

Neural Networks and Brain Function

Edmund T. Rolls

University of Oxford,
Department of Experimental Psychology,
Oxford OX1 3UD, England

and

Alessandro Treves

International School of Advanced Studies
Programme in Neuroscience
34013 Trieste, Italy

OXFORD
UNIVERSITY PRESS

1 Introduction

1.1 Introduction

It is now becoming possible to develop an understanding not only of *what* functions are performed by each region of the brain, but of *how* they are performed. To understand what function is performed by a given brain region requires evidence from the deficits that follow damage to that region; evidence of which brain regions are selectively activated during the performance of particular tasks, as shown by brain imaging, using techniques such as Positron Emission Tomography (PET), functional magnetic resonance imaging, and optical imaging; and evidence of what inputs reach that region, as shown by recording the activity of single neurons in that region. Understanding what function is performed by a given brain region also involves knowing what is being performed in the regions that provide inputs to a given region, and that receive outputs from the region being considered. Knowledge of the connective anatomy of the brain helps in this. This is the systems-level analysis of brain function. To understand *how* a brain region performs these functions involves quantitative neuroanatomy, to show how many inputs are received from each source by neurons in a given brain region, and where they terminate on a cell; studies of the rules and mechanisms (including their pharmacology) that determine synaptic connectivity and modifiability; analysis of the manner in which information is represented by neuronal responses in different brain regions; and computational approaches to *how* the system could operate. The latter, the computational approach, is essential to understanding how each part of the brain functions. Even if we had good systems-level evidence on the operation of a given brain region, and good evidence on what is represented in it, we would still not know how that brain region operated. For example, we would not know how many memories it could store, or how it might solve computationally difficult problems such as how objects can be recognized, despite presentation in different positions on the retina, in different sizes, and from different views. To understand these problems, some notion of how each area of the brain computes is needed. It is the aim of this book to provide an introduction to how different areas of the brain may perform their computations. In order to understand this, we need to take into account the nature of the computation performed by single neurons, how the connections between neurons alter in order to store information about the problem to be solved, how neurons in a region interact, and how these factors result in useful computations being performed.

To approach this aim, we describe a number of fundamental operations that can be

performed by networks in the brain. We take the approach of describing how networks that are biologically plausible operate. In Chapters 2–4 we describe networks that can learn to associate two inputs, that can store patterns (e.g. memories), and can recall the patterns from a fragment; and that can perform categorization of the inputs received, and thus perform feature analysis. In later chapters, we show how these principles can be applied to develop an understanding of how some regions of the brain operate.

Without an understanding of how the brain performs its computations, we will never understand *how* the brain works. Some of the first steps to this understanding are described in this book.

In the rest of this chapter, we introduce some of the background for understanding brain computation, such as how single neurons operate, how some of the essential features of this can be captured by simple formalisms, and some of the biological background to what it can be taken happens in the nervous system, such as synaptic modification based on information available locally at each synapse.

1.2 Neuronal network approaches versus connectionism

The approach taken in this book is to introduce how real neuronal networks in the brain may compute, and thus to achieve a fundamental and realistic basis for understanding brain function. This may be contrasted with connectionism, which aims to understand cognitive function by analysing processing in neuron-like computing systems. Connectionist systems are neuron-like in that they analyse computation in systems with large numbers of computing elements in which the information which governs how the network computes is stored in the connection strengths between the nodes (or 'neurons') in the network. However, in many connectionist models the individual units or nodes are not intended to model individual neurons, and the variables that are used in the simulations are not intended to correspond to quantities that can be measured in the real brain. Moreover, connectionist approaches use learning rules in which the synaptic modification (the strength of the connections between the nodes) is determined by algorithms which require information which is not local to the synapse, that is, evident in the pre- and postsynaptic firing rates (see further Chapter 5). Instead, in many connectionist systems, information about how to modify synaptic strengths is propagated backwards from the output of the network to affect neurons hidden deep within the network. Because it is not clear that this is biologically plausible, we have instead in this text concentrated on introducing neuronal network architectures which are more biologically plausible, and which use a local learning rule. Connectionist approaches (see e.g. McClelland and Rumelhart, 1986; McLeod, Plunkett and Rolls, 1998) are very valuable, for they show what can be achieved computationally with networks in which the connection strength determines the computation that the network achieves with quite simple computing elements. However, as models of brain function, many connectionist networks achieve almost too much, by solving problems with a carefully limited number of 'neurons' or nodes, which contributes to the ability of such networks to generalize successfully over the problem space. Connectionist schemes thus make an important start on understanding how complex computations (such as language) could be implemented in brain-like systems. In doing this, connectionist models often use simplified representations of the inputs and outputs, which are often crucial to the way in which the

problem is solved. In addition, they may use learning algorithms that are really too powerful for the brain to perform, and therefore they can be taken only as a guide to how cognitive functions might be implemented by neuronal networks in the brain. In this book, we focus on more biologically plausible neuronal networks.

1.3 Neurons in the brain, and their representation in neuronal networks

Neurons in the vertebrate brain typically have large dendrites extending from the cell body, which receive inputs from other neurons through connections called synapses. The synapses operate by chemical transmission. When a synaptic terminal receives an all-or-nothing action potential from the neuron of which it is a terminal, it releases a transmitter which crosses the synaptic cleft and produces either depolarization or hyperpolarization in the postsynaptic neuron, by opening particular ionic channels. (A textbook such as Kandel *et al.*, 1991 gives further information on this process.) Summation of a number of such depolarization or excitatory inputs within the time constant of the receiving neuron, which is typically 20–30 ms, produces sufficient depolarization that the neuron fires an action potential. There are often 5000–20 000 inputs per neuron. An example of a neuron found in the brain is shown in Fig. 1.1, and there are further examples in Chapter 10. Once firing is initiated in the cell body (or axon initial segment of the cell body), the action potential is conducted in an all-or-nothing way to reach the synaptic terminals of the neuron, whence it may affect other neurons. Any inputs the neuron receives which cause it to become hyperpolarized make it less likely to fire (because the membrane potential is moved away from the critical threshold at which an action potential is initiated), and are described as inhibitory. The neuron can thus be thought of in a simple way as a computational element which sums its inputs within its time constant and, whenever this sum, minus any inhibitory effects, exceeds a threshold, produces an action potential which propagates to all of its outputs. This simple idea is incorporated in many neuronal network models using a formalism of a type described in the next section.

1.4 A formalism for approaching the operation of single neurons in a network

Let us consider a neuron i as shown in Fig. 1.2 which receives inputs from axons which we label j through synapses of strength w_{ij} . The first subscript (i) refers to the receiving neuron, and the second subscript (j) to the particular input¹. j counts from 1 to C , where C is the number of synapses or connections received. The firing rate of the i th neuron is denoted r_i , and of the j th input to the neuron r'_j . (The prime is used to denote the input or presynaptic term. The letter r is used to indicate that the inputs and outputs of real neurons are firing rates.) To express the idea that the neuron makes a simple linear summation of the inputs it receives, we can write the activation of neuron i , denoted h_i , as

$$h_i = \sum_j r'_j w_{ij} \quad (1.1)$$

where \sum_j indicates that the sum is over the C input axons indexed by j . The multiplicative form here indicates that activation should be produced by an axon only if it is firing, and depending

¹ This convention, that i refers to the receiving neuron, and j refers to a particular input to that neuron via a synapse of weight w_{ij} , is used throughout this book, except where otherwise stated.

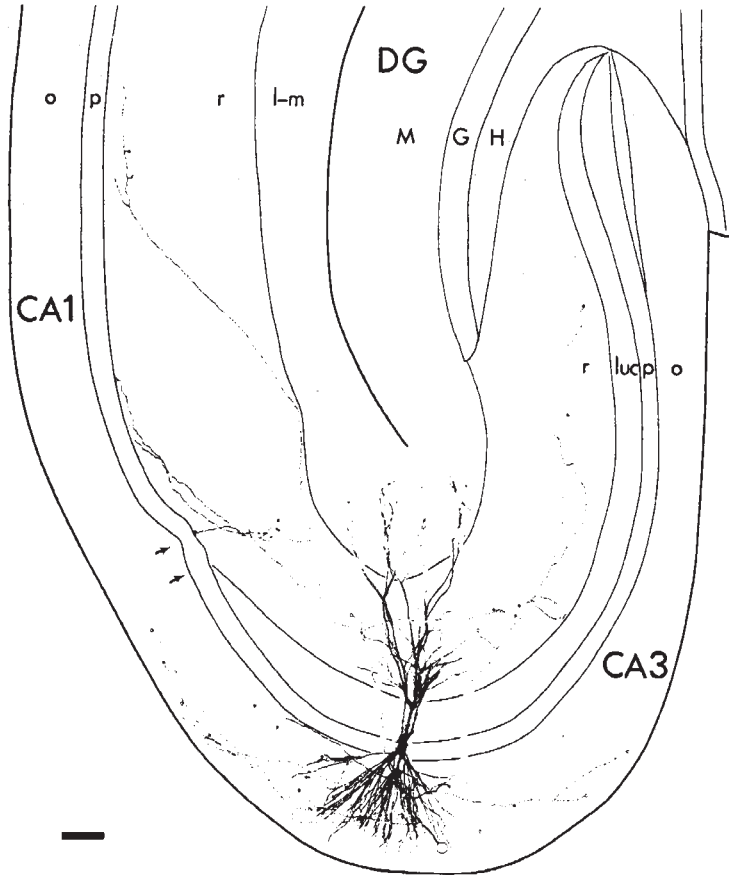


Fig. 1.1 An example of a real neuron found in the brain. The neuron is a CA3 neuron from the hippocampus. The thick extensions from the cell body or soma are the dendrites, which form an extensive dendritic tree receiving in this case approximately 12 000 synapses. The axon is the thin connection leaving the cell. It divides into a number of collateral branches. Two axonal branches can be seen in the plane of the section to travel to each end of the population of CA3 cells. One branch (on the left) continues to connect to the next group of cells, the CA1 cells. The junction between the CA3 and CA1 is shown by the two arrows. The diagram shows a camera lucida drawing of a single CA3 pyramidal cell intracellularly labelled with horseradish peroxidase. DG, dentate gyrus. The small letters refer to the different strata of the hippocampus. Scale bar = 100µm. (Reprinted with permission from Ishizuka, Weber and Amaral, 1990.)

