysteresis with this model because the dynamics of the network are not simulated, only its stable states. One observation worth mentioning is that the biasing effect illustrated in Figure 5.18 only occurs if the network is initially in an inactive state (all shape-fragment units at rest). If the network starts off in the state representing the unfavoured depth-organization, it remains in that state; the slightly greater feedforward excitation favouring the other organization is not sufficient to overthrow the initial stable coalition. (From a psychological perspective the Necker cube is not the best stimulus to investigate hysteresis because it is exceptionally liable to flip from one organization to the other.)

Clearly, the model is highly simplified; it is intended primarily to illustrate the concepts involved rather than as a serious theory of the Necker cube. An obvious development would be to simulate the network dynamics and to incorporate firing rate adaptation, so as to model some of the adaptation and rate-of-reversal studies in the literature (Rock, 1975, pp. 263–270).
Figure 5.17: (A) A typical stable state resulting from a Necker cube stimulus when lateral connections are disconnected. This does not represent a coherent 3-D interpretation of the pattern. (B) Demonstration that lateral connections do not need to be powerful. A shape-fragment unit lacking feedforward input remains silent in spite of excitation from other shape-fragment units.

Figure 5.18: Demonstration of how contextual enhancement can resolve local ambiguity in the stimulus. A small bias in the feedforward input to one shape-fragment unit is propagated over the lateral connections to determine the final organization.
5.7 Stereopsis

Stereopsis is the derivation of shape information from binocular disparity (reviewed by Poggio & Poggio, 1984). Human stereopsis has been extensively investigated using a particular kind of stimulus, the random dot stereogram (RDS) introduced by Julesz (1960, 1971). An example is shown in Figure 5.19. The two random dot patterns appear to contain no structure when viewed monocularly, but when the patterns are presented separately to the two eyes they ‘fuse’, yielding a vivid percept of shape. The two patterns are identical, except that a central shape is shifted a short distance to the left in one pattern relative to the other, the gap being filled by more random dots. Thus, when viewed stereoscopically, the dots in the central region have a different disparity from the dots in the surrounding region. Accordingly, the central region is perceived to lie in a different depth plane. In a sense, the RDS presents the visual system with a ‘worst case’ stereo stimulus: first, there is no shape information other than disparity, and second, the visual system must match each dot in the left image with the corresponding dot in the right image in order to compute disparity correctly. This matching task is difficult, because each dot considered individually could match any of numerous other dots in the other image. This is the correspondence problem of stereopsis. The human visual system solves the correspondence problem remarkably efficiently (Julesz, 1971). Many algorithms have been proposed for solving the correspondence problem for RDSs, none of which accounts for all the psychophysical data (reviewed by Blake & Wilson, 1991). In this section I outline how the present theory is applied to this particular problem. The basic premise is that there is nothing special about stereopsis—that is, there are no specialized stereo mechanisms as such. Instead, humans perceive stereograms using a general shape perception system.

The first point to note is that an RDS is an example of a texture, and is analysed as such in the early processing stage. As explained in the previous chapter, there are three stages in texture perception: first the texture elements are identified by neurons with small RFs; second, patches of texture are made explicit by neurons with larger RFs; third, discontinuities are detected between patches differing significantly in texture (p. 127). Simple cells have oriented RFs, and so are activated by clumps of dots that happen to be aligned in the RDS. As Barlow, Blakemore and Pettigrew (1967) point out, the use of oriented features rather than dots itself goes some way towards solving the correspondence problem, because a match can be made only between features that have a similar orientation in both eyes. Many of these simple cells are disparity-selective. Each neuron responds when a clump of dots of the appropriate orientation occurs in the neuron’s RF in both eyes, at the neuron’s preferred disparity. An observer viewing an RDS makes vergence eye movements, varying the relative position of the images on the two retinas—in effect, exploring a range of disparities until fusion is perceived. Many of the simple cells can be expected to signal false matches during this process, whenever non-corresponding clumps of dots happen to activate them. These cells feed into higher-order neurons that represent patches of texture of a particular disparity. Because each of these texture cells integrates information from several simple cells, strong activation is likely to occur only...
when a correct match is obtained between the left and right images. This is because a group of simple cells is unlikely to be activated simultaneously by false matches. Thus, the correspondence problem is almost entirely solved at this stage. Complex cells responding to random dot patterns at a preferred disparity have been reported in areas V1 and V2 of the macaque (Poggio, 1984; Poggio, Motter, Squatrito & Trotter, 1985). The third stage in the early processing of texture is the representation of texture discontinuities. In this case, the cells involved would respond to the disparity discontinuities between the various regions of the RDS. Once these contour segments have been made explicit they can be organized by the activation of shape fragments, as with any other kind of contour.

It was explained in the previous section that feedback and lateral connections between neurons that tend to be simultaneously active are responsible for a cooperative effect which can result in hysteresis phenomena. Such connections can be expected to interlink many of the cells encoding textures of the same disparity, since these would presumably be activated together by an extended textured surface that was not strongly slanted. Once a group of these texture cells was activated by an RDS, therefore, the cooperative effect would stabilize the group, causing hysteresis. Such an effect was found by Fender and Julesz (1967). To control for eye movements they used stabilized images (of sufficiently high contrast to prevent their disappearance). They found that the two patterns of an RDS had to be brought within 6° of disparity for fusion to occur, but once they had fused the pair could be separated by up to 2° before fusion broke, provided they were not pulled apart too quickly. A much smaller effect was obtained for simple line stimuli than for the RDS, presumably because much fewer neurons were involved.

The hysteresis effect in stereopsis is not merely a psychophysical curiosity but is of great functional importance in normal vision. During fixation, the eyes drift at a rate of about 1° per second, and these movements are not perfectly correlated in the two eyes (Fender & Julesz, 1967). This means that small disparity errors are constantly building up. The hysteresis allows stable fusion to be maintained in spite of these small movements. Also, after fusion has been achieved for one region in an RDS, the hysteresis enables fusion to be maintained for that region while further vergence eye movements are made to find the correct disparity for another region. Because hysteresis is evidence for cooperative interactions between neurons (Wilson & Cowan, 1973), Fender and Julesz’s discovery led several investigators to propose algorithms for solving the correspondence problem that employed such cooperative interactions (Julesz, 1971; Nelson, 1975; Dev, 1975; Julesz & Chang, 1976; Marr & Poggio, 1976; see Marr, 1982, for a critical review). Unlike these models, the present theory uses cooperativity mainly to preserve fusion after the correspondence problem has already been solved. Cooperativity does not play a major role in the matching process itself, which is basically feedforward. However, the contextual enhancement effect exerted by the feedback and lateral connections may well serve to help resolve ambiguities during matching. Evidence consistent with this was found by Julesz and Chang (1976). Using an ambiguous RDS, in which the central region consisted of a repeating ‘wallpaper’ pattern which could be fused at two different disparities, either nearer or further than the surrounding region, they found that setting just a few per cent of the dots to an unambiguous disparity was sufficient to determine the disparity at which fusion was perceived.

5.8 Figure and Ground

The distinction between figure and ground has long been considered to be of fundamental importance in visual perception. The classical description of the phenomenon is the monograph by Rubin (1915). According to him, the pattern that is perceived as the figure has a definite shape, a ‘thing-like’ character, whereas the ground, comprising the remainder of the visual field, is relatively formless. Compared with the ground, the figure dominates consciousness, is much better remembered, and is much more likely to recall associations from memory. Rubin illustrated these observations with the ambiguous pictures in Figure 5.20, which can be separated into figure and ground in two different ways. (Some modern authors use the term ‘figure-ground segregation’ to mean the process of locating contours in the image, but this is sloppy terminology. We seem to be aware of contours both in the figure and in the ground. I will adhere to Rubin’s phenomenological insights here.)

In my view, the distinction between figure and ground is simply the difference between the attended and unattended patterns in the image. The main instrument of selective attention in the
Magpie theory is the focusing mechanism. Recall that this mechanism can enhance the activation of neurons in selected parts of the visual cortex corresponding to spatially-restricted regions of the visual field. If the mechanism is used to focus on a particular pattern in the image, the analysis of that pattern can proceed to the highest levels of the visual cortex, including the activation of shape fragments and the process of object recognition. This accounts for the properties of the figure mentioned above. In the parts of the visual cortex not enhanced by the focusing mechanism, the activation of neurons is progressively more attenuated at successive levels of the hierarchy of areas. The unselected region of the visual field, therefore, does not receive such a full analysis. Neurons encoding features and simpler shape fragments are activated reasonably well, but the high-level neurons that represent larger and more elaborate shape fragments are only weakly activated by this part of the image. This explains the relative formlessness of the ground. The direction of the focusing mechanism is under the control of central cognition, and is independent of eye movements. This explains why the selection of which part of the image to see as figure is a voluntary one, and why one does not need to be looking directly at a pattern to see it as figure.

One aspect of figure-ground segregation that is difficult to explain, however, is that even if two patterns are spatially overlapping in the image it is still possible to see one as figure and the other as part of the ground, provided the patterns can be distinguished by some elementary feature, such as a difference in colour, depth, or velocity (e.g., Rock & Gutman, 1981). For example, given a pattern of scattered dots in which each dot has one of two binocular disparities it is possible to see the constellation of ‘near’ dots as figure and the intermingled ‘far’ dots as ground, or vice versa. (This ability may explain the Gestalt effect of grouping by similarity in sparse dot patterns, as suggested earlier.) Some recent experiments by Motter (1994a, 1994b) might demonstrate a neurophysiological correlate of this effect. He trained monkeys to select an item from a display based on its luminance or colour, the criterial luminance or colour being indicated to the monkey by the luminance or colour of a cue prior to the onset of the display. He found that most neurons in area V4 that responded to the items in the display were activated about twice as strongly when the item matched the luminance or colour of the cue than when the item did not match the cue. For example, a neuron responding to a red item did so much more strongly if the cue was red, although the cue was not in the neuron’s RF. The most interesting result was that this differential activation did not seem to be restricted to a particular location in the display: neurons that matched the cue were enhanced right across the visual field. Thus, if the cue was red, many neurons responding to red items were enhanced, whereas intermingled neurons responding to items of other colours were not.

