Many scientists, and many others, are interested in how the brain works. In order to understand this, we need to know what computations are performed by different brain systems; and how they are computed by each of these systems.

The aim of this book is to elucidate what is computed in different brain systems; and to describe current computational approaches and models of how each of these brain systems computes.

To understand how our brains work, it is essential to know what is computed in each part of the brain. That can be addressed by utilising evidence relevant to computation from many areas of neuroscience. Knowledge of the connections between different brain areas is important, for this shows that the brain is organised as systems, with whole series of brain areas devoted for example to visual processing. That provides a foundation for examining the computation performed by each brain area, by comparing what is represented in a brain area with what is represented in the preceding and following brain area, using techniques of for example neurophysiology and functional neuroimaging. Neurophysiology at the single neuron level is needed because this is the level at which information is transmitted between the computing elements of the brain, the neurons. Evidence from the effects of brain damage, including that available from neuropsychology, is needed to help understand what different parts of the system do, and indeed what each part is necessary for. Functional neuroimaging is useful to indicate where in the human brain different processes take place, and to show which functions can be dissociated from each other. So for each brain system, evidence on what is computed at each stage, and what the system as a whole computes, is essential.

To understand how our brains work, it is also essential to know how each part of the brain computes. That requires a knowledge of what is computed by each part of the brain, but it also requires knowledge of the network properties of each brain region. This involves knowledge of the connectivity between the neurons in each part of the brain, and knowledge of the synaptic and biophysical properties of the neurons. It also requires knowledge of the theory of what can be computed by networks with defined connectivity.

There are at least three key goals of the approaches described here. One is to understand ourselves better, and how we work and think. A second is to be better able to treat the system when it has problems, for example in mental illnesses. Medical applications are a very important aim of the type of research described here. A third, is to be able to emulate the operation of parts of our brains, which some in the field of artificial intelligence (AI) would like to do to produce useful machines. All of these goals require, and cannot get off the ground, without a firm foundation in what is computed by brain systems, and theories and models of how it is computed. To understand the operation of the whole brain, it is necessary to show how the different brain systems operate together: but a necessary foundation for this is to know what is computed in each brain system.

Part of the enterprise here is to stimulate new theories and models of how parts of the brain work. The evidence on what is computed in different brain systems had advanced rapidly in the last 50 years, and provides a reasonable foundation for the enterprise, though there is much that remains to be learned. Theories of how the computation is performed are less advanced, but progress is being made, and current models are described in this book for many brain systems, in the expectation that before further advances are made, knowledge of
the considerable current evidence on how the brain computes provides a useful starting point, especially as current theories do take into account the limitations that are likely to be imposed by the neural architectures present in our brains.

The simplest way to define brain computation is to examine what information is represented at each stage of processing, and how this is different from stage to stage. For example in the primary visual cortex (V1), neurons respond to simple stimuli such as bars or edges or gratings and have small receptive fields. Little can be read off from the firing rates about for example whose face is represented from a small number of neurons in V1. On the other hand, after four or five stages of processing, in the inferior temporal cortex, information can be read from the firing rates of neurons about whose face is being viewed, and indeed there is remarkable invariance with respect to the position, size, contrast and even in some cases view of the face. That is a major computation, and indicates what can be achieved by neural computation.

These approaches can only be taken to understand brain function because there is considerable localization of function in the brain, quite unlike a digital computer. One fundamental reason for localization of function in the brain is that this minimizes the total length of the connections between neurons, and thus brain size. Another is that it simplifies the genetic information that has to be provided in order to build the brain, because the connectivity instructions can refer considerably to local connections. These points are developed in my book Cerebral Cortex: Principles of Operation (Rolls, 2016b).

That brings me to what is different about the present book and Cerebral Cortex: Principles of Operation (Rolls, 2016b). The previous book took on the enormous task of making progress with understanding how the major part of our brains, the cerebral cortex, works, by understanding its principles of operation. The present book builds on that approach, and uses it as background, but has the different aim of taking each of our brain systems, and describing what they compute, and then what is known about how each system computes. The issue of how they compute relies for many brain systems on how the cortex operates, so Cerebral Cortex: Principles of Operation provides an important complement to the present book.