on the strength of the synapse w_{ij} from input axon j onto the dendrite of the receiving neuron i . Equation 1.1 indicates that the strength of the activation reflects how fast the axon j is firing (that is r'_j), and how strong the synapse w_{ij} is. The sum of all such activations expresses the idea that summation (of synaptic currents in real neurons) occurs along the length of the dendrite, to produce activation at the cell body, where the activation h_i is converted into firing rate r_i . This conversion can be expressed as

$$r_i = f(h_i) \tag{1.2}$$

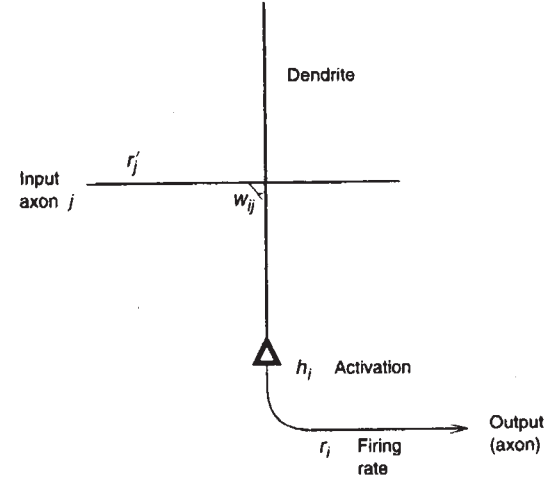


Fig. 1.2 Notation used to describe an individual neuron in a network model. By convention, we generally represent the dendrite as thick, and vertically oriented (as this is the normal way that neuroscientists view cortical pyramidal cells under the microscope), and the axon as thin. The cell body or soma is indicated between them. The firing rate we also call the (firing rate) activity of the neuron.

which indicates that the firing rate is a function of the postsynaptic activation. The function is called the activation function in this case. The function at its simplest could be linear, so that the firing rate would be proportional to the activation (see Fig. 1.3a). Real neurons have thresholds, with firing occurring only if the activation is above the threshold. A threshold linear activation function is shown in Fig. 1.3b. This has been useful in formal analysis of the properties of neural networks. Neurons also have firing rates which become saturated at a maximum rate, and we could express this as the sigmoid activation function shown in Fig. 1.3c. Another simple activation function, used in some models of neural networks, is the binary threshold function (Fig. 1.3d), which indicates that if the activation is below threshold, there is no firing, and that if the activation is above threshold, the neuron fires maximally. Some non-linearity in the activation function is an advantage, for it enables many useful computations to be performed in neuronal networks, including removing interfering effects of similar memories, and enabling neurons to perform logical operations, such as firing only if several inputs are present simultaneously.

A property implied by Eq. 1.1 is that the postsynaptic membrane is electrically short, and so summates its inputs irrespective of where on the dendrite the input is received. In real neurons, the transduction of current into firing frequency (the analogue of the transfer function of Eq. 1.2) is generally studied not with synaptic inputs but by applying a steady current through an electrode into the soma. An example of the resulting curves, which illustrate the additional phenomenon of firing rate adaptation, is reproduced in Fig. 1.3e.

1.5 Synaptic modification

For a neuronal network to perform useful computation, that is to produce a given output when it receives a particular input, the synaptic weights must be set up appropriately. This is often performed by synaptic modification occurring during learning.

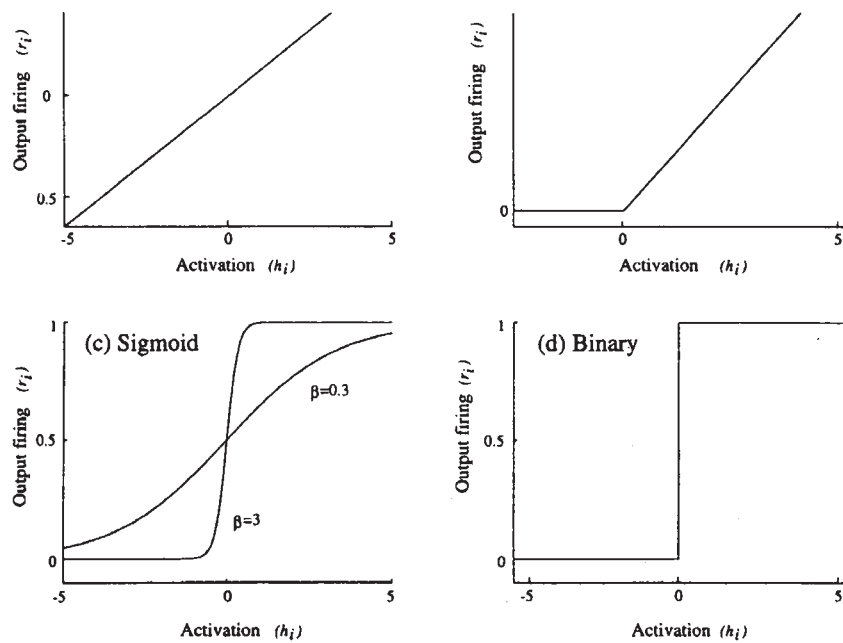


Fig. 1.3 Different types of activation function. The activation function relates the output activity (or firing rate), r_i , of the neuron (i) to its activation, h_i . (a) Linear. (b) Threshold linear. (c) Sigmoid. (One mathematical exemplar of this class of activation function is $r_i = 1 / (1 + \exp(-2\beta h_i))$. The output of this function, also sometimes known as the logistic function, is 0 for an input of $-\infty$, 0.5 for 0, and 1 for $+\infty$. The function incorporates a threshold at the lower end, followed by a linear portion, and then an asymptotic approach to the maximum value at the top end of the function. The parameter β controls the steepness of the almost linear part of the function round $h_i = 0$. If β is small, the output goes smoothly and slowly from 0 to 1 as h_i goes from $-\infty$ to $+\infty$. If β is large, the curve is very steep, and approximates a binary threshold activation function.) (d) Binary threshold.

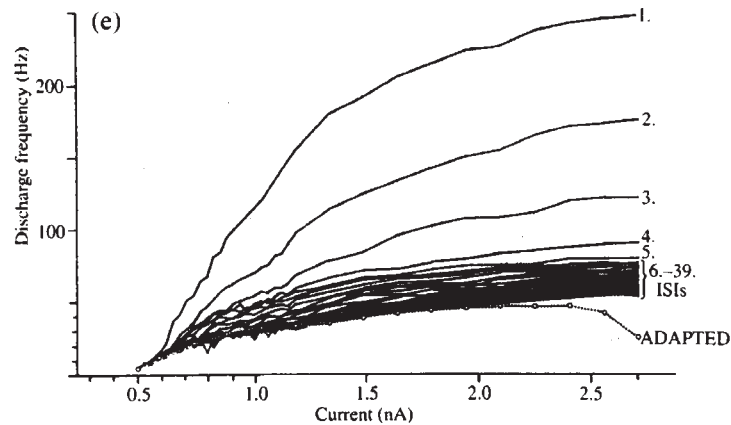


Fig. 1.3(e) Frequency-current plot (the closest experimental analogue of the activation function) for a CA1 pyramidal cell. The firing frequency (in Hz) in response to the injection of 1.5 s long, rectangular depolarizing current pulses has been plotted against the strength of the current pulses (in nA) (abscissa). The first 39 interspike intervals (ISIs) are plotted as instantaneous frequency ($1/ISI$, where ISI is the inter-stimulus interval), together with the average frequency of the adapted firing during the last part of the current injection (circles and broken line). The plot indicates a current threshold at approximately 0.5 nA, a linear range with a tendency to saturate, for the initial instantaneous rate, above approximately 200 Hz, and the phenomenon of adaptation, which is not reproduced in simple non-dynamical models (see further Appendix 5). (Reprinted with permission from Lanthorn, Storm and Andersen, 1984.)

A simple learning rule that was originally presaged by Donald Hebb (1949) proposes that synapses increase in strength when there is conjunctive presynaptic and postsynaptic activity. The Hebb rule can be expressed more formally as follows:

$$\delta w_{ij} = k r_i r'_j \quad (1.3)$$

where δw_{ij} is the change of the synaptic weight w_{ij} which results from the simultaneous (or conjunctive) presence of presynaptic firing r'_j and postsynaptic firing r_i (or strong depolarization), and k is a learning rate constant which specifies how much the synapses alter on any one pairing. The presynaptic and postsynaptic activity must be present approximately simultaneously (to within perhaps 100–500 ms in the real brain).