The ability to attend to just one of two overlapping patterns presents a problem for the Magpie theory because it is assumed that the focusing mechanism can only select visual patterns on the basis of direction in the visual field, in the manner of a spotlight beam of adjustable width. It is hard to explain how overlapping patterns could be selected individually. One possibility is that the mechanism does not need to focus on an entire pattern in order for the whole pattern to be analysed to a high level (that is, to be perceived as figure). In the earlier discussion of hysteresis in grouping I suggested that the feedback and lateral connections between neurons could be responsible for a cooperative effect, sustaining a stable coalition of active neurons. Because these connections are
assumed to interlink neurons that are often active simultaneously, such a stable coalition would be most easily formed if the neurons encoded similar features, such as the same colour, disparity or velocity. If the focusing mechanism were used serially to focus on a succession of small parts of one pattern, with a sufficiently narrow focus to exclude parts of the overlapping pattern, a stable coalition of neurons representing the former pattern could be built up, and competitive interaction would suppress the neurons responding to the latter pattern. This is illustrated schematically in Figure 5.21. This explanation of the effect is hardly an elegant one, but the situation is, after all, rarely encountered outside the psychology laboratory. In natural scenes objects are usually opaque and spatially continuous. Two aspects of Motter’s (1994a, 1994b) experiments are pertinent. First, the differential response of neurons matching the cue over non-matching neurons did not begin until 150–200 msec after the display onset, and did not reach full strength until about 500 msec; this is much longer than the latency of the neurons to visual stimulation (about 50 msec) and suggests a relatively slow organizing process. Second, the monkeys were given a considerable amount of practice at the task before the experiments began.

Two further aspects of the figure-ground phenomenon deserve mention. First, there is a tendency for patterns to appear lighter when they are seen as figure than when they are part of the ground (Coren, 1969). This is to be expected, because lightness is related to contrast (p. 131) which is signalled by the absolute firing rates of some neurons in early processing. Because focusing increases the firing rates of the affected neurons, they effectively signal greater contrast than otherwise. Second, Rubin (1915) was much impressed by the tendency for the outline of the figure to be perceived as belonging only to the figure, not to the ground (see Fig. 5.20). The ground is instead perceived as continuing behind the figure. This probably indicates that some of the neurons activated by this contour encode depth discontinuities. This is supported by Julesz’s (1971) finding that when Figure 5.20B is presented in the form of a stereogram, so that the depth relation of the two regions is unambiguous, the picture can be seen only as a vase and not as a pair of faces.

The Gestalt psychologists formulated a number of laws predicting what would be seen as figure in simple drawings, such as an enclosed area being seen as figure more often than the surrounding area (Pomerantz & Kubovy, 1986). These should probably be regarded as heuristics rather than laws, because the selection of the figure is clearly under voluntary control. Perhaps they indicate biases on the part of the central cognitive processes that control the focusing mechanism.

Figure 5.21: Using the focusing mechanism to help establish a stable coalition of strongly-active neurons, perhaps explaining how overlapping patterns can be segregated as figure and ground on the basis of a colour, depth or motion difference. (A) Red and green patterns activate separate but intermingled clusters of neurons (open and filled squares). (B) Focusing (circle) on a part of the red pattern enhances some of its neurons (larger square). These neurons help to activate other neurons responding to the red pattern via lateral excitatory connections (arrows). (C) Moving the focus around the red pattern builds up a stable coalition of neurons responding to that pattern; neurons responding to the green pattern are inhibited.
Visual search is the process of finding an item that meets some criterion in a visual display containing many different items, such as finding a familiar face in a school photograph. Psychologists have studied visual search to try to elucidate the characteristics of visual selective attention. In a typical visual search experiment the subject is presented with a visual display containing a number of simple items, as in Figure 5.22, and is required to decide as quickly as possible whether one of the items has some distinguishing characteristic. The item to be found, if present, is called the target, and the other items are called distractors; the total number of items is the display size. The subject responds by pressing one of two buttons to indicate whether a target is present or absent, and the time taken between the onset of the display and the button-press is recorded. The data of interest is how reaction time varies as a function of the display size and of the distinguishing characteristics of the targets. The basic finding is that if the target possesses a feature which all the distractors lack then reaction time is roughly constant, independent of display size. For example, a vertical line segment among horizontal distractors, or a large item among small distractors, or a brown item among green distractors, are all detected easily, regardless of how many distractors there are. Phenomenologically, the target appears to ‘pop out’ of the display (Fig. 5.22A). On the other hand, if the target is defined by a conjunction of features, such that it shares at least one feature with every distractor, then reaction time generally increases steeply with display size. For example, a green-and-vertical target among green-and-diagonal and brown-and-vertical distractors does not pop out. Reaction time is a linear function of display size, and the slope of the function for positive responses, when the target is present, is about half that of the function for negative responses, when the target is absent (Treisman & Gelade, 1980). This suggests that the subject is performing a serial, item-by-item search of the display. The 2:1 slope ratio suggests that the subject terminates the search when the target is found, since, on average, only half the items need to be searched when the target is present.

Anne Treisman and her collaborators have developed a theory of visual processing to account for these and other experimental results, called feature integration theory (FIT) (reviewed by Treisman, 1988). As is usually the case in cognitive psychology, the original proposal of this theory (Treisman, Sykes & Gelade, 1977) and the striking results of the first supporting experiments (Treisman & Gelade, 1980; Treisman & Schmidt, 1982) prompted a deluge of empirical investigation, which showed the phenomena to be much more complex than was at first thought. FIT has had to be made more complicated to accommodate these results (Treisman & Sato, 1990), while critics are still able to question almost every aspect of the theory (e.g., Green, 1991). In this section I try to explain the main empirical phenomena of this field in terms of the Magpie theory, and argue that it provides a better account than FIT in that it is more consistent with the known physiology of the visual cortex. However, I make no attempt to address all the findings in the mountainous feature integration literature, and further research will be required to show that the present theory gives a better quantitative account of the data. (For other alternatives to FIT see Duncan & Humphreys, 1989; Wolfe, Cave & Franzel, 1989; Grossberg, Mingolla & Ross, 1994.)
I will not describe the details of FIT here (see Treisman, 1988), but I will summarize its main assumptions. Following Neisser (1967), visual perception is proposed to consist of two main processing stages. In the first stage, the retinal image is processed to make explicit several types of elementary feature. This stage exploits a massive amount of parallel processing, and is automatic in the sense that it is not affected by attention. In the second stage, features are conjoined to form temporary representations of the visible objects. It is proposed that attention is generally required for this process. In particular, attention is required to accurately locate the positions of features so that they can be correctly joined together. Like a zoom lens, attention can be distributed over a broad region of the image, or it can be narrowly-focused on a small part of it. Within the selected region, features are combined into object representations. Outside the selected region, features are not accurately localized (they are ‘free floating’ with respect to position) and, if they are conjoined into object representations at all, there is a strong possibility that the representations are inaccurate, containing a mixture of features that actually belong to different objects.

FIT accounts for the visual search results as follows. When the target is identified by a feature which all the distractors lack, the subject can note the presence of the feature simply from the activity of any feature detector unit of the appropriate type. The theory proposes that all feature detectors that respond to a particular feature are grouped together, so that the subject can note the presence of the feature just on the basis of significant activity in that group. This test does not require the use of selective attention, and is independent of the display size, since the distractors do not affect that group of feature detectors. If, however, the target is defined by a conjunction of features, this simple test cannot be used, because all the relevant groups of feature detectors are activated both by the target and by some of the distractors. It is assumed that there are no ‘conjunction detectors’, units that respond only to particular conjunctions of features. This necessitates a serial, item-by-item search for the target using selective attention, because attention is required to accurately conjoin the features of each item. If attention were not applied to each item individually, it would be possible for the ‘free floating’ features from two distractors to conjoin to give a representation of a target, causing the subject to identify a target that was not actually present.

There are some additional lines of evidence that support this theory. First, Treisman and Gelade (1980) found that in feature search, subjects could sometimes identify the presence of a target without being able to report where it was in the display. This supports the proposal that features are detected preattentively without their locations being accurately registered. In conjunction search, however, if subjects correctly identified a target then they were always able to report its location, consistent with the view that attention serves to accurately localize features. However, Green (1991) points out that this particular finding has not been replicated. Second, if attention is overloaded, illusory conjunctions of features can occur (Treisman & Schmidt, 1982; Cohen & Ivry, 1989). Subjects who are required to identify items in a briefly-presented display often report items that are not present, but which appear to result from a combination of features from other items. For example, if the display contains a pink triangle and a yellow circle, a subject might report seeing a yellow triangle. This phenomenon, too, is open to alternative interpretations (Tsal, 1989; Green, 1991). Third, Treisman (1985) and Prinzmetal, Presti and Posner (1986) found that pre-cuing the location of a target with a symbol appearing just before the onset of the display greatly speeds conjunction search times, but has only a small effect on feature search times. This is presumably because the cue directs attention to the correct location in the display (cf. Posner, Snyder & Davidson, 1980).

There are many similarities between FIT and the present theory, particularly with regard to the nature of the selective attention mechanism. However, there are important differences. In particular, there is disagreement over the information signalled by individual processing units. The use of the term feature must be clarified here; I will write ‘Feature’ (capital F) when I use the word in Treisman’s sense and ‘feature’ when I use the word as defined in the previous chapter. In FIT, the exact nature of the Features is to be determined empirically, but it is assumed that they are single values on dimensions like orientation, size, colour, lightness, depth, and velocity. There are no units signalling combinations of these Features. In the present theory, features are operationally defined as the preferred stimuli of neurons in the low-level areas of the visual cortex. Neurophysiological evidence indicates that there are no neurons in the visual cortex that encode individual Features in the FIT sense. Instead, each neuron seems to be maximally activated by a particular combination of values on several dimensions—for example, a particular orientation, spatial frequency and binocular
disparity. That is, every feature is a conjunction of Features—precisely what should not exist according to FIT! Beyond early processing, neurons signalling shape fragments and views of objects are, of course, conjunction detectors *par excellence*. Treisman carefully states that her processing units should not be interpreted as single neurons (Treisman & Sato, 1990), but this clearly weakens the theory, particularly as physiological evidence is often given in support of it (e.g., Treisman & Gormican, 1988).