With its focus on what and how each brain system computes, a field that includes computational neuroscience, this book is distinct from the many excellent books on neuroscience that describe much evidence about brain structure and function, but do not aim to provide an understanding of how the brain works at the computational level. This book aims to forge an understanding of how some key brain systems may operate at the computational level, so that we can understand how the brain actually performs some of its complex and necessarily computational functions in memory, perception, attention, decision-making, cognitive functions, and actions.

Indeed, as one of the key aims of this book is to describe what computations are performed by different brain systems, I have chosen to include in this book some of the key discoveries in neuroscience that I believe help to define what computations are performed in different brain systems.

That makes this book very different from many of the textbooks of neuroscience (such as Principles of Neuroscience (Kandel et al., 2013)), and some of the textbooks of theoretical neuroscience that describe principles of operation of neurons or of networks of neurons (Dayan and Abbott, 2001; Hertz et al., 1991; Gerstner et al., 2014), but not in general what is computed in different brain systems, and how it is computed. Further, there are likely to be great developments in our understanding of how the brain computes, and this book is intended to set out a framework for new developments, by providing an analysis of what is computed by different brain systems, and providing some current approaches to how these computations may be performed.
A test of whether one’s understanding is correct is to simulate the processing on a computer, and to show whether the simulation can perform the tasks performed by the brain, and whether the simulation has similar properties to the real brain. The approach of neural computation leads to a precise definition of how the computation is performed, and to precise and quantitative tests of the theories produced. How memory systems in the brain work is a paradigm example of this approach, because memory-like operations which involve altered functionality as a result of synaptic modification are at the heart of how many computations in the brain are performed. It happens that attention and decision-making can be understood in terms of interactions between and fundamental operations in memory systems in the cortex, and therefore it is natural to treat these areas of cognitive neuroscience in this book. The same fundamental concepts based on the operation of neuronal circuitry can be applied to all these functions, as is shown in this book.

One of the distinctive properties of this book is that it links the neural computation approach not only firmly to neuronal neurophysiology, which provides much of the primary data about how the brain operates, but also to psychophysical studies (for example of attention); to neuropsychological studies of patients with brain damage; and to functional magnetic resonance imaging (fMRI) (and other neuroimaging) approaches. The empirical evidence that is brought to bear is largely from non-human primates and from humans, because of the considerable similarity of their cortical systems, and the major differences in their systems-level computational organization from that of rodents, as set out in Section 19.10.

The overall aims of the book are developed further, and the plan of the book is described, in Chapter 1. Appendix B describes the fundamental operation of key networks of the type that are likely to be the building blocks of brain function. Appendix C describes quantitative, information theoretic, approaches to how information is represented in the brain, which is an essential framework for understanding what is computed in a brain system, and how it is computed. Appendix D describes Matlab software that has been made available with this book to provide simple demonstrations of the operation of some key neuronal networks related to cortical function; to show how the information represented by neurons can be measured; and to provide a tutorial version of the VisNet program for invariant visual object recognition described in Chapter 2. The neural networks programs are also provided in Python. The programs are available at https://www.oxcns.org.

Part of the material described in the book reflects work performed in collaboration with many colleagues, whose tremendous contributions are warmly appreciated. The contributions of many will be evident from the references cited in the text. Especial appreciation is due to Alessandro Treves, Gustavo Deco, and Simon M. Stringer, who have contributed greatly in always interesting and fruitful research collaborations on computational aspects of brain function, and to many neurophysiology and functional neuroimaging colleagues who have contributed to the empirical discoveries that provide the foundation to which the computational neuroscience must always be closely linked, and whose names are cited throughout the text. Charl Ning (University of Warwick) is thanked for help with translating the Matlab neural network programs described in Appendix D into Python. Professor Lorraine Tyler (University of Cambridge) is thanked for guidance to some of the literature on language in the brain. Dr Patrick Mills is thanked for very helpful comments on an earlier version of this book. Much of the work described would not have been possible without financial support from a number of sources, particularly the Medical Research Council of the UK, the Human Frontier Science Program, the Wellcome Trust, and the James S. McDonnell Foundation. I am also grateful to many colleagues who I have consulted while writing this book. The book was typeset by the author using LATEX and WinEdt.