The Hebb rule is expressed in this multiplicative form to reflect the idea that *both* presynaptic and postsynaptic activity must be present for the synapses to increase in strength. The multiplicative form also reflects the idea that strong pre- and postsynaptic firing will produce a larger change of synaptic weight than smaller firing rates. The Hebb rule thus captures what is typically found in studies of associative long-term potentiation (LTP) in the brain, described in Section 1.6.

One useful property of large neurons in the brain, such as cortical pyramidal cells, is that with their short electrical length, the postsynaptic term, r_i , is available on much of the dendrite of a cell. The implication of this is that once sufficient postsynaptic activation has been produced, any active presynaptic terminal on the neuron will show synaptic strengthening. This enables associations between coactive inputs, or correlated activity in input axons, to be learned by neurons using this simple associative learning rule.

If, in contrast, a group of coactive axons made synapses close together on a small dendrite, then the local depolarization might be intense, and these synapses only would modify onto the dendrite. (A single distant active synapse might not modify in this type of neuron, because of the long electronic length of the dendrite.) The computation in this case is described as Sigma-Pi ($\Sigma\Pi$), to indicate that there is a local product computed during learning, this allows a particular set of locally active synapses to modify together, and then the output of the neuron can reflect the sum of such local multiplications (see Rumelhart and McClelland, 1986). This idea is not used in most neuronal networks that have been studied.

1.6 Long-term potentiation and long-term depression as biological models of synaptic modifications that occur in the brain

Long-term potentiation (LTP) and long-term depression (LTD) provide useful models of some of the synaptic modifications that occur in the brain. The synaptic changes found appear to be synapse-specific, and to depend on information available locally at the synapse. LTP and LTD may thus provide a good model of biological synaptic modification involved in real neuronal network operations in the brain. We next therefore describe some of the properties of LTP and LTD, and evidence which implicates them in learning in at least some brain systems. Even if they turn out not to be the basis for the synaptic modifications that occur during learning, they have many of the properties that would be needed by some of the synaptic modification systems used by the brain.

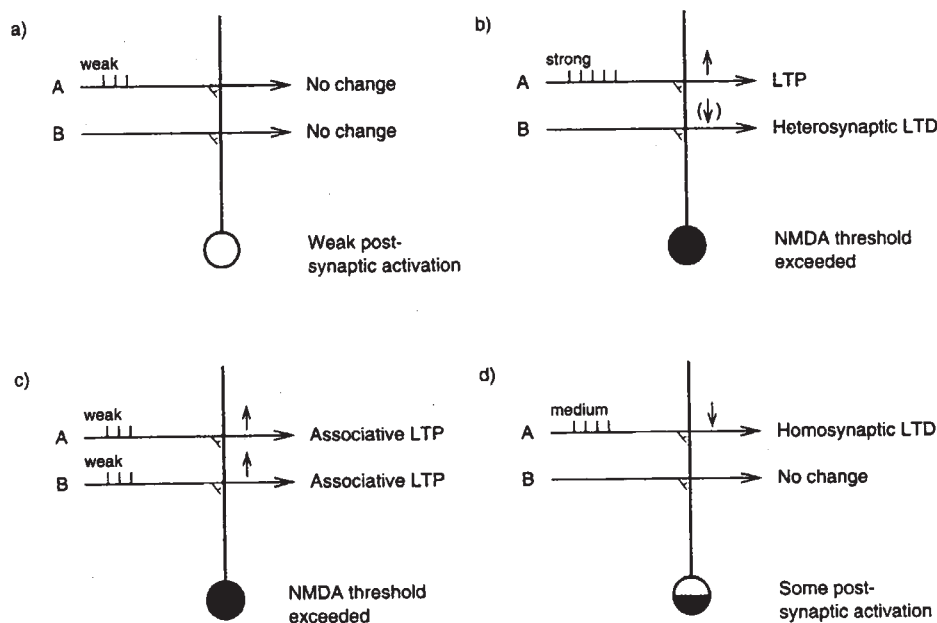


Fig. 1.4 Schematic illustration of synaptic modification rules as revealed by long-term potentiation (LTP) and long-term depression (LTD). The activation of the postsynaptic neuron is indicated by the extent to which its soma is black. There are two sets of inputs to the neuron: A and B. (a) A weak input (indicated by 3 spikes) on the set A of input axons produces little postsynaptic activation, and there is no change in synaptic strength. (b) A strong input (indicated by 5 spikes) on the set A of input axons produces strong postsynaptic activation, and the active synapses increase in strength. This is LTP. It is homosynaptic in that the synapses that increase in strength are the same as those through which the neuron is activated. LTP is synapse-specific, in that the inactive axons, B, do not show LTP. They either do not change in strength, or they may weaken. The weakening is called heterosynaptic LTD, because the synapses that weaken are other than those through which the neuron is activated (*hetero-* is Greek for other). (c) Two weak inputs present simultaneously on A and B summate to produce strong postsynaptic activation, and both sets of active synapses show LTP. (d) Intermediate strength firing on A produces some activation, but not strong activation, of the postsynaptic neuron. The active synapses become weaker. This is homosynaptic LTD, in that the synapses that weaken are the same as those through which the neuron is activated (*homo-* is Greek for same).

Long-term potentiation (LTP) is a use-dependent and sustained increase in synaptic strength that can be induced by brief periods of synaptic stimulation. It is usually measured as a sustained increase in the amplitude of electrically evoked responses in specific neural pathways following brief trains of high frequency stimulation (see Fig. 1.4b). For example, high frequency stimulation of the Schaffer collateral inputs to the hippocampal CA1 cells results in a larger response recorded from the CA1 cells to single test pulse stimulation of the pathway. LTP is *long-lasting*, in that its effect can be measured for hours in hippocampal slices, and in chronic *in vivo* experiments in some cases may last for months. LTP becomes evident rapidly, typically in less than 1 minute. LTP is in some brain systems *associative*. This is illustrated in Fig. 1.4c, in which a weak input to a group of cells (e.g. the commissural input to CA1) does not show LTP unless it is given at the same time as (i.e. associatively with) another input (which could be weak or strong) to the cells. The associativity arises because it is only when sufficient activation of the postsynaptic neuron to exceed the threshold of NMDA receptors (see below) is produced that any learning can occur. The two weak inputs

summate to produce sufficient depolarization to exceed the threshold. This associative property is shown very clearly in experiments in which LTP of an input to a single cell only occurs if the cell membrane is depolarized by passing current through it at the same time as the input arrives at the cell. The depolarization alone or the input alone is not sufficient to produce the LTP, and the LTP is thus associative. Moreover, in that the presynaptic input and the postsynaptic depolarization must occur at about the same time (within approximately 500 ms), the LTP requires **temporal contiguity**. LTP is also **synapse-specific**, in that for example an inactive input to a cell does not show LTP even if the cell is strongly activated by other inputs (Fig. 1.4b, input B).

These spatiotemporal properties of LTP can be understood in terms of actions of the inputs on the postsynaptic cell, which in the hippocampus has two classes of receptor, NMDA (*N*-methyl-D-aspartate) and K-Q (kainate-quisqualate), activated by the glutamate released by the presynaptic terminals. Now the NMDA receptor channels are normally blocked by Mg^{2+} , but when the cell is strongly depolarized by strong tetanic stimulation of the type necessary to induce LTP, the Mg^{2+} block is removed, and Ca^{2+} entering via the NMDA receptor channels triggers events that lead to the potentiated synaptic transmission (see Fig. 1.5). Part of the evidence for this is that NMDA antagonists such as AP5 (*D*-2-amino-5-phosphonopentanoate) block LTP. Further, if the postsynaptic membrane is voltage clamped to prevent depolarization by a strong input, then LTP does not occur. The voltage-dependence of the NMDA receptor channels introduces a threshold and thus a non-linearity which contributes to a number of the phenomena of some types of LTP, such as cooperativity (many small inputs together produce sufficient depolarization to allow the NMDA receptors to operate), associativity (a weak input alone will not produce sufficient depolarization of the postsynaptic cell to enable the NMDA receptors to be activated, but the depolarization will be sufficient if there is also a strong input), and temporal contiguity between the different inputs that show LTP (in that if inputs occur non-conjunctively, the depolarization shows insufficient summation to reach the required level, or some of the inputs may arrive when the depolarization has decayed). Once the LTP has become established (which can be within one minute of the strong input to the cell), the LTP is expressed through the K-Q receptors, in that AP5 blocks only the establishment of LTP, and not its subsequent expression (see further Bliss and Collingridge, 1993; Nicoll and Malenka, 1995; Fazeli and Collingridge, 1996).

There are a number of possibilities about what change is triggered by the entry of Ca^{2+} to the postsynaptic cell to mediate LTP. One possibility is that somehow a messenger reaches the presynaptic terminals from the postsynaptic membrane and, if the terminals are active, causes them to release more transmitter in future whenever they are activated by an action potential. Consistent with this possibility is the observation that after LTP has been induced, more transmitter appears to be released from the presynaptic endings. Another possibility is that the postsynaptic membrane changes just where Ca^{2+} has entered, so that K-Q receptors become more responsive to glutamate released in future. Consistent with this possibility is the observation that after LTP, the postsynaptic cell may respond more to locally applied glutamate (using a microiontophoretic technique).