In order to account for the results of visual search experiments in the framework of the Magpie theory it is necessary to make certain assumptions about the capabilities of central cognition. This is unavoidable because the strategy used by the subject in performing the experimental task is, at least to some extent, a voluntary one—the subject chooses to look for the target and to press the appropriate button. My first assumption is that there is a difference between early processing and later processing stages in the type of information provided to central cognitive processes. I assume that early processing delivers information directly to central cognition about the *locations* of items in the viewed scene. This is reasonable, because neurons in early processing have relatively small RFs, and cells with simple RFs, in particular, are very fussy about the location of the contours that they respond to. Neurons in higher areas of the visual cortex have much larger RFs and are much less selective for stimulus position, and presumably provide central cognition with much poorer information about the location of items. However, I assume that early processing does not provide central cognition with much information about the *identity* of items. This, too, is reasonable, because features are rather unreliable indicators of the nature of objects in the viewed scene. For example, wavelength-selective neurons in area V1 do not exhibit colour constancy, and so presumably do not contribute directly to the conscious experience of colour. I assume instead that central cognition gets information about the identity of items from later processing stages, where the activity of a neuron can be taken to be a reliable indicator of the nature of an item in the environment.

Thus, it is likely that most of the visual neurons that send information about the identity of objects directly to central cognition have relatively large RFs. This means that the focusing mechanism is needed to prevent interference between the representations of different objects, as explained in Chapter 3. In particular, in a conjunction search experiment, a neuron that responds to the conjunction of Features that identifies a target could be activated in two ways, either by a target item in its RF, or by two distractor items that possess those Features separately and that both fall in its RF. For example, a neuron responding to a vertical red line in a large RF, which receives connections from lower-level neurons separately encoding the colour red and a vertical luminance contour, could be activated by a vertical-green distractor and a horizontal-red distractor both falling in its RF. This would be an illusory conjunction (Treisman & Schmidt, 1982). The neuron might not be activated as strongly by two distractors as by a target, but its activity could nevertheless be sufficiently great to make it hard for central cognition to discriminate between the two possibilities. Focusing must be used to restrict full analysis to individual items in the display. In this example, if the vertical-green item was focused, but not the horizontal-red item, then neurons encoding the colour red would not be strongly activated, and neither would any neuron responding to vertical red lines (refer back to the discussion of focusing in Chapter 3). Hence, the item would be correctly identified as a vertical-green distractor, not a target. This explanation predicts that illusory conjunctions are more likely to occur between nearby items, since they would then be more likely to fall within a single neuron’s RF. This is counter to Treisman and Schmidt’s (1982) claim that illusory conjunctions can be formed from very widely-separated items, but consistent with Cohen and Ivry’s (1989) finding that illusory conjunctions of colour and form only occur at above-chance levels between items that are separated by 1° or less outside the focus of attention. Note that, although early processing is assumed not to provide central cognition directly with information about the identity of items, it presumably does provide information about their locations, so that the focusing mechanism can be accurately directed to individual items.

This explanation accounts for the need for a serial search in the case of a conjunction target, and for the possibility of illusory conjunctions under conditions in which the subject is unable to deploy the focusing mechanism effectively. It remains to be explained why search for a target defined by a single Feature is independent of display size. I propose that central cognition is able to spread the influence of the focusing mechanism across the whole display, and to produce a target-present response if any out of a whole group of neurons that detect the target Feature and that have RFs
collectively covering the whole display are strongly activated. This is a similar explanation to that given by FIT, but with the difference that the neurons in question are assumed not to be strict ‘Feature detectors’. As pointed out above, there is no evidence for neurons that signal only individual Features. Rather, the pool of neurons that mediate Feature search responses is assumed to consist of cells that encode conjunctions of Features. Each individual neuron in the pool responds to a particular conjunction, consisting of the target Feature along with a number of other Features from other dimensions. In accordance with the assumption that central cognition obtains information about the identity of objects only from the higher stages of the visual cortex, these neurons have large RFs. This means that illusory conjunctions can occur in the absence of focusing—a neuron can respond to two items in its RF that separately possess the Features needed to activate the cell. But in Feature search, this is not a problem. The only information needed by central cognition in this task is whether the target Feature is present in the display. Other information about the target item, such as its location or its other Features, is not relevant to the task, so illusory conjunctions do not matter.

Treisman and Gelade (1980) reported that subjects could sometimes respond to the presence of a target without being able to report its location in Feature search, although, as noted above, this claim is particularly controversial (Green, 1991). The above explanation predicts that subjects should have rough but not exact knowledge of where the target is, since the neurons responding to the target have large RFs, and the focusing mechanism is spread over the whole display rather than concentrated on the target item. However, when the target has been detected by the means described above, the subject may focus on that item to make sure that it is indeed a target before producing a response. This would reduce the number of errors made in the experiment, and the increase in reaction time would not be great because there is no need for a serial, item-by-item search. The initial detection of the target gives approximate information on its location, so the focusing mechanism can directly ‘zoom in’ on the region of the display containing the target. This might account for the phenomenological pop-out effect. Although the number of distractors would have no effect on the time taken by this focusing operation, the density of items in the display might have a small effect, since the nearer the distractors are to the target, the more narrow the focus needs to be to isolate the target. The use of focusing to check the presence of the target is supported by Treisman’s (1985) finding that pre-cuing the location of the target 100 msec before the onset of the display significantly facilitates the speed of Feature search.

A variant of Feature search in which the target is an item that lacks a Feature possessed by all the distractors produces very different results. Reaction time increases steeply with display size, and the same 2:1 ratio of the average reaction times of negative and positive responses is found as in conjunction search (Treisman & Souther, 1985; Treisman & Gormican, 1988). This suggests that a serial search of the items needs to be performed for this task. This may be because all the neurons that respond to the target can also be activated by the distractors. Thus, the method proposed above for producing the fast responses in Feature search cannot be used in this case. There is no pool of neurons that reliably signals target items only.

It has often been assumed that the properties of items that determine whether a target will pop out in a visual search display are the same as the properties of texture elements that determine whether two textures perceptually segregate (e.g., Treisman & Gelade, 1980; Bergen & Julesz, 1983). However, Wolfe (1992) has produced counter-examples to this claim. According to the present theory, the perception of texture contours (p. 126) does not play an essential part in visual search. However, if the items in a feature search display are quite densely packed, it is certainly conceivable that the presence of a texture contour ringing the target item helps to attract the focusing mechanism to that location.

Conjunction search does not necessarily require a serial, item-by-item search for the target. Treisman (1982) tested conjunction search using displays in which similar distractors were grouped together in spatially-contiguous regions, and found that search was much faster than when different distractors were randomly intermixed. She argued that this was because the subjects could perform the serial search group-by-group instead of item-by-item. Within each group, the target, if present, differs from the distractors in just one Feature, and so the task reduces to Feature search. This explanation is easily translated into the terms of the present theory: if the focus is narrowed to cover a group of items which all share one of the two target Features, a strong response by neurons encoding the other Feature signals the presence of a target in that group; lack of such response indicates the absence of a
target, so the focus is moved to another group. Several laboratories, however, have reported that conjunction search can be very fast in some circumstances even if the distractors are randomly intermixed—so fast that it appears to be independent of display size. For example, Nakayama and Silverman (1986) tested search for targets defined by the conjunction of binocular disparity and velocity, and of disparity and colour. The function relating reaction time to display size was flat in both cases. Phenomenologically, all the items with the same disparity can be segregated as a single plane in depth, and the target is then easily perceived as the only item in that depth plane with mismatched velocity or colour. Similarly, McLeod, Driver and Crisp (1988) found that certain conjunctions of motion and form can be easily perceived. In general, conjunction search can be very fast if the items are highly discriminable (Wolfe, Cave & Franzel, 1989; Treisman & Sato, 1990). The phenomenology suggests that all the items having one target Feature in common can be perceived as a figure, the remaining distractors being ignored as part of the ground; the target then pops out of the figure, if it is present, because it is the only item in the figure having the other target Feature. In the previous section I proposed that the figure is simply the pattern selected by the focusing mechanism. Thus, the explanation of the fast conjunction search in these cases is basically the same as that given above for Treisman’s (1982) grouped displays: all the items with a common Feature (e.g., the target disparity) are focused, and any activity within this focused group by neurons encoding the other Feature (e.g., the target colour) can then be taken by central cognition to signal the presence of a target. The problem with this explanation, as McLeod et al. (1988) point out, is that the items seen as figure can be spatially intermingled with the other distractors. This conflicts with the assumption that only spatially-contiguous regions of the image can be focused. In addressing this problem in the previous section I suggested that feedback and lateral excitatory connections between neurons encoding similar features (e.g., similar disparities or velocities) could be important in figure-ground segregation in the case of overlapping patterns (Fig. 5.21).

In conclusion, the account of visual search given by the present theory incorporates some of the basic assumptions of FIT: the hypotheses that illusory conjunctions are a problem in conjunction search, that a selective attention mechanism is used to overcome the problem, and that Feature search is performed by detecting activity in a large group of neurons that collectively cover the visual field. However, the detailed architectures proposed by the two theories are very different, and that of the present theory is much more consistent with the known physiology of the visual cortex. In particular, the theory allows for the fact that most if not all neurons in the visual cortex encode conjunctions of Features.