The cover shows on the left a human brain with a lateral view above and a medial view below. The numbers refer to Brodmann areas, and a summary of the functions of each area
is provided in Section 1.11 together with in Fig.1.9 a more fully labelled version of these images. These images refer to one of the aims of this book, to describe what computations are performed in each brain area and especially in brain systems that consist of a set of closely connected brain areas. It should be noted that there is not an exact correspondence between Brodmann areas and computationally relevant brain areas, and an aim of this book is to delineate more exactly the brain’s computational systems, and the areas and subareas that are involved in each computational system. (The images are from Purves,D., Augustine,G.J., Fitzpatrick,D. et al., editors (2019) Neuroscience. International Edition. © Oxford University Press: Oxford.) The different neural networks on the right refer to some of the types of biologically plausible neuronal network architectures that are important in how the computations are performed in different cortical areas. These networks are introduced in Section 1.9, and described computationally in Appendix B.

Updates to and .pdfs of many of the publications cited in this book are available at https://www.oxcns.org. Updates and corrections to the text and notes are also available at https://www.oxcns.org.

I dedicate this work to the overlapping group: my family, friends, and colleagues – in salutem praesentium, in memoriam absentium.
Contents

1 Introduction 1
  1.1 What and how the brain computes: introduction 1
  1.2 What and how the brain computes: plan of the book 3
  1.3 Neurons 5
  1.4 Neurons in a network 6
  1.5 Synaptic modification 8
  1.6 Long-term potentiation and long-term depression 10
  1.7 Information encoding by neurons, and distributed representations 14
    1.7.1 Definitions 16
    1.7.2 Advantages of different types of coding 16
  1.8 Neuronal network approaches versus connectionism 18
  1.9 Introduction to three neuronal network architectures 18
  1.10 Systems-level analysis of brain function 21
  1.11 Brodmann areas 23
  1.12 The fine structure of the cerebral neocortex 26
    1.12.1 The fine structure and connectivity of the neocortex 27
    1.12.2 Excitatory cells and connections 27
    1.12.3 Inhibitory cells and connections 29
    1.12.4 Quantitative aspects of cortical architecture 32
    1.12.5 Functional pathways through the cortical layers 34
    1.12.6 The scale of lateral excitatory and inhibitory effects, and modules 38

2 The ventral visual system 40
  2.1 Introduction and overview 40
    2.1.1 Introduction 40
    2.1.2 Overview of what is computed in the ventral visual system 40
    2.1.3 Overview of how computations are performed in the ventral visual system 43
    2.1.4 What is computed in the ventral visual system is unimodal, and is related to other 'what' systems after the inferior temporal visual cortex 45
  2.2 What: V1 – primary visual cortex 47
  2.3 What: V2 and V4 – intermediate processing areas in the ventral visual system 48
  2.4 What: Invariant representations of faces and objects in the inferior temporal visual cortex 49
    2.4.1 Reward value is not represented in the ventral visual system 49
    2.4.2 Translation invariant representations 50
    2.4.3 Reduced translation invariance in natural scenes 50
    2.4.4 Size and spatial frequency invariance 53
    2.4.5 Combinations of features in the correct spatial configuration 54
    2.4.6 A view-invariant representation 54
    2.4.7 Learning in the inferior temporal cortex 58
    2.4.8 A sparse distributed representation is what is computed in the ventral visual system 60
    2.4.9 Face expression, gesture, and view 66
    2.4.10 Specialized regions in the temporal cortical visual areas 66
### Contents

2.5 How the computations are performed: approaches to invariant object recognition 71
   2.5.1 Feature spaces 72
   2.5.2 Structural descriptions and syntactic pattern recognition 73
   2.5.3 Template matching and the alignment approach 75
   2.5.4 Invertible networks that can reconstruct their inputs 76
   2.5.5 Deep learning 76
   2.5.6 Feature hierarchies 77

2.6 Hypotheses about how the computations are performed in a feature hierarchy approach 82

2.7 VisNet: a model of how the computations are performed in the ventral visual system 86
   2.7.1 The architecture of VisNet 87
   2.7.2 Initial experiments with VisNet 96
   2.7.3 The optimal parameters for the temporal trace used in the learning rule 102
   2.7.4 Different forms of the trace learning rule, and error correction 104
   2.7.5 The issue of feature binding, and a solution 112
   2.7.6 Operation in a cluttered environment 125
   2.7.7 Learning 3D transforms 132
   2.7.8 Capacity of the architecture, and an attractor implementation 137
   2.7.9 Vision in natural scenes – effects of background versus attention 140
   2.7.10 The representation of multiple objects in a scene 148
   2.7.11 Learning invariant representations using spatial continuity 150
   2.7.12 Lighting invariance 151
   2.7.13 Deformation-invariant object recognition 153
   2.7.14 Learning invariant representations of scenes and places 154
   2.7.15 Finding and recognising objects in natural scenes 156
   2.7.16 Non-accidental properties, and transform invariant object recognition 159