The rule which underlies associative LTP is thus that synapses connecting two neurons become stronger if there is conjunctive presynaptic and (strong) postsynaptic activity. This

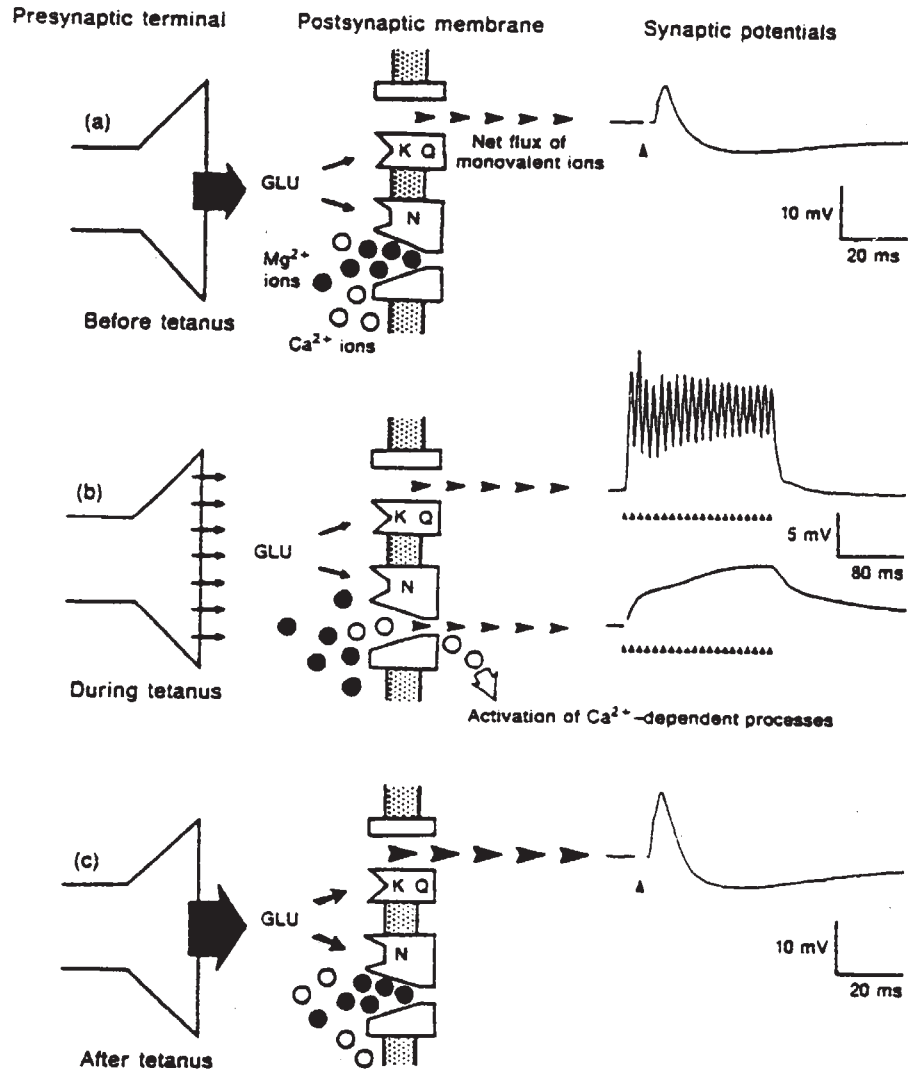


Fig. 1.5 The mechanism of induction of LTP in the CA1 region of the hippocampus. (a) Neurotransmitter (e.g. L-glutamate) is released and acts upon both K/Q (kainate/quisqualate) and NMDA (N) receptors. The NMDA receptors are blocked by magnesium and the excitatory synaptic response (EPSP) is therefore mediated primarily by ion flow through the channels associated with K/Q receptors. (b) During high-frequency activation, the magnesium block of the ion channels associated with NMDA receptors is released by depolarization. Activation of the NMDA receptor by transmitter now results in ions moving through the channel. In this way, calcium enters the postsynaptic region to trigger various intracellular mechanisms which eventually result in an alteration of synaptic efficacy. (c) Subsequent low-frequency stimulation results in a greater EPSP. See text for further details. (Reprinted with permission from Collingridge and Bliss, 1987.)

learning rule for synaptic modification is sometimes called the Hebb rule, after Donald Hebb of McGill University who drew attention to this possibility, and its potential importance in learning, in 1949.

In that LTP is long-lasting, develops rapidly, is synapse-specific, and is in some cases associative, it is of interest as a potential synaptic mechanism underlying some forms of memory. Evidence linking it directly to some forms of learning comes from experiments in which it has been shown that the drug AP5 infused so that it reaches the hippocampus to block NMDA receptors blocks spatial learning mediated by the hippocampus (see Morris, 1989). The task learned by the rats was to find the location relative to cues in a room of a platform submerged in an opaque liquid (milk). Interestingly, if the rats had already learned where the platform was, then the NMDA infusion did not block performance of the task. This is a close parallel to LTP, in that the learning, but not the subsequent expression of what had been learned, was blocked by the NMDA antagonist AP5. Although there is still some uncertainty about the experimental evidence that links LTP to learning (see e.g. Morris, 1996), there is a need for a synapse-specific modifiability of synaptic strengths on neurons if neuronal networks are to learn (see Rolls, 1996a and examples throughout this book), and if LTP is not always an exact model of the synaptic modification that occurs during learning, then something with many of the properties of LTP is nevertheless needed, and is likely to be present in the brain given the functions known to be implemented in many brain regions (see Chapters 6–10).

In another model of the role of LTP in memory, Davis (1992) has studied the role of the amygdala in learning associations to fear-inducing stimuli. He has shown that blockade of NMDA synapses in the amygdala interferes with this type of learning, consistent with the idea that LTP provides a useful model of this type of learning too (see further Chapter 7).

Long-term depression (LTD) can also occur. It can in principle be associative or non-associative. In associative LTD, the alteration of synaptic strength depends on the pre- and postsynaptic activities. There are two types. Heterosynaptic LTD occurs when the postsynaptic neuron is strongly activated, and there is low presynaptic activity (see Fig. 1.4b input B, and Table 2.1). Heterosynaptic LTD is so-called because the synapse that weakens is other than (hetero-) the one through which the postsynaptic neuron is activated. Heterosynaptic LTD is important in associative neuronal networks (see Chapters 2 and 3), and in competitive neuronal networks (see Chapter 4). In competitive neural networks it would be helpful if the degree of heterosynaptic LTD depended on the existing strength of the synapse, and there is some evidence that this may be the case (see Chapter 4). Homosynaptic LTD occurs when the presynaptic neuron is strongly active, and the postsynaptic neuron has some, but low, activity (see Fig. 1.4d and Table 2.1). Homosynaptic LTD is so-called because the synapse that weakens is the same as (homo-) the one that is active. Heterosynaptic and homosynaptic LTD are found in the neocortex (Artola and Singer, 1993; Singer, 1995) and hippocampus (Christie, 1996 and other papers in *Hippocampus* (1996) 6(1)), and in many cases are dependent on activation of NMDA receptors (see also Fazeli and Collingridge, 1996). LTD in the cerebellum is evident as weakening of active parallel fibre to Purkinje cell synapses when the climbing fibre connecting to a Purkinje cell is active (see Ito, 1984, 1989, 1993a,b).

1.7 Distributed representations

When considering the operation of many neuronal networks in the brain, it is found that many useful properties arise if each input to the network (arriving on the axons, r') is encoded in the activity of an ensemble or population of the axons or input lines (distributed encoding), and is not signalled by the activity of a single input, which is called local encoding. We start off with some definitions, and then highlight some of the differences, and summarize some evidence which shows the type of encoding used in some brain regions. Then in Chapter 2 (e.g. Table 2.2) on, we show how many of the useful properties of the neuronal networks described depend on distributed encoding. In Chapter 10, we review evidence on the encoding actually found in different regions of the cerebral cortex.

1.7.1 Definitions

A *local* representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of one of the neurons. In a famous example, a single neuron might be active only if one's grandmother was being seen. An implication is that most neurons in the brain regions where objects or events are represented would fire only very rarely. A problem with this type of encoding is that a new neuron would be needed for every object or event that has to be represented. There are many other disadvantages of this type of encoding, many of which will become apparent in this book. Moreover, there is evidence that objects are represented in the brain by a different type of encoding.

A *fully distributed* representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of the full set of neurons. If the neurons are binary (e.g. either active or not), the most distributed encoding is when half the neurons are active for any one stimulus or event.

A *sparse distributed* representation is a distributed representation in which a small proportion of the neurons is active at any one time. In a sparse representation with binary neurons, less than half of the neurons are active for any one stimulus or event. For binary neurons, we can use as a measure of the sparseness the proportion of neurons in the active state. For neurons with real, continuously variable, values of firing rates, the sparseness of the representation a can be measured, by extending the binary notion of the proportion of neurons that are firing, as

$$a = (\sum_{i=1,N} r_i / N)^2 / \sum_{i=1,N} (r_i^2 / N) \quad (1.4)$$

where r_i is the firing rate of the i th neuron in the set of N neurons (Treves and Rolls, 1991).