5.10 Summary
This chapter has attempted to explain shape perception and object recognition in terms of the computational mechanisms introduced in Chapter 3. It was assumed that these operations are basically achieved by the feedforward activation of a hierarchy of pattern-detecting neurons in the visual cortex. Various more complicated processes are required to explain certain effects, such as the use of imagery in recognizing objects from unfamiliar views and the need for selective attention when there are many objects in the viewed scene. The remainder of the chapter gave an account of a number of special topics: the Gestalt grouping effects, illusory contours and interpolation, the perception of three-dimensional structure in pictures, multistability and hysteresis, stereopsis, the figure-ground phenomenon, and visual search.

Apart from a need for a more detailed account of these topics, the theory is limited in that many important aspects of visual perception have been neglected. One of these is the perception of the spatial layout of the viewed scene and of the distance of objects from the observer. A closely related subject is the coordination of visual perception with motor performance, in particular with eye movements. Any complete account of vision will have to address these topics.
Chapter 6
Conclusion

As stressed in Chapter 1, the purpose of this thesis is not to describe a unified theory of cognition but merely to present an argument that such a theory is a feasible goal, based on the fundamental hypothesis that the cerebral neocortex is an approximately uniform cognitive architecture. In support of this hypothesis I have presented a theory of the general-purpose computational mechanisms provided by this architecture, a theory that is computational in character but firmly grounded in current knowledge of neuroanatomy and neurophysiology. Perhaps many readers will feel that I have overstated the case for cortical uniformity. Certainly there are likely to be significant differences between cortical areas in computational mechanisms, and it is important to remember that other parts of the brain with architectures quite different from the neocortex also make a vital contribution to human cognition. It is only claimed that the importance of the cortex and the degree to which it is computationally uniform are both sufficient for the development of a general theory to be a worthwhile exercise. This may be the case even if the details of the Magpie theory turn out to be quite wrong. Several other researchers have proposed general theories of brain function that are consistent with the hypothesis. These include theories that are directly inspired by the anatomical uniformity of the cortex (Marr, 1970; Eccles, 1984; Barlow, 1985; Mel & Koch, 1990; Singer, 1990; Mumford, 1991, 1992; James & Hoang, 1993), more general theories that are proposed to be applicable to many parts of the brain including the cortex (Grossberg, 1980; Carpenter & Grossberg, 1987; Edelman, 1987; Poggio, 1990), as well as general theories in the symbolist tradition (Anderson, 1983; Laird, Newell & Rosenbloom, 1987; Newell, 1990).

With regard to the present theory, one of its fundamental assumptions is that perception is basically a ‘bottom-up’ process, characterized by a mainly feedforward flow of information from the thalamus up through the hierarchy of cortical areas. This can be contrasted with theories that emphasize ‘top-down’ processing—in particular, proposals that visual perception requires the active interpretation of the retinal image by elaborate problem-solving computations (e.g., Gregory, 1970; Rock, 1983). The present theory also has a different emphasis from some other connectionist models of vision that make heavy use of competitive and cooperative interactions between neurons at the same hierarchical level (e.g., Marr & Poggio, 1976; Grossberg & Mingolla, 1985; Finkel & Edelman, 1989; Grossberg, 1994). Of course, there are top-down and interactive processes in the Magpie theory, especially the feedback and lateral connections and the focusing mechanism, but their function, although certainly important, is supplementary to that of feedforward activation. Also, it is not denied that central cognitive processes can have important influences on perceptual processing, especially at the higher stages. The use of visual imagery was noted in the discussion of object recognition, and there is good reason to suppose that the perception of abstract visual properties, like number and connectedness, necessarily makes use of more sophisticated computations than those discussed here (Ullman, 1984; Minsky & Papert, 1988). But timing considerations strongly suggest that, in most circumstances, perception is basically a simple, feedforward process. The time constants of neurons, synaptic delays, and limited axonal propagation speeds place strong constraints on the amount of sequential computation that can be performed between the reception of the visual image by the photoreceptors and the extraction of useful information about the layout of the viewed scene and the identity of objects (Feldman & Ballard, 1982; Thorpe & Imbert, 1989).

Another feature of the theory is the proposal that the cortex achieves its prodigious capabilities by the use of a vast number of neurons and connections. Put this way it might seem a truism, but it is
surprising how often this simple fact is overlooked by psychologists. For example, several theories of the representation of information in human memory are based on the assumption that economical use of storage capacity is of fundamental importance. I suggest that, although the need for economical use of storage was indeed an important factor in the evolution of the brain, it was only one factor among many. An analogy can be drawn with the design of algorithms in computer programming, where it is common to face a trade-off between economy and speed. For example, when implementing a mathematical function, one might write a complicated algorithm for computing the function accurately, or one might store a table of numbers such that the value of the function can be approximated by interpolation. The former solution might be more economical in that the program code occupies less storage than the table of numbers, but it is also likely to take more time to produce the output. It is conceivable that there was an analogous trade-off between storage economy and processing speed in the evolution of the brain. Clearly, natural selection must have exerted a strong pressure for the perceptual systems to deliver their information quickly, especially as their biological components are inherently rather slow, as noted above. This suggests that the cortex is most likely to use fast, feedforward algorithms, at least for those purposes for which it originally evolved, even if this is at the cost of a rather profligate use of neurons.

A further major characteristic of the Magpie theory is its assumption of substantial plasticity in the functional properties of cortical neurons, and the prominent role that it gives to experience in shaping those properties, as enshrined in the Empiricist Principle (p. 61). The fundamental hypothesis of the theory is that the same computational mechanisms are used throughout the cortex, and this implies that the same mechanisms of plasticity are used everywhere, so that the diversity observed in the adult cortex must be a result of diversity in the information processed and stored. The assumption of uniform computational mechanisms does not logically imply that the cortex of the newborn animal is a tabula rasa, with no innate information. Nevertheless, it would seem a little inconsistent, in spirit if not in logic, to argue both against faculty-specific mechanism and for innate faculty-specific information. In other words, there is a natural tendency for those who advocate uniformity of the cognitive architecture to adopt an empiricist approach to knowledge. By the way, the same is true the other way around: this is why the British Empiricist school of philosophy, founded on the premise that there is no innate knowledge, tried to explain the mind in terms of a uniform mechanism, namely the association of ideas. Interestingly, the founder of Empiricism, John Locke, was no associationist; he assumed the mind to be equipped with certain (presumably innate) powers with which knowledge could be acquired—a capacity for perceiving, remembering, reasoning, comparing, willing, and the like (Locke, 1706, Bk. 2). In fact, Locke conceived of the newborn child as having a rich and complex cognitive architecture with no stored information, in modern language. This is a perspective I have sympathy with, although I would not endorse Empiricism without reservation (p. 61).

One of the most important assumptions of the Magpie theory is that the cortex is a perceptual-motor architecture that is largely dedicated to the task of pattern recognition. This assumption has a strong evolutionary foundation, as explained in Chapter 2, and as a consequence I have neglected processes such as reasoning, problem solving, imagining and remembering throughout the thesis, although of course the cortex is assumed to be largely responsible for these tasks. Even so, I have had to invoke unspecified central cognitive processes at various points in the account of visual perception. Indeed, the entire theory of visual perception might be thought unsatisfactory on the grounds that I have not specified what happens next after the neurons representing shape fragments or object views have been activated. The reader may feel that the ghost has been chased further back into the machine but has still not been exorcised (Fodor, 1983, p. 127). For this reason I would like to finish the thesis with some brief speculations on consciousness.

Consciousness has been the Cinderella of psychology since the rise of behaviourism in the early years of this century, but in recent years it has become scientifically respectable once again, not to say fashionable (e.g., Jackendoff, 1987; Edelman, 1989; Crick, 1994). Defining consciousness is a mug’s game: I will assume the meaning of the term to be intuitively obvious, except to note that I do not identify consciousness with self-awareness. (The latter is merely a special case; we can have thoughts about ourselves just as we can have thoughts about anything else.) Consciousness is not, I assume, the function of a particular module in the human cognitive architecture, separate from the modules responsible for perception, memory, language, etc., but is instead produced by all of these other modules working together. Consciousness is therefore produced by the collective operation of many
regions of the cortex (and perhaps other parts of the brain). (See Dennett (1991) for extensive
discussion of the merits and implications of this point of view.) Consciousness cannot simply be
identified with central cognition: many relatively peripheral processes (of the kind considered in
Chapters 4 and 5) probably contribute directly to consciousness, whereas there may be processes
involved in problem solving, for example, that are unconscious. It is useful to divide the problem of
explaining consciousness into two separate parts (cf. Jackendoff, 1987): the computational problem is
to explain the processes in the brain that underlie consciousness, and the phenomenological problem
is to explain the relation between these physical processes and the sensations, images, feelings and
thoughts (what philosophers call qualia) that make up conscious experience. Generally speaking,
philosophers have ignored the former problem and scientists have ignored the latter, and I will respect
this tradition: my assumptions regarding the phenomenological problem were stated in the
introduction to Chapter 5 and are not elaborated here. I will further restrict the computational problem
by making an undoubtedly crude division of conscious experience into three components. One
component that is not further discussed here is consciousness of the external world (including the
body): that is, the vivid sensations of colour, sound, movement, warmth, pain, and so on that
constitute perception. The computations underlying these experiences in the visual modality have
already been considered in detail in Chapters 4 and 5. A second component of conscious experience is
appetitive and emotional feelings: hunger, thirst, anger, love, sadness, joy, envy, and so on. It is
plausible that certain centres in the limbic system are at least as important as the neocortex in
generating these feelings. The third component, which is addressed here, is the internally-generated
succession of memories, images, ideas, intentions, speculations, judgments, comparisons, fantasies,
deductions, conceptions and the like that make up ‘the stream of thought’ (James, 1890). It is likely
that the prefrontal cortex is especially important for this type of consciousness: this region comprises
about 30% of the whole cortex in humans and there is much neuropsychological and
neurophysiological evidence linking it to central thought processes (reviewed by Fuster, 1989). Other
parts of the cortex certainly contribute to the stream of thought, however. For example, visual imagery
has been shown to involve parts of the visual cortex (Farah, 1988).