2.8 Further approaches to invariant object recognition 161
   2.8.1 Other types of slow learning 161
   2.8.2 HMAX 161
   2.8.3 Hierarchical convolutional deep neural networks 166
   2.8.4 Sigma-Pi synapses 167

2.9 Visuo-spatial scratchpad memory, and change blindness 167

2.10 Processes involved in object identification 169

2.11 Top-down attentional modulation is implemented by biased competition 170

2.12 Highlights on how the computations are performed in the ventral visual system 173

3 The dorsal visual system 176

3.1 Introduction, and overview of the dorsal cortical visual stream 176

3.2 Global motion in the dorsal visual system 177

3.3 Invariant object-based motion in the dorsal visual system 179

3.4 What is computed in the dorsal visual system: visual coordinate transforms 181
   3.4.1 The transform from retinal to head-based coordinates 182
   3.4.2 The transform from head-based to allocentric bearing coordinates 183
   3.4.3 A transform from allocentric bearing coordinates to allocentric spatial view coordinates 184

3.5 How visual coordinate transforms are computed in the dorsal visual system 185
   3.5.1 Gain modulation 185
   3.5.2 Mechanisms of gain modulation using a trace learning rule 186
   3.5.3 Gain modulation by eye position to produce a head-centered representation in Layer 1 of VisNetCT 187
   3.5.4 Gain modulation by head direction to produce an allocentric bearing to a landmark in Layer 2 of VisNetCT 188
   3.5.5 Gain modulation by place to produce an allocentric spatial view representation in Layer 3 of VisNetCT 189
   3.5.6 The utility of the coordinate transforms in the dorsal visual system 190
The taste and flavour system

4.1 Introduction and overview

4.2 Taste and related pathways: what is computed

4.3 Taste and related pathways: how the computations are performed

The olfactory system

5.1 Introduction

5.2 What is computed in the olfactory system

5.3 How computations are performed in the olfactory system

The somatosensory system

6.1 What is computed in the somatosensory system

6.2 The anterior somatosensory cortex, areas 1, 2, 3a, and 3b, in the anterior parietal cortex

6.3 The ventral somatosensory stream: areas S2 and PV, in the lateral parietal cortex

6.4 The dorsal somatosensory stream to area 5 and then 7b, in the posterior parietal cortex
6.1.5 Somatosensory representations in the insula 236
6.1.6 Somatosensory and temperature inputs to the orbitofrontal cortex, affective value, pleasant touch, and pain 236
6.1.7 Decision-making in the somatosensory system 240

6.2 How computations are performed in the somatosensory system 242
6.2.1 Hierarchical computation in the somatosensory system 242
6.2.2 Computations for pleasant touch and pain 243
6.2.3 The mechanisms for somatosensory decision-making 243

7 The auditory system 244
7.1 Introduction, and overview of computations in the auditory system 244
7.2 Auditory Localization 245
7.3 Ventral and dorsal cortical auditory pathways 248
7.4 The ventral cortical auditory stream 249
7.5 The dorsal cortical auditory stream 251
7.6 How the computations are performed in the auditory system 251

8 The temporal cortex 253
8.1 Introduction and overview 253
8.2 Middle temporal gyrus and face expression and gesture 253
8.3 Semantic representations in the temporal lobe neocortex 255
8.3.1 Neurophysiology 255
8.3.2 Neuropsychology 256
8.3.3 Functional neuroimaging 256
8.3.4 Brain stimulation 257
8.4 The mechanisms for semantic learning in the human anterior temporal lobe 259

9 The hippocampus, memory, and spatial function 260
9.1 Introduction and overview 260
9.1.1 Overview of what is computed by the hippocampal system 260
9.1.2 Overview of how the computations are performed by the hippocampal system 262
9.2 What is computed in the hippocampus 263
9.2.1 Systems-level anatomy 263
9.2.2 Evidence from the effects of damage to the hippocampus 265
9.2.3 Episodic memories need to be recalled from the hippocampus, and can be used to help build neocortical semantic memories 267
9.2.4 Systems-level neurophysiology of the primate hippocampus 270
9.2.5 Head direction cells in the presubiculum