Coarse coding utilizes overlaps of receptive fields, and can compute positions in the input space using differences between the firing levels of coactive cells (e.g. colour-tuned cones in the retina). The representation implied is distributed. **Fine coding** (in which for example a neuron may be 'tuned' to the exact orientation and position of a stimulus) implies more local coding.

1.7.2 Advantages of different types of coding

One advantage of distributed encoding is that the similarity between two representations can be reflected by the correlation between the two patterns of activity which represent the different stimuli. We have already introduced the idea that the input to a neuron is represented by the activity of its set of input axons r'_j , where j indexes the axons, numbered from $j = 1, C$ (see Fig. 1.2 and Eq. 1.1). Now the set of activities of the input axons is a vector (a vector is an ordered set of numbers; Appendix 1 provides a summary of some of the concepts involved). We can denote as r^1 the vector of axonal activity that represents stimulus 1, and r^2 the vector that represents stimulus 2. Then the similarity between the two vectors, and thus the two stimuli, is reflected by the correlation between the two vectors. The correlation will be high if the activity of each axon in the two representations is similar; and will become more and more different as the activity of more and more of the axons differs in the two representations. Thus the similarity of two inputs can be represented in a graded or continuous way if (this type of) distributed encoding is used. This enables generalization to similar stimuli, or to incomplete versions of a stimulus (if it is for example partly seen or partly remembered), to occur. With a local representation, either one stimulus or another is represented, and similarities between different stimuli are not encoded.

Another advantage of distributed encoding is that the number of different stimuli that can be represented by a set of C components (e.g. the activity of C axons) can be very large. A simple example is provided by the binary encoding of an 8-element vector. One component can code for which of two stimuli has been seen, 2 components (or bits in a computer byte) for 4 stimuli, 3 components for 8 stimuli, 8 components for 256 stimuli, etc. That is, the number of stimuli increases exponentially with the number of components (or in this case, axons) in the representation. (In this simple binary illustrative case, the number of stimuli that can be encoded is 2^C .) Put the other way round, even if a neuron has only a limited number of inputs (e.g. a few thousand), it can nevertheless receive a great deal of information about which stimulus was present. This ability of a neuron with a limited number of inputs to receive information about which of potentially very many input events is present is probably one factor that makes computation by the brain possible. With local encoding, the number of stimuli that can be encoded increases only linearly with the number C of axons or components (because a different component is needed to represent each new stimulus). (In our example, only 8 stimuli could be represented by 8 axons.)

In the real brain, there is now good evidence that in a number of brain systems, including the high-order visual and olfactory cortices, and the hippocampus, distributed encoding with the above two properties, of representing similarity, and of exponentially increasing encoding capacity as the number of neurons in the representation increases, is found (Rolls and Tovee, 1995a; Abbott, Rolls and Tovee, 1996; Rolls, Treves and Tovee, 1997; Rolls, Treves, Robertson, Georges-François and Panzeri, 1998). For example, in the high-order visual cortex in the temporal lobe of the primate brain, the number of faces that can be represented increases approximately exponentially with the number of neurons in the population (see Fig. 1.6a). If we plot instead the information about which stimulus is seen, we see that this rises approximately linearly with the number of neurons in the representation (Fig. 1.6b). This corresponds to an exponential rise in the number of stimuli encoded, because information is a

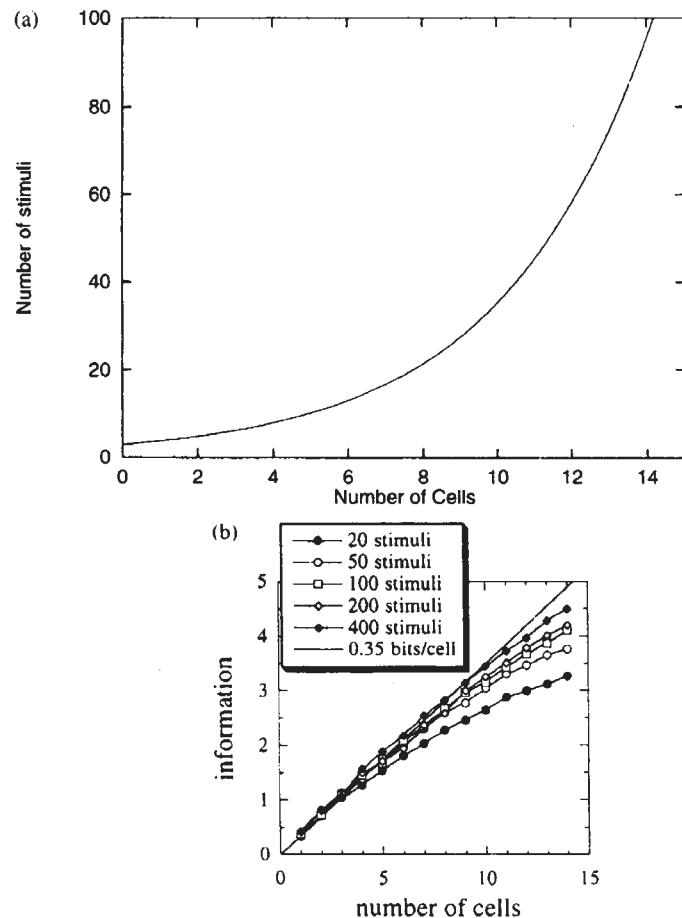


Fig. 1.6 (a) The number of stimuli (in this case from a set of 20 faces) that are encoded in the responses of different numbers of neurons in the temporal lobe visual cortex (after Rolls, Treves and Tovee, 1997; Abbott, Rolls and Tovee, 1996; see Section 10.4.3 and Appendix A2.) (b) The amount of information (in bits) available from the responses of different numbers of temporal cortex neurons about which of 20 faces has been seen (from Abbott, Rolls and Tovee, 1996). In (b), it is shown by simulation that for numbers of stimuli greater than 20, the information rises more and more linearly as the number of cells in the sample increases (see Appendix A2). In both (a) and (b), the analysis time was 500 ms.

log measure (see Appendix A2). A similar result has been found for the encoding of position in space by the primate hippocampus (Rolls, Treves, Robertson, Georges-François and Panzeri, 1998). It is particularly important that the information can be read from the ensemble of neurons using a simple measure of the similarity of vectors, the correlation (or dot product, see Appendix 1) between two vectors. The importance of this is that it is essentially vector similarity operations that characterize the operation of many neuronal networks (see e.g. Chapters 2–4, and Appendix 1). The neurophysiological results show that both the ability to reflect similarity by vector correlation, and the utilization of exponential coding capacity, are a property of real neuronal networks found in the brain.

To emphasize one of the points being made here, although the binary encoding used in the 8-bit vector described above has optimal capacity for binary encoding, it is not optimal for vector similarity operations. For example, the two very similar numbers 127 and 128 are represented by 01111111 and 10000000 with binary encoding, yet the correlation or bit overlap of these vectors is 0. The brain in contrast uses a code which has the attractive property of exponentially increasing capacity with the number of neurons in the representation, though it is different from the simple binary encoding of numbers used in computers; and at the same time codes stimuli in such a way that the code can be read off with simple dot product or correlation-related decoding, which is what is specified for the elementary neuronal network operation shown in Eq. 1.1.

1.8 Introduction to three simple neuronal network architectures

With neurons of the type outlined in Section 1.4, and an associative learning rule of the type described in Section 1.5, three neuronal network architectures arise which appear to be used in many different brain regions. The three architectures will be described in Chapters 2–4, and a brief introduction is provided here.

In the first architecture (see Fig. 1.7a,b), pattern associations can be learned. The output neurons are driven by an unconditioned stimulus. A conditioned stimulus reaches the output neurons by associatively modifiable synapses w_{ij} . If the conditioned stimulus is paired during learning with activation of the output neurons produced by the unconditioned stimulus, then later, after learning, due to the associative synaptic modification, the conditioned stimulus alone will produce the same output as the conditioned stimulus. Pattern associators are described in Chapter 2.

In the second architecture, the output neurons have recurrent associatively modifiable synaptic connections w_{ij} to other neurons in the network (see Fig. 1.7c). When an external input causes the output neurons to fire, then associative links are formed through the modifiable synapses that connect the set of neurons that is active. Later, if only a fraction of the original input pattern is presented, then the associative synaptic connections or weights allow the whole of the memory to be retrieved. This is called completion. Because the components of the pattern are associated with each other as a result of the associatively modifiable recurrent connections, this is called an autoassociative memory. It is believed to be used in the brain for many purposes, including episodic memory in which the parts of a memory of an episode are associated together, and helping to define the response properties of cortical neurons, which have collaterals between themselves within a limited region.