The problem to be addressed, then, is how the theory of cortical function described in Chapter 3
may be applied to explaining the stream of thought. Consider the following suggestion. Symbolist
architectures known as production systems have been very popular among cognitive psychologists
modelling central cognitive tasks (e.g., Newell & Simon, 1972; Anderson, 1983; Newell, 1990). A
typical production system consists of a working memory containing temporary symbol structures, and
a collection of production rules which act on the working memory. Each rule has two parts: a pattern
which can be matched against the contents of the working memory, and a specification of one or more
actions to be performed. The system operates in cycles. In each cycle the patterns of all the rules are
matched against the current contents of the working memory, one matching rule is selected, and the
actions of that rule are carried out, modifying the working memory and perhaps producing output.
This process repeats sequentially. The evolving symbol structures in the working memory can be
taken to model the changing contents of thought.

Production systems are appealing because they have a number of properties that are analogous
to the character of human thought. First, their operation is basically serial, yet the pattern-match phase
is well-suited to implementation in massively parallel hardware. Second, the sequence of operations
carried out by a production system is not fixed in advance but is sensitive to changing circumstances.
For example, if a separate perceptual processor deposits a symbol structure in the working memory
asserting that something important has happened in the environment, the production system can
immediately abandon the task it is working on and start processing the new information. This models
one of the most prominent characteristics of human thinking, its susceptibility to distraction. Third, a
production system is a very suitable architecture in which to implement a particular form of learning
by practice known as chunking (Newell & Rosenbloom, 1981; Anderson, 1983; Newell, 1990). If a
production system performs a particular sequence of actions repeatedly, a new production rule can be
created which recognizes the preconditions of that sequence and yields the appropriate actions in a
single cycle. For example, given the problem ‘what is 7 times 9?’ the system might at first carry out a
laborious sequence of calculations, but with practice a new rule is formed that recognizes this problem
and immediately puts the answer into the working memory. This is analogous to the shift from an ‘analytical’ to a ‘perceptual’ mode of thought during cognitive skill acquisition that was noted in Chapter 2 (p. 30).

Of course, it is not my view that the human cognitive architecture is a production system. Aside from the general argument against symbolist architectures presented in Chapter 2, the symbolic data structures postulated by theories like Soar (Newell, 1990) do not fit well with the visual imagery and ‘inner speech’ that fill consciousness when we are thinking, and the discrete cycles of a production system contrast with the continuous flow of consciousness that James (1890) emphasized (although neither of these arguments can be called conclusive). On the other hand, it is not difficult to see how a process qualitatively similar to that performed by a production system could be carried out by the cortex, assuming there is some truth in the Magpie theory. Suppose that the output from the deep layers of a particular cortical area is returned as feedforward input to the middle layers of that same area, via one or more subcortical nuclei, forming a closed loop. This concept is illustrated in Figure 6.1. The circulation of information around the loop could function as a working memory, and the pattern recognition by layers 2–4 and the generation of output in the deep layers of the cortical area would not be too dissimilar from the function of the rules in a production system.

There is substantial anatomical evidence for the existence of loops of this kind, involving several different subcortical centres. One example is the reciprocal pathways between the cortex and the claustrum; the connections with this nucleus were described in the review of cortical anatomy in Chapter 2 (p. 24). As a second example, many cortical areas in the frontal lobe are reciprocally connected with the basal ganglia. In detail, layer 5 of each area projects to a part of the striatum (caudate nucleus or putamen), which projects to a part of the globus pallidus or substantia nigra, which projects to one of the ventral nuclei of the thalamus, which projects back to the original cortical area again. Several different frontal lobe areas (both motor and prefrontal) are involved in loops of this kind, and there appears to be rather little cross-talk between the different loops within the basal ganglia (Alexander, DeLong & Strick, 1986; Middleton & Strick, 1994). A third example of looping circuitry is the reciprocal connections between many cortical areas and the so-called association nuclei of the thalamus—that is, nuclei like the medial dorsal nucleus and the pulvinar, in which the input from the cortex appears to play a more important role than the input from subcortical sources. These various subcortical loops may all contribute to the stream of thought: it is highly unlikely, of course, that any single cortical area is crucial for consciousness.

The correspondence between the circuitry of Figure 6.1 and the structure of a production system should not be taken too literally. There are at least three important differences, apart from the fact that one architecture is connectionist and the other symbolist. First, the contents of working memory (or of consciousness) should probably not be identified with activity in the subcortical relay

![Figure 6.1: Circuitry possibly contributing to the stream of thought. Output generated by the deep layers returns as the feedforward input to the middle layer, after relay in one or more subcortical nuclei. To simplify the diagram only the most significant elements of the cortical circuit are shown.](image-url)
nucleus, but with the contents of active storage in the relevant cortical areas. Whereas a production system makes a clean spatial separation between the production rule memory and the working memory, in the cortex active storage and latent storage are inextricably linked. Second, it would be a mistake to regard the subcortical loop as a simple reverberatory circuit, merely feeding the cortical output back into the cortex after a certain time delay. Instead, the subcortical nuclei in question surely perform complex spatial and temporal transformations on the information they receive. Third, in most production systems (although not Soar) only a single production rule is allowed to be applied to the working memory in each cycle. It seems doubtful that any such constraint would apply in the cortex. In spite of these differences, the hypothetical scheme of Figure 6.1 does appear to provide a feasible account of how a connectionist architecture that evolved for perceptual-motor processing could be responsible for the stream of thought. It shares with production systems the three characteristics of thought listed earlier: serial processing in a parallel architecture, interruptibility, and the potential to substitute cumbersome sequential processing with a more efficient strategy based on pattern recognition. Evaluating this proposal in detail will require further research.
Glossary

This glossary provides brief definitions of technical terms and abbreviations that are used throughout the text. Some words, such as architecture, neuropsychology, and just about any word ending in -ism, are used by different people to mean very different things; this glossary defines how they are used in this thesis.

Accommodation. Progressive reduction in the firing rate of a neuron to a maintained stimulus, over a time course of tens to hundreds of milliseconds.

Acetylcholine. A neurotransmitter that has a modulatory effect on cortical neurons. Acetylcholine is released in the cortex by fibres originating in the nucleus basalis of Meynert.

ACh. Acetylcholine.

Achromatic. Without colour.

Action potential. A brief electrical impulse used for long-distance signalling in the nervous system. An action potential is initiated at the start of the axon when the membrane potential in the soma of the neuron reaches a certain threshold. The impulse propagates along the axon and all the axon collaterals, with no reduction in its magnitude, until it reaches the synapses at the axon terminals. The generation of each action potential is followed by a brief refractory period during which the neuron cannot produce another one, placing an upper limit on the firing rate. In some parts of the brain dendrites as well as axons can produce action potentials.

Active storage. Representations in a connectionist architecture encoded by the activity of the units. See latent storage.

Acuity. Ability to discriminate fine details in a sensory pattern.

Adaptation. In psychophysics, refers to a reduction in sensitivity to a stimulus with prolonged or repeated presentation. In physiology, refers to a reduction in response of a sensory neuron to a maintained stimulus over a time course of seconds to minutes. The latter is assumed to underlie the former.

Agonist. A substance that binds to a certain type of receptor and causes the associated membrane channels to open. Not necessarily used as a neurotransmitter.

Allocortex. Primitive, three-layered region of the cerebral cortex. Includes olfactory cortex and the hippocampus.

Amnesia. Loss of memory. Retrograde amnesia is inability to recall memories acquired before the event that caused the amnesia; anterograde amnesia is inability to acquire new memories after the event.

Amygdala. Collection of nuclei in the forebrain, part of the limbic system. Has reciprocal connections with certain areas of the cortex as well as the hypothalamus and other parts of the brain. Important in control of instinctive and emotional behaviour.

Antagonist. A substance that blocks a certain type of receptor, thus preventing the effects of the neurotransmitter that normally activates that type.

Anterior. Towards the front.
Apical dendrite. The main dendrite of a cortical pyramidal neuron, in most cases ascending through the cortex to the uppermost layers where it splits into a tree of dendritic branches.

APV. 2-amino-5-phosphonovalerate, an antagonist of NMDA receptors.

Architecture. See cognitive architecture.

Area. A region of the cortex in which neurons have similar functional properties and similar connections with other areas and other parts of the brain.

Auditory cortex. Region of the temporal lobe responsible for auditory perception.

Axon. Cable emanating from the soma of a neuron that carries action potentials from that neuron to other neurons. Forms synapses onto the dendrites or somas of other neurons.

Axon collateral. An axon branch emanating from the main axon.

Basal ganglia. Collection of nuclei in the forebrain receiving connections from broad regions of the cortex and returning connections to certain areas in the frontal lobe. Includes the striatum (caudate nucleus and putamen), globus pallidus, substantia nigra, and subthalamic nucleus. Involved in motor and central cognitive functions. (Fig. 2.2.)

Basket cell. Type of inhibitory neuron in the cortex with smooth dendrites and quite widespread axon collaterals within the cortex. The collaterals form specialized ‘baskets’ around the somas of pyramidal neurons, making multiple inhibitory synapses. (Figs. 2.5H, 2.6H.)

Bilateral. Relating to both sides of the brain.

Binaural. Relating to both ears.

Binding problem. Given that the visual attributes of objects are represented in different areas of the cortex, and that there are usually many objects in the viewed scene, the binding problem is the problem of associating together all the attributes of each object in active storage.

Binocular disparity. A difference between the images of an object on the two retinas, in particular a difference in position, resulting from the fact that the two eyes look at the object from different directions. The brain is able to use these disparities to obtain information on distance and shape.

Bipolar cell. Type of neuron in the cortex with a spindle-shaped soma and dendrites ascending and descending within a narrow column. Many of these neurons release peptide neurotransmitters, which may have a modulatory effect on neighbouring cells. (Figs. 2.5G, 2.6E.)

Brain stem. Ancient part of the brain between the spinal cord and the forebrain, including the reticular formation, pons, cerebellum, and tectum. Involved in many sensory, motor, and regulatory functions. (Fig. 2.1.)

Broad tuning. See tuning.

Broca's area. Region in the frontal lobe of the left hemisphere of the human cortex, important for language. Probably not a single area.

Canonical cortical circuit. A hypothetical idealization of the neural circuitry present in all areas of the neocortex. The circuitry actually present in any particular area is a more-or-less elaborated version of the canonical circuit.

Capacitance. Ability of a neuron’s membrane to store electric charge. Means that changes in the membrane potential cannot occur instantaneously but require time to build up and decay. Measured in farads.

Caudal. Towards the back of the brain.

Caudate nucleus. Part of the striatum in the basal ganglia. Receives connections from large regions of the cortex, especially prefrontal cortex, and sends connections to parts of the globus pallidus and substantia nigra. (Fig. 2.2.)
Central cognition. Cognitive processes other than perception and motor control, such as remembering, imagining, language processing, reasoning and problem solving.

Cerebellum. An outgrowth of the brain stem, consisting of a number of nuclei surrounded by a cortex which has a very regular internal structure. Important in the control of skilled movements. (Fig. 2.1.)

Chandelier cell. Type of inhibitory neuron in the cortex with smooth dendrites and axon collaterals which form multiple inhibitory synapses onto the initial segments of the axons of neighbouring pyramidal neurons. (Fig. 2.6D.)

Channel. See membrane channel, visual channels.

Characteristic pattern. In the Magpie theory, the pattern of firing rates in the feedforward excitatory input connections of a neuron to which the neuron gives a maximal response.

Cholinergic. Relating to the neurotransmitter acetylcholine.

Chromatic. Relating to colour.

Claustrum. Nucleus in the forebrain having reciprocal connections with the whole cortex, but apparently having few connections with any other part of the brain. (Fig. 2.2.)

Cochlea. Part of the inner ear, the sense organ for hearing.

Cognition. Thought processes in general, conscious and unconscious, regarded as a form of information processing. Includes perceptual and motor control processes.

Cognitive architecture. All the structures and mechanisms of the cognitive system that are genetically specified and not modifiable by learning. Assumed to determine the gross division of the system into processing modules and communication pathways, as well as the basic computational mechanisms within each module that are directly implemented in the neural circuitry.

Collateral. See axon collateral.

Column. A small block of cortex, extending through all the cell layers but no more than a millimetre or so across, in which neurons have similar physiological properties.

Commissural connections. Connections between two cortical areas in opposite cerebral hemispheres, passing through the corpus callosum.

Complex cell. Neuron in the visual cortex responding selectively to a contour of a particular orientation, but not very selective for the position of the contour in the receptive field. See simple cell.

Computational map. A cortical area regarded as a sheet-like processor in which the same computations are performed in parallel across the sheet, and in which some parameter of the information being processed (e.g., location on the surface of the body) is topographically mapped across the sheet.

Computational mechanisms. General term for the representations and elementary computational operations provided by the cognitive architecture.

Computer. In its most general sense, a computer is a machine in which certain physical states can be regarded as representations, and in which the transformations of those representations performs some useful function.

Conductance. Measure of the ease with which electric current can flow across a neuron’s membrane. Reciprocal of resistance. Measured in siemens (reciprocal ohms).

Cone. Type of photoreceptor in the retina, less sensitive to light than the rods but necessary for high acuity vision. In humans and some other primates there are three types of cone with different wavelength selectivities, responsible for colour vision.
**Conjunction problem.** The problem of bringing diverse items of information into physical proximity in the cortex so that a record of their combination can be formed in latent storage.

**Connectionist architecture.** A type of cognitive architecture consisting of a large number of interconnected processing units, each processor having neuron-like functional properties. Includes both highly simplified and biologically realistic models of the brain.

**Connection plasticity.** Long-lasting changes in the effectiveness of connections between neurons, resulting from changes in the potency of synapses or from the formation of new synapses or the elimination of old ones. Thought to underlie learning and memory.

**Contextual enhancement.** Increase in the response of a neuron to an input pattern when the pattern is consistent with the surrounding pattern. Assumed in the Magpie theory to be caused by connections of the feedback and lateral type.

**Contour.** Border between two regions of a visual image differing in some quantity, such as luminance, colour or texture.

**Contralateral.** The opposite side of the brain.

**Contrast.** Difference in luminance between the brightest and dimmest parts of a visual pattern. Usually expressed as the ratio of this difference to the mean luminance of the pattern.

**Convolution.** A mathematical operation that can be thought of as applying a two-dimensional function (modelling a receptive field) to every point of a visual image. For each point in the image, the intensity values around that point are multiplied by the corresponding function value, and all these products are summed to give the convolution value at the point.

**Corpus callosum.** The massive bundle of fibres passing between the cortices of the two cerebral hemispheres. (Figs. 2.1, 2.2.)

**Cortex.** In general, any thin layered sheet of grey matter, such as the cerebral cortex and the cortex of the cerebellum. However, the term is usually used to refer to the neocortex, except where confusion may arise.

**Corticocortical connections.** Connections between different areas of the cortex.

**Critical period.** A period during the brain’s development when the neurons in a given region are especially susceptible to connection plasticity.

**CSF.** Contrast sensitivity function.

**C-unit.** In the Magpie theory, a neuron that keeps a relatively long-lasting trace of the activity of its feedforward input synapses. See S-unit and Trace.

**Dendrite.** Cable emanating from the soma of a neuron. Cortical neurons may have several dendrites, and each dendrite may split into many branches. Most of the synapses that provide inputs to the neuron are located on its dendrites.

**Dendritic spines.** Small protuberances from the dendrites of certain types of neuron, notably pyramidal neurons in the cortex, on which excitatory input synapses are preferentially located.

**Depolarization.** Bringing a neuron’s membrane potential closer to zero than the resting potential, i.e., exciting the neuron. See hyperpolarization.

**Depth.** Visual percept of the distance to a surface.

**Difference of Gaussians.** A ‘Mexican hat’ shaped function, obtained by subtracting one Gaussian function from another where one Gaussian is tall but narrow and the other is shallow but broad. Used to model the distribution of sensitivity in the receptive fields of ganglion cells in the retina. (Fig. 4.2.)

**Directional selectivity.** Property of some neurons in the visual cortex that respond more strongly to a stimulus moving in one direction across the receptive field than the opposite direction.
Disparity. See binocular disparity.

DOG. Difference of Gaussians.

Dopamine. Neurotransmitter released by fibres that innervate the whole cortex and that originate from neurons in the substantia nigra.

Dorsal. Towards the top of the brain.

Dualism. The theory that consciousness resides in a spiritual substance distinct from the physical brain.

Early processing. The initial processing stages of sensory information, characterized by highly localized and parallel computations.

Eccentricity. Angle between a point on the retina and the centre of the fovea.

Electrotonic spread. Dendrites are poorly-insulated cables, so a local change in the membrane potential caused by activation of a synapse attenuates as it spreads passively along the dendrite. This is called electrotonic spread of potential.

Empiricist Principle. The principle that perceptual patterns encoded as the characteristic patterns of cortical neurons should reflect the statistical properties of normal sensory input, so that patterns that occur frequently in the input should be well-represented. This does not necessarily imply that the characteristic patterns of all neurons are acquired from experience.

End-stopping. Property of some simple and complex cells in the visual cortex which give a good response to a contour contained within the receptive field but a much weaker response to a contour that extends beyond the receptive field in either or both directions.

Excitation. Tending to increase the rate at which a neuron generates action potentials.

Extracellular. Outside the cell.

Faculty. A hypothetical division of the mind, such as visual perception, imagination, motor control, language production, reasoning, and the like.

Feature. Operationally defined as simple visual stimuli encoded by the characteristic patterns of neurons in the low-level areas of the visual cortex.

Feedback connections. Connections from higher to lower stages of a processing hierarchy. In the case of connections between cortical areas, feedback connections typically originate mainly from the deep layers of the source area and terminate outside layer 4 of the target area. (Fig. 2.8.)

Feedforward connections. Connections from lower to higher stages of a processing hierarchy. In the case of connections between cortical areas, feedforward connections typically originate in the upper layers of the source area and terminate in and around layer 4 of the target area. (Fig. 2.8.)

Fibre. Another word for axon.

Firing rate. The rate at which a neuron generates action potentials.

FIT. Feature integration theory.

Focusing mechanism. Mechanism of the Magpie theory that selectively enhances the activity of neurons analysing a restricted part of the sensory input. Needed to overcome the problem of interference between representations of simultaneous stimuli, and for solving the binding problem.

Folk psychology. Common sense knowledge of the mind, as distinct from scientific psychology. Not meant to be derogatory.

Forebrain. The brain above the brain stem. Includes the cerebral cortex, thalamus, basal ganglia and limbic system.
Fourier analysis. A mathematical procedure for decomposing a complex waveform into the sum of a number of sine waves differing in frequency, amplitude, and phase.

Fovea. Small depression in the retina of primates where cones are concentrated and where acuity is highest.

Frontal lobe. The part of the cerebral cortex anterior to the central sulcus, consisting of motor cortex and prefrontal cortex. (Fig. 2.1.)

Functionalism. The theory that consciousness results from certain functions performed by the brain. It is assumed to be the functions themselves, rather than their material instantiation, that are crucial.

GABA. \(\gamma\)-aminobutyric acid.

GABA\(_A\), GABA\(_B\) receptors. Two types of receptor for the inhibitory neurotransmitter \(\gamma\)-aminobutyric acid. In cortical pyramidal neurons it appears that GABA\(_A\) receptors are used in synapses around the soma whereas GABA\(_B\) receptors are used in synapses on the dendrites. The functional significance of the two different types is unknown.

\(\gamma\)-aminobutyric acid. Neurotransmitter with an inhibitory effect on cortical neurons, released by several types of smooth neuron.

Ganglion cells. The output neurons of the retina. Their axons form the optic tract. Most of the ganglion cells that send connections to the lateral geniculate nucleus have circularly-symmetric receptive fields.