In the third architecture, the main input to the output neurons is received through associatively modifiable synapses w_{ij} (see Fig. 1.7d). Because of the initial values of the synaptic strengths, or because every axon does not contact every output neuron, different patterns tend to activate different output neurons. When one pattern is being presented, the most strongly activated neurons tend via lateral inhibition to inhibit the other neurons. For this reason the network is called competitive. During the presentation of that pattern, associative modification of the active axons onto the active postsynaptic neuron takes place. Later, that or similar patterns will have a greater chance of activating that neuron or set of neurons. Other neurons learn to respond to other input patterns. In this way, a network is built which can

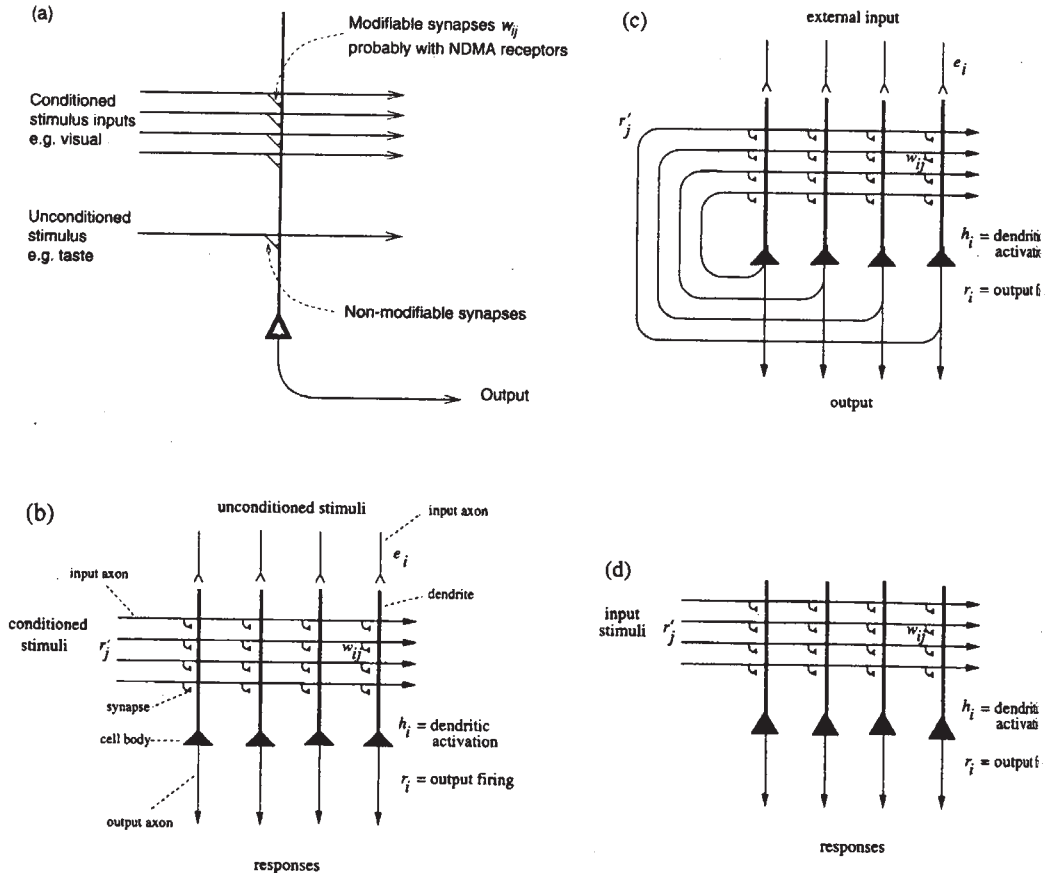


Fig. 1.7 Three network architectures that use local learning rules: (a) Pattern association introduced with a single output neuron; (b) Pattern association network; (c) Autoassociation network; (d) Competitive network.

categorize patterns, placing similar patterns into the same category. This is useful as a preprocessor for sensory information, and finds use in many other parts of the brain too.

These are three fundamental building blocks for neural architectures in the brain. They are often used in combination with each other. Because they are some of the building blocks of some of the architectures found in the brain, they are described in Chapters 2–4.

1.9 Systems-level analysis of brain function

To understand the neuronal network operations of any one brain region, it is useful to have an idea of the systems-level organization of the brain, in order to understand how the networks in each region provide a particular computational function as part of an overall computational scheme.

Some of the pathways followed by sensory inputs to reach memory systems and eventually motor outputs are shown in Fig. 1.8. Some of these regions are shown in the drawings of the primate brain in Figs 1.9–1.13. Each of these routes is described in turn. The description is based primarily on studies in non-human primates, for they have well-developed cortical areas which in many cases correspond to those found in humans, and it has been possible to analyse their connectivity and their functions by recording the activity of neurons in them. Further evidence on these systems is provided in Section 10.5.

Information on the ventral visual cortical processing stream projects after the primary visual cortex, area V1, to the secondary visual cortex (V2), and then via area V4 to the posterior and then to the anterior inferior temporal visual cortex (lower stream in Fig. 1.8;

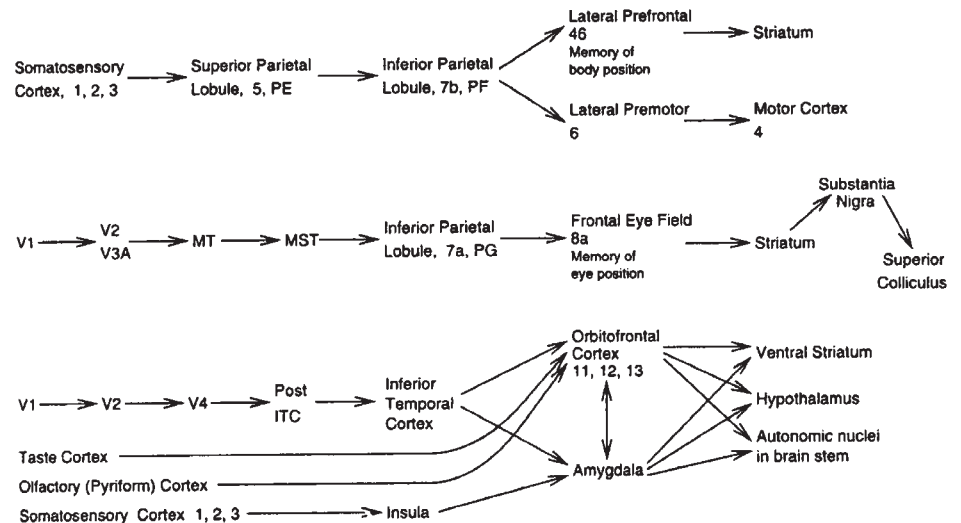


Fig. 1.8 Schematic diagram showing the major connection of three separate processing systems in the brain (see text). The top pathway, also shown in Fig. 1.13 on a lateral view of the macaque brain, shows the connections from the primary somatosensory cortex, areas 1, 2 and 3, via area 5 in the parietal cortex, to area 7b. The middle pathway, also shown in Fig 1.11, shows the connections in the 'dorsal visual system' from V1 to V2, MST, etc., with some connections reaching the frontal eye fields. The lower pathway, also shown in Fig. 1.9, shows the connections in the 'ventral visual system' from V1 to V4, the inferior temporal visual cortex, etc., with some connections reaching the amygdala and orbitofrontal cortex.

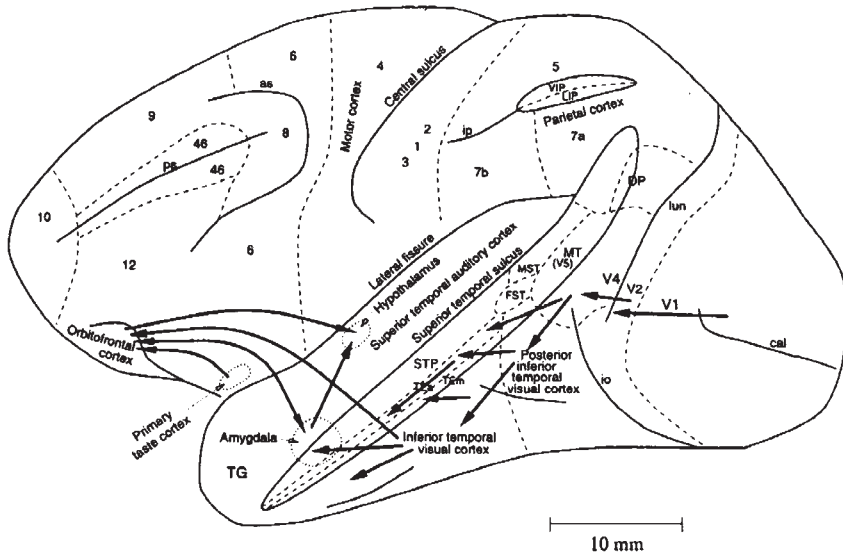


Fig. 1.9 Lateral view of the macaque brain showing the connections in the 'ventral visual system' from V1, V2 and V4, the inferior temporal visual cortex, etc., with some connections reaching the amygdala and orbitofrontal cortex. as, arcuate sulcus; cal, calcarine sulcus; cs, central sulcus; lf, lateral (or Sylvian) fissure; lun, lunate sulcus; ps, principal sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus (which has been opened to reveal some of the areas it contains); sts, superior temporal sulcus (which has been opened to reveal some of the areas it contains). AIT, anterior inferior temporal cortex; FST, visual motion processing area; LIP, lateral intraparietal area; MST, visual motion processing area; MT, visual motion processing area (also called VS); PIT, posterior inferior temporal cortex; STP, superior temporal plane; TA, architectonic area including auditory association cortex; TE, architectonic area including high order visual association cortex, and some of its subareas TEa and TE_m; TG, architectonic area in the temporal pole; V1-V4, visual areas 1-4; VIP, ventral intraparietal area; TEO, architectonic area including posterior visual association cortex. The numerals refer to architectonic areas, and have the following approximate functional equivalence: 1, 2, 3, somatosensory cortex (posterior to the central sulcus); 4, motor cortex; 5, superior parietal lobule; 7a, inferior parietal lobule, visual part; 7b, inferior parietal lobule, somatosensory part; 6, lateral premotor cortex; 8, frontal eye field; 12, part of orbitofrontal cortex; 46, dorsolateral prefrontal cortex.