Gaussian function. The one-dimensional Gaussian is familiar as the Normal distribution in statistics. The two-dimensional Gaussian is obtained by rotating this curve about its central axis, yielding a bell-shaped surface. Formally it is \(G(x, y) = k \exp\left(-\frac{x^2 + y^2}{r^2}\right)\) where \(k\) is the peak height and \(r\) the radius.

Globus pallidus. Part of the basal ganglia, receiving connections from the striatum and sending connections to the thalamus. (Fig. 2.2.)

Grandmother cell. A hypothetical neuron in the visual cortex that is strongly active if and only if one perceives (or perhaps imagines) one’s grandmother.

Grey matter. Brain tissue that is densely packed with the somas and dendrites of neurons. See white matter.

Hard weight. A connection weight that cannot be altered by connection plasticity.

Hebbian plasticity. A form of connection plasticity in which the effectiveness of a synapse in activating a neuron increases if there is a conjunction of high presynaptic and high postsynaptic activity. For example, the presynaptic activity might be the firing rate and the postsynaptic activity the membrane potential.

Heterogeneous architecture. A type of cognitive architecture in which different faculties make use of computational mechanisms that have little in common with one another. See uniform architecture.

Hippocampus. A part of the allocortex, consisting of the subiculum, dentate gyrus, and CA fields. Has few direct connections with the neocortex, but substantial connections with the parahippocampal cortex. Also sends output to the hypothalamus and septum. Important for memory. (Fig. 2.2.)

Homunculus. A ‘little man in the head’, a module in a hypothetical cognitive architecture which has abilities equivalent to those of the architecture as a whole.

Hyperpolarization. Making a neuron’s membrane potential more negative than the resting potential, i.e., inhibiting the neuron. See depolarization.
**Hypothalamus.** A small but vital collection of nuclei at the base of the forebrain. Part of the limbic system. Involved in regulating body temperature, water and food balance, blood pressure, hormones, and the like. Important in motivation and emotion. (Fig. 2.1.)

**Hysteresis.** Persistence of an effect after changes in the conditions that caused it.

**Inferior colliculus.** Nucleus in the brain stem, a stage in the auditory pathway from the cochlea to the auditory cortex. (Fig. 2.1.)

**Inferotemporal cortex.** Large region of visual cortex in the temporal lobe in which many neurons respond to elaborate visual patterns such as faces.

**Inhibition.** Tending to reduce the rate at which a neuron generates action potentials.

**Intracellular.** Inside the cell.

**Introspection.** Observation of one’s own conscious experience.

**In vitro.** Literally, ‘in glass’. Refers to a small slice of living brain tissue outside the body. Convenient for some kinds of experiment, but caution must be applied in generalizing the results to the intact animal.

**In vivo.** In the living animal. See in vitro.

**Ion.** An electrically-charged molecule. The flow of certain kinds of ion (chiefly sodium, potassium, chloride and calcium) between the inside and outside of a neuron through membrane channels affects the membrane potential.

**Iontophoresis.** Technique for injecting ions very precisely into brain tissue by passing current through a micropipette.

**Ipsilateral.** The same side of the brain.

**Isoluminant.** Refers to a visual pattern in which luminance is uniform but colour varies.

**IT.** Inferotemporal cortex.

**Latency.** Time delay between the presentation of a stimulus and the response of a neuron.

**Latent storage.** Representations in a connectionist architecture encoded by the weights of the connections between the units, which change slowly and persist a long time. See active storage.

**Lateral.** Towards the side.

**Lateral connections.** Connections within a single level of a processing hierarchy, as opposed to feedforward and feedback connections. Lateral connections between cortical areas seem to have no particular laminar organization.

**Lateral geniculate nucleus.** The nucleus in the thalamus that relays visual information from the retina to the visual cortex. (Figs. 2.7, 4.1.)

**LGN.** Lateral geniculate nucleus.

**Lightness.** Visual percept related to the reflectance of a surface: a surface that reflects a large fraction of the light falling on it appears light, whereas a surface that absorbs most of the light appears dark or black. Not the same as brightness, which is more related to luminance.

**Limbic system.** Collection of interconnected structures in the forebrain, including the hypothalamus, amygdala, septum, and hippocampus. The term ‘system’ might be a misnomer.

**Linear.** A mathematical function or physical device is linear if its output for a complex input is just the sum of its outputs for the isolated components of that input.

**Local circuitry.** In the Magpie theory, the local circuitry consists of short-range excitatory connections and longer-range inhibitory connections between neurons within a single cortical layer and within a horizontal distance of about a millimetre.
Locus coeruleus. Nuclei in the brain stem containing neurons that send widespread fibres to many parts of the brain, including the cortex, and that release the neurotransmitter norepinephrine.

Long-term depression. Long-lasting (hours or more) decrease in the effectiveness of a connection in exciting a neuron. See long-term potentiation.

Long-term potentiation. Long-lasting (hours or more) increase in the effectiveness of a connection in exciting a neuron. Can generally be induced by high-frequency stimulation of the presynaptic fibres. There are different types of long-term potentiation in different parts of the brain, but a common type is a form of Hebbian plasticity in which a high presynaptic firing rate must be paired with strong postsynaptic depolarization for the potentiation to occur.

LTD. Long-term depression.

LTP. Long-term potentiation.

Luminance. The intensity of light reflected to the eye from a surface.

Macaque. A genus of Old World monkeys, including rhesus monkeys (Macaca mulatta), cynomolgus monkeys (Macaca fascicularis), and other species. Macaques have visual abilities very similar to ours.

Magnocellular layers. The two layers of the primate lateral geniculate nucleus in which neurons have relatively large somas. Part of the M channel.

Magpie theory. The eclectic theory of cortical function introduced in this thesis.

M channel. A visual channel in the primate’s visual system in which neurons have large receptive fields, are poorly selective for wavelength, give transient responses to maintained stimuli, and are very sensitive to moving stimuli and to stimuli of low contrast. See P channel.

Medial. Towards the plane that divides the cerebral hemispheres.

Membrane channel. Opening in the membrane of a neuron through which selected types of ion can flow, thereby affecting the membrane potential. Some channels are associated with receptors in synapses; such channels are normally closed, but open transiently when the associated receptors are activated by neurotransmitter.

Membrane potential. The electric potential across the membrane of a neuron. Affected by the flow of different types of ion through membrane channels. In the absence of inputs from synapses, the potential remains at an equilibrium level called the resting potential. Transient opening of channels at synapses causes the potential in the dendrites and soma to deviate from the resting level. If the membrane potential at the soma is raised above the resting level sufficiently to reach a certain threshold, an action potential is generated in the axon.

Modular architecture. A cognitive architecture consisting of a collection of processors that are fairly independent of one another, each processor dedicated to one faculty.

Modulation. Effect of a neurotransmitter on a neuron that is more subtle than simple excitation or inhibition, such as affecting the neuron’s response to other neurotransmitters.

Monoamine. The neurotransmitters dopamine, serotonin and norepinephrine.

Monosynaptic connection. A direct synaptic connection between two neurons, as opposed to an indirect connection involving one or more intermediate neurons.

Motor cortex. Cortex in the posterior part of the frontal lobe that is especially important in the control of movement.

MT. The middle temporal area of the visual cortex.

Myelinated axon. An axon with an insulating sheath, greatly increasing the speed at which action potentials are propagated.

NBM. Nucleus basalis of Meynert.
NE. **Norepinephrine.**

**Neocortex.** Six-layered part of the cerebral cortex, unique to mammals. Relative to total brain size it is very extensive in primates, especially humans. Usually just called cortex. (Figs. 2.1, 2.2.)

**Neuropsychology.** Study of neurological patients with the methods of experimental psychology, with the aim of elucidating the normal functions of the damaged or dysfunctioning brain structures.

**Neurotransmitter.** Substance used for signalling between neurons. See *synapse*.

**NMDA receptor.** A type of receptor for the excitatory neurotransmitter glutamate, named after its agonist N-methyl-D-aspartate. The membrane channel associated with the NMDA receptor is voltage-dependent: it opens only if the receptor is activated by neurotransmitter and the postsynaptic membrane potential is above a certain threshold. The channel allows calcium ions into the postsynaptic cell, which may trigger the events leading to long-term potentiation.

**Nonlinear.** See *linear*.

**Norepinephrine.** Neurotransmitter released by fibres that innervate the whole cortex and originate from neurons in the locus coeruleus.

**Nucleus.** A lump of grey matter.

**Nucleus basalis of Meynert.** A nucleus at the base of the forebrain containing neurons that send connections to the whole of the cortex and which release the neurotransmitter acetylcholine.

**Occipital lobe.** The region of cortex posterior to the parietal and temporal lobes, consisting entirely of visual cortex. (Fig. 2.1.)

**Ocular dominance.** Measure of the relative importance of the two eyes in stimulating a neuron in the visual cortex. In area V1, neurons dominated by the left eye and by the right eye are segregated in alternating slabs.

**OD.** *Ocular dominance.*

**OFF-response.** A type of response of a visual neuron such that the firing rate increases when luminance is reduced and decreases when luminance is increased. An OFF-subfield of a receptive field is a region in which OFF-responses can be elicited.

**ON-response.** A type of response of a visual neuron such that the firing rate increases when luminance is increased and decreases when luminance is reduced. An ON-subfield of a receptive field is a region in which ON-responses can be elicited.

**Parahippocampal cortex.** Cortex adjacent to the hippocampus, with structure that appears to be intermediate between allocortex and neocortex. Includes the presubiculum, parasubiculum, entorhinal cortex, and perirhinal cortex. Has reciprocal connections both with the hippocampus and with parts of the neocortex. (Fig. 2.2.)

**Parietal lobe.** Cortex between the frontal and occipital lobes. Consists of high-level visual cortex that appears to be more concerned with spatial relations than with object recognition, and also somatosensory cortex. (Fig. 2.1.)

**Parvocellular layers.** The four layers of the primate lateral geniculate nucleus in which neurons have relatively small somas. Part of the P channel.

**P channel.** A visual channel in the primate’s visual system in which neurons have small receptive fields, are selective for wavelength, give sustained responses to maintained stimuli, and are poorly sensitive to stimuli of low contrast. See *M channel*.

**Phenomenology.** Description of conscious experience.

**Photoreceptors.** Light-sensitive cells in the retina, the rods and cones.

**Plasticity.** See *connection plasticity.*
**Pons.** Structure in the brain stem that relays information from the cerebral cortex to the cerebellum. (Fig. 2.1.)

**Population coding.** A system of representation in which a single item is represented by a pattern of activity (e.g., firing rates) distributed over several neurons. Contrasts with a system in which a single item is represented by the activity of a single neuron.

**Posterior.** Towards the back.

**Postsynaptic.** The side of a synapse that receives the neurotransmitter.

**Prefrontal cortex.** Region of the frontal lobe anterior to the motor cortex. Very extensive in humans.

**Presynaptic.** The side of a synapse that releases the neurotransmitter.

**Psychophysics.** Study of the relation between objectively-measured sensory stimuli and the corresponding percepts.

**Pulvinar.** A large nucleus in the thalamus, connected with most of the visual cortex. Receives connections from the superior colliculus. (Fig. 2.7.)

**Putamen.** Part of the striatum in the basal ganglia. Receives connections from sensorimotor cortex and sends connections to parts of the globus pallidus and substantia nigra. (Fig. 2.2.)

**Pyramidal neuron.** The most numerous type of neuron in the cortex. Has a single large apical dendrite, a few smaller basal dendrites, and an axon that goes to another cortical area or another part of the nervous system. (Fig. 2.5C–E.)

**Raphe nuclei.** Nuclei in the brain stem containing neurons that send widespread fibres to many parts of the brain, including the cortex, and that release the neurotransmitter serotonin.

**RDS.** Random dot stereogram.

**Receptive field.** The region of the sensory periphery in which stimulation can affect the firing rate of a sensory neuron; for example, a region of the retina in the case of a visual neuron, or a region of the skin in the case of a cutaneous neuron.

**Receptors.** (1) Specialized sensory neurons that convert stimulus energy into electrical signals, such as the light-sensitive cells in the retina and pressure-sensitive cells in the skin. (2) Specialized molecules in the membranes of neurons, especially on the postsynaptic side of synapses, which cause certain chemical events to occur when they bind with molecules of neurotransmitter, such as opening associated membrane channels.

**Representation.** A state in a computer that can be taken to mean or stand for something, i.e., which can be given a semantic interpretation.

**Resting potential.** The equilibrium level of the membrane potential in the absence of synaptic inputs, typically about \(-75\) mV in cortical neurons.

**Retina.** Layer of photoreceptors and associated neural circuitry at the back of the eye. (Fig. 4.1.)

**Retinotopic map.** Organization found in many areas of the visual cortex such that the visual field is topographically mapped across the area. Nearby neurons in such an area have receptive fields covering nearby parts of the retina.

**Reversal potential.** For a given type of membrane channel, the reversal potential is the particular value of the membrane potential at which the flow of ions through the channel has no net effect on the potential.

**RF.** Receptive field.

**Rod.** Type of photoreceptor in the retina, much more sensitive than the cones. Responsible for vision in very poor light.

**Rostral.** Towards the front of the brain.

Serotonin. Neurotransmitter released by fibres that innervate the whole cortex and originate from neurons in the raphe nuclei.

Shape. The three-dimensional organization of the edges and surfaces of an object.

Shape fragment. Visual patterns, corresponding to commonly-occurring pieces of shape, assumed to be encoded as the characteristic patterns of many neurons in the higher-level areas of the visual cortex.

Sharp tuning. See tuning.

Short excitatory connections. See local circuitry.

Simple cell. Neuron in the visual cortex responding selectively to a contour of a particular orientation at a particular position in the receptive field. See complex cell.

Smooth neuron. Neuron with few or no dendritic spines. Such neurons generally have inhibitory effects in the cortex.

Soft weight. A connection weight that can be modified by connection plasticity.

Soma. The cell body of a neuron, as opposed to the dendrites and axon.

Somatosensory cortex. Cortex in the parietal lobe responsible for somesthesis.

Somatotopic map. Organization found in many areas of the somatosensory cortex such that the surface of the body is topographically mapped across the area. Nearby neurons in such an area have receptive fields covering nearby parts of the skin. (Fig. 3.15.)

Somesthesis. Perception of the body, including the senses of touch, temperature, pain, posture, and movement.

Spatial frequency. Loosely, the fine details of a visual pattern are its high-frequency components whereas coarse variations in luminance are its low-frequency components. See Fourier analysis.

Spectral composition. The amounts of different wavelengths in light.

Spike. Another term for action potential.

Spine. See dendritic spine.

Spiny stellate neuron. Type of excitatory neuron in certain areas of the cortex, with a dense coverage of dendritic spines and an axon that arborizes locally within the cortex. (Figs. 2.5B, 2.6F.)

Stabilized image. Image artificially projected onto the retina so that there is no movement in the image.

Stereopsis. The perception of depth and shape based on binocular disparity.

Strabismus. Misalignment of the eyes, such that the two retinal images are not in register.

Striatum. Part of the basal ganglia. Consists of the putamen and the caudate nucleus. (Fig. 2.2.)

Subcortical. Refers to the nervous system other than the cerebral cortex.

Substantia nigra. Part of the basal ganglia. Divided into the pars reticulata, which has similar connections to the globus pallidus, and the pars compacta, which contains neurons that send fibres to the striatum and to the cortex and which release the neurotransmitter dopamine. (Fig. 2.2.)

S-unit. In the Magpie theory, a neuron that keeps a relatively brief trace of the activity of its feedforward input synapses. See C-unit and Trace.

Superior colliculus. A nucleus in the brain stem that receives connections from the retina, visual cortex, and elsewhere. Important for visual-motor functions such as controlling the direction of gaze. (Figs. 2.1, 4.1.)
**Symbolist architecture.** A type of cognitive architecture in which the representations are sentence-like structures composed of discrete, word-like symbols. Inspired by conventional digital computers, many artificial intelligence systems are of this type.

**Synapse.** Junction between two neurons. Nearly all of the synapses in the cortex are chemical synapses between the axon of the presynaptic neuron and the dendrite or soma of the postsynaptic neuron. When an action potential arrives in the axon, a small amount of neurotransmitter is released which diffuses across the narrow cleft between the cells and comes into contact with receptors in the postsynaptic membrane. This causes membrane channels to open which affects the membrane potential. Excitatory synapses raise the potential, inhibitory synapses lower it.

**Tabula rasa.** The mind of a newborn infant considered to contain no innate knowledge.

**Tectum.** The superior colliculus and inferior colliculus.

**Temporal lobe.** The lateral part of the cerebral hemisphere, consisting of auditory cortex, high-level visual cortex involved in object-recognition, and cortex involved in memory. (Fig. 2.2.)

**Texton.** Term used to refer to an individual ‘grain’ of a visual texture with a limited number of significant properties. Two textures can only be discriminated easily if they are composed of different textons.

**Thalamus.** Large collection of nuclei in the forebrain. Nearly all the information reaching the cortex is relayed in the thalamus; for example, visual information from the retina passes through the lateral geniculate nucleus. Each nucleus sends connections to a restricted region of the cortex, and receives connections back from the same region. (Fig. 2.7, Table 2.1.)

**Threshold.** A level that a neuron’s membrane potential must exceed for some effect to occur, such as the generation of an action potential.

**Time constant.** Measure of the rate at which a continuous variable (such as membrane potential) can change value in time. The time constant of the membrane potential depends on the capacitance and conductance of the membrane.

**Trace.** In the Magpie theory, a signal associated with a synapse which is initiated by activation of the synapse and persists after presynaptic firing has ceased, and which is used by the postsynaptic cell in adapting the weight of the connection.

**Tract.** A bundle of fibres between two parts of the brain.

**Tuning.** The selectivity of a neuron’s response. A sharply tuned neuron responds with a high firing rate only if there is a comparatively close match between the input and its characteristic pattern. A broadly tuned neuron is less selective.

**Uniform architecture.** A type of cognitive architecture in which similar computational mechanisms are used in many different faculties. See heterogeneous architecture.

**V1, V2, V3, V4.** Areas of the visual cortex. V1 is the primary visual area that receives information directly from the lateral geniculate nucleus.

**Ventral.** Towards the bottom of the brain.

**Vergence movements.** Eye movements that change the focal distance without changing the average direction of gaze.

**Visual channels.** Parallel pathways from the retina to the visual cortex that are independent of one another and which carry different types of visual information. The primate visual system contains the M channel and the P channel, whereas cats and related species have the X channel and Y channel.

**Visual cortex.** Cortex responsible for visual perception. Occupies a large fraction of the cortex in primates, and consists of many different areas in the occipital, parietal and temporal lobes.
**Voltage-dependent channel.** A type of membrane channel that becomes more permeable to ions as the membrane potential becomes more depolarized. An example is the channel associated with the NMDA receptor.

**Von Neumann computer.** A digital computer with a single main processing unit and communication channels of very limited capacity. Nearly all the artificial computers in use today are von Neumann computers.

**W cells.** Ganglion cells in the retina of cats and related species that have functional properties different from those of cells in the X and Y channels. Most W cells project to the superior colliculus.

**Weight.** The potency of a connection in exciting or inhibiting a neuron.

**Wernicke’s area.** Region in the temporal lobe of the left hemisphere of the human cortex, important for language. Probably not a single area.

**White matter.** Brain tissue consisting mainly of axons, with few somas. See grey matter.

**X channel.** Visual channel in cats and related species in which neurons have small receptive fields, give sustained responses to a maintained stimulus, and are poorly sensitive to stimuli of low contrast.

**Y channel.** Visual channel in cats and related species in which neurons have large receptive fields, give transient responses to a maintained stimulus, are very sensitive to stimuli of low contrast, and have some complex nonlinear summation properties.
Bibliography