Fig. 1.9; Fig. 1.12). Information processing along this stream is primarily unimodal, as shown by the fact that inputs from other modalities (such as taste or smell) do not anatomically have significant inputs to these regions, and by the fact that neurons in these areas respond primarily to visual stimuli, and not to taste or olfactory stimuli, etc. (see Rolls, 1997d; Baylis, Rolls and Leonard, 1987; Ungerleider, 1995). The representation built along this pathway is mainly about what object is being viewed, independently of exactly where it is on the retina, of its size, and even of the angle with which it is viewed (Rolls, 1994a, 1995a, 1997d). The representation is also independent of whether the object is associated with reward or punishment, that is the representation is about objects *per se* (Rolls *et al.*, 1977). The computation which must be performed along this stream is thus primarily to build a representation of objects which shows invariance. After this processing, the visual representation is interfaced to other sensory systems in areas in which simple associations must be learned between stimuli in different modalities (see Fig. 1.10). The representation must thus be in a form in which the simple generalization properties of associative networks can be useful. Given that the association is about which object is present (and not where it is on the retina), the representation computed in sensory systems must be in a form which allows the simple correlations computed by associative

networks to reflect similarities between objects, and not between their positions on the retina. The way in which such invariant sensory representations could be built in the brain is the subject of Chapter 8. A similar mainly unimodal analysis of taste inputs is carried out by the primary taste cortex, to represent what taste is present, independently of the pleasantness or aversiveness of the taste (Rolls, 1995b, 1997c).

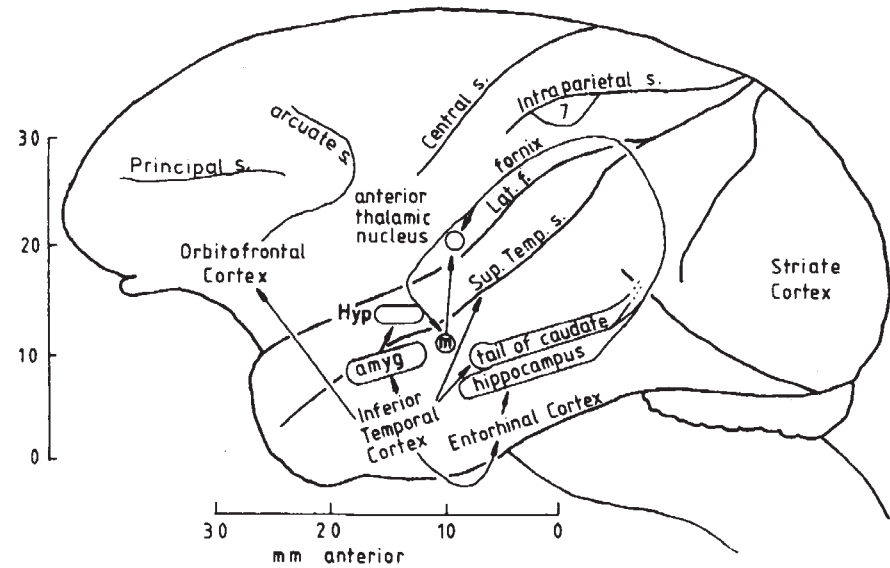


Fig. 1.10 Lateral view of the primate (rhesus monkey) brain showing some of the brain regions described. This diagram shows some of the outputs of the inferior temporal visual cortex. Abbreviations: Amyg, amygdala; central s, central sulcus; Hyp, hypothalamus/substantia innominata/basal forebrain; Lat f, lateral (or Sylvian) fissure; m, mammillary body; Sup Temp s, superior temporal sulcus; 7, posterior parietal cortex, area 7.

After mainly unimodal processing stages, these information processing streams converge together into a number of areas, particularly the amygdala and orbitofrontal cortex (see Figs 1.8, 1.9 and 1.10). These areas appear to be necessary for learning to associate sensory stimuli with other reinforcing (rewarding or punishing) stimuli. For example, the amygdala is involved in learning associations between the sight of food and its taste. (The taste is a primary or innate reinforcer.) The orbitofrontal cortex is especially involved in rapidly relearning these associations, when environmental contingencies change (see Rolls, 1990a, 1995b, 1996b; Chapter 7). They thus are brain regions in which the computation at least includes simple pattern association (e.g. between the sight of an object and its taste). In the orbitofrontal cortex, this association learning is also used to produce a representation of flavour, in that neurons are found in the orbitofrontal cortex which are activated by both olfactory and taste stimuli (Rolls and Baylis, 1994), and in that the neuronal responses in this region reflect in some cases olfactory to taste association learning (Rolls, Critchley, Mason and Wakeman, 1996; Critchley and Rolls, 1996b). In these regions too, the representation is concerned not only with what sensory stimulus is present, but for some neurons, with its hedonic or reward-related properties, which are often

computed by association with stimuli in other modalities. For example, many of the visual neurons in the orbitofrontal cortex respond to the sight of food only when hunger is present. This probably occurs because the visual inputs here have been associated with a taste input, which itself in this region only occurs to a food if hunger is present, that is when the taste is rewarding (see Chapter 7 and Rolls, 1993, 1994b, 1995b, 1996b). The outputs from these associative memory systems, the amygdala and orbitofrontal cortex, project onwards to structures such as the hypothalamus, through which they control autonomic and endocrine responses such as salivation and insulin release to the sight of food; and to the striatum, including the ventral striatum, through which behaviour to learned reinforcing stimuli is produced. The striatal output pathways for the cerebral cortex are described in Chapter 9, on motor systems. Somatosensory (touch) inputs also reach the amygdala and orbitofrontal cortex, via projections from the somatosensory cortical areas (S1, 2 and 3) to the insula (see Fig. 1.8).

Another processing stream shown in Figs 1.8, 1.11 and 1.12 is that from V1 to MT, MST and thus to the parietal cortex (see Ungerleider, 1995; Ungerleider and Haxby, 1994; Section 10.5). This 'where' pathway for primate vision is involved in representing where stimuli are relative to the animal, and the motion of these stimuli. Neurons here respond for example to stimuli in visual space around the animal, including the distance from the observer, and also respond to optic flow or to moving stimuli. The outputs of this system control eye movements to visual stimuli (both slow pursuit and saccadic eye movements). One output of these regions is to the frontal eye fields which are important as a short term memory for where fixation should occur next, as shown by the effects of lesions to the frontal eye fields on saccades to

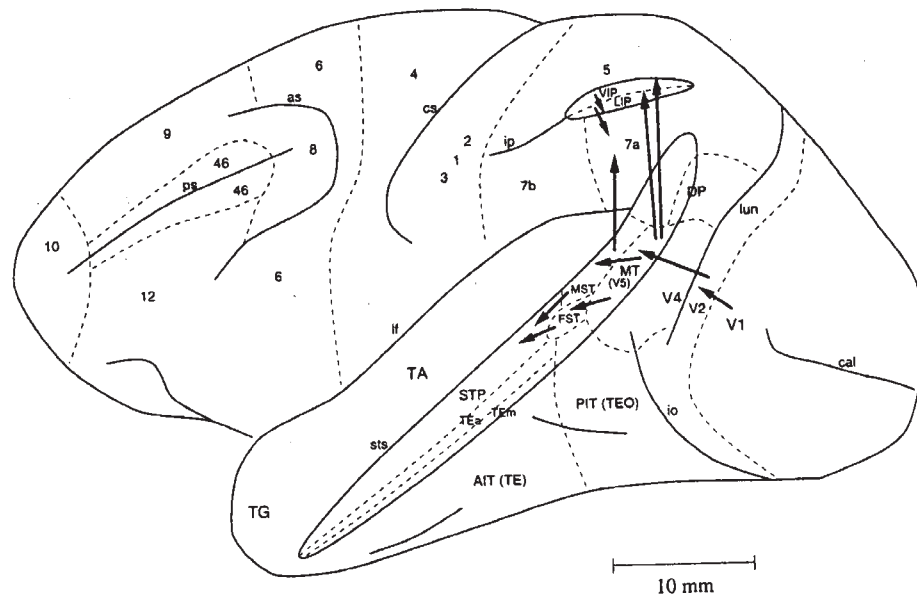


Fig. 1.11 Lateral view of the macaque brain showing the connections in the 'dorsal visual system' from V1 to V2, MST, etc. with some connections reaching the frontal eye fields. Abbreviations as in Fig. 1.9.

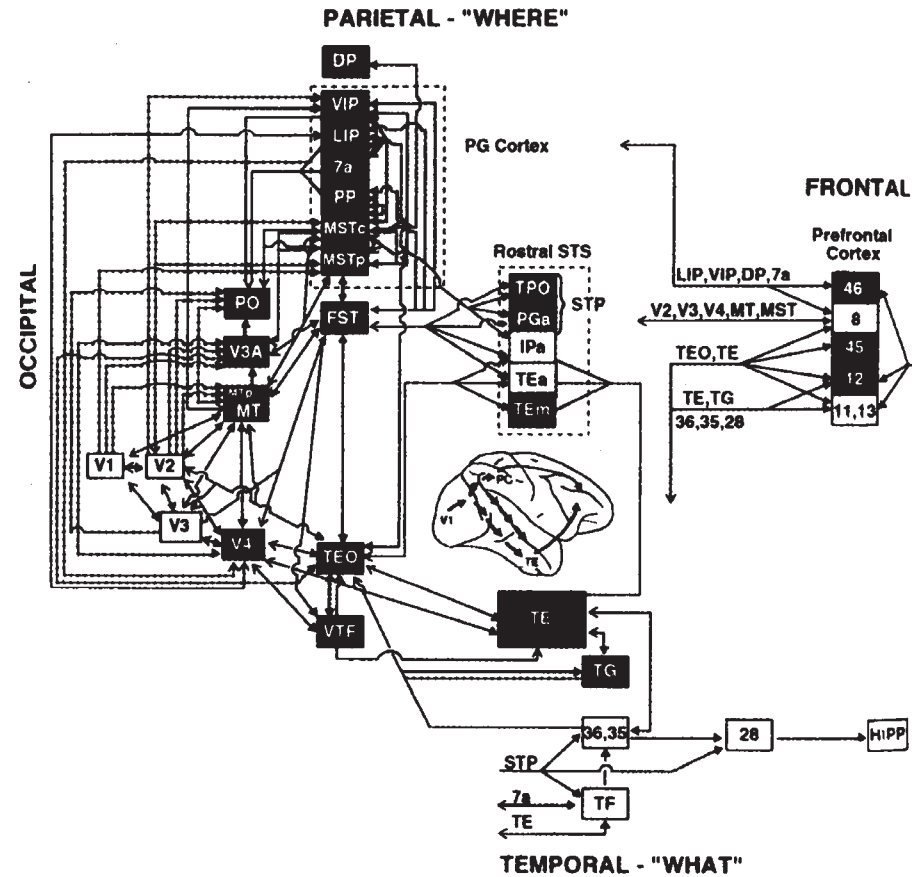


Fig. 1.12 Visual processing pathways in monkeys. Solid lines indicate connections arising from both central and peripheral visual field representations; dotted lines indicate connections restricted to peripheral visual field representations. Shaded boxes in the ventral (lower) stream indicate visual areas related primarily to object vision; shaded boxes in the dorsal stream indicate areas related primarily to spatial vision; and white boxes indicate areas not clearly allied with only one stream. The shaded region on the lateral view of the brain represents the extent of the cortex included in the diagram. Abbreviations: DP, dorsal prefrontal area; LIP, lateral intraparietal area; PP, posterior parietal area; MSTc, medial superior temporal area, central visual field representation; MSTp, medial superior temporal area, peripheral visual field representation; MT, middle temporal area; MTp, middle temporal area, peripheral visual field representation; PO, parieto-occipital area; V1, primary visual cortex; V2, visual area 2; V3, visual area 3; V3A, visual area 3, part A; V4, visual area 4; and VIP, ventral intraparietal area. Inferior parietal area 7a; prefrontal areas 8, 11 to 13, 45 and 46 are from Brodmann (1925). Inferior temporal areas TE and TEO, parahippocampal area TF, temporal pole area TG, and inferior parietal area PG are from Von Bonin and Bailey (1947). Rostral superior temporal sulcal (STS) areas are from Seltzer and Pandya (1978) and VTF is the visually responsive portion of area TF (Boussaoud, Desimone and Ungerleider, 1991). (Reprinted with permission from Ungerleider, 1995.)

remembered targets, and by neuronal activity in this region (see Section 10.5.2). Outputs from the frontal eye fields again reach the striatum, and then progress through the basal ganglia (that is, via the substantia nigra) to reach the superior colliculus.

A related processing stream is that from the somatosensory cortical areas to parietal cortex

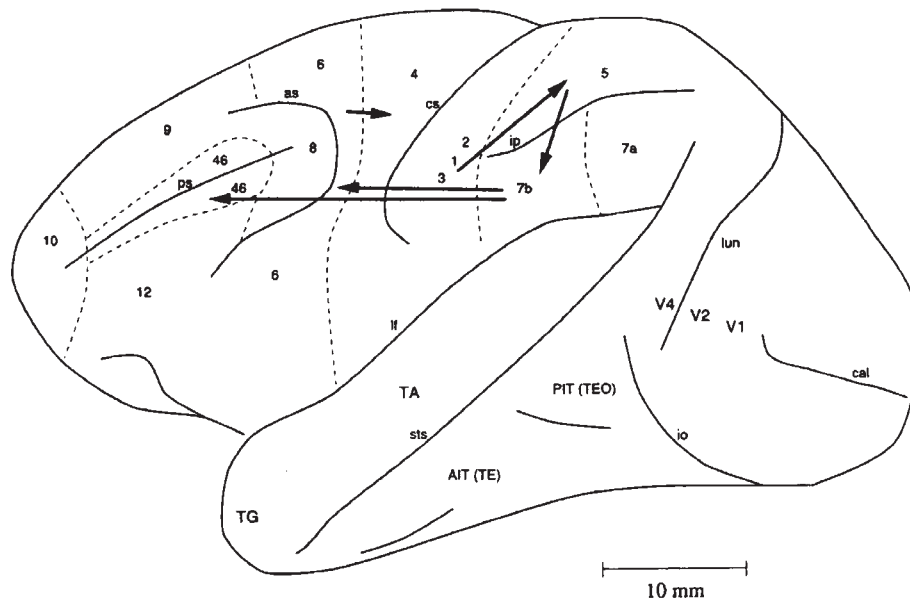


Fig. 1.13 Lateral view of the macaque brain showing the connections from the primary somatosensory cortex, areas 1, 2 and 3, via area 5 in the parietal cortex, to area 7b. Abbreviations as in Fig. 1.9.

area 5, and thus to parietal cortex area 7b (see Figs 1.8 and 1.13, and Section 10.5). Along this pathway more complex representations are formed of objects touched (Iwamura, 1993), and of where the limbs are in relation to the body (by combining information from different joint proprioceptors). In area 7, some neurons respond to visual and to related somatosensory stimuli. Outputs from the parietal cortex project to the premotor areas and to the basal ganglia, which both provide routes to behavioural output. In addition, the parietal cortex projects to the dorsolateral prefrontal cortex, which provides a short term or working memory for limb responses, as shown by the effects of lesions to the dorsolateral prefrontal cortex, and by recordings of neuronal activity in it during delayed response tasks (Goldman-Rakic, 1996). The dorsolateral prefrontal cortex can influence behaviour through basal ganglia outputs, and through outputs to premotor areas.

The hippocampus receives inputs from both the 'what' and the 'where' systems (see Chapter 6 and Fig. 1.12). By rapidly learning associations between conjunctive inputs in these systems, it is able to form memories of particular events occurring in particular places at particular times. To do this, it needs to store whatever is being represented in each of many cortical areas at a given time, and to later recall the whole memory from a part of it. The types of network it contains which are involved in this simple memory function are described in Chapter 6.

With this overview of some of the main processing streams in the cerebral cortex, it is now time to consider in Chapters 2–5 the operation of some fundamental types of biologically plausible network. Then in Chapter 6–10 we will consider how these networks may contribute to the particular functions being performed by different brain regions.

2 Pattern association memory

A fundamental operation of most nervous systems is to learn to associate a first stimulus with a second which occurs at about the same time, and to retrieve the second stimulus when the first is presented. The first stimulus might be the sight of food, and the second stimulus the taste of food. After the association has been learned, the sight of food would enable its taste to be retrieved. In classical conditioning, the taste of food might elicit an unconditioned response of salivation, and if the sight of the food is paired with its taste, then the sight of that food would by learning come to produce salivation. More abstractly, if one idea is associated by learning with a second, then when the first idea occurs again, the second idea will tend to be associatively retrieved.

2.1 Architecture and operation

The essential elements necessary for pattern association, forming what could be called a prototypical pattern associator network, are shown in Fig. 2.1. What we have called the second or unconditioned stimulus pattern is applied through unmodifiable synapses generating an input to each unit which, being external with respect to the synaptic matrix we focus on, we can call the external input e_i for the i th neuron. (We can also treat this as a vector, e , as indicated in the legend to Fig. 2.1. Vectors and simple operations performed with them are summarized in Appendix A1). This unconditioned stimulus is dominant in producing or forcing the firing of the output neurons (r_i for the i th neuron, or the vector r). At the same time, the first or conditioned stimulus pattern r'_j for the j th axon (or equivalently the vector r') present on the horizontally running axons in Fig. 2.1 is applied through *modifiable* synapses w_{ij} to the dendrites of the output neurons. The synapses are modifiable in such a way that if there is presynaptic firing on an input axon r'_j paired during learning with postsynaptic activity on neuron i , then the strength or weight w_{ij} between that axon and the dendrite increases. This simple learning rule is often called the Hebb rule, after Donald Hebb who in 1949 formulated the hypothesis that if the firing of one neuron was regularly associated with another, then the strength of the synapse or synapses between the neurons should increase in strength. (In fact, the terms in which Hebb put the hypothesis were a little different from an association memory, in that he stated that if one neuron regularly comes to elicit firing in another, then the strength of the synapses should increase. He had in mind the building of what he called cell assemblies. In a pattern associator, the conditioned stimulus need not produce before learning any significant activation of the output neurons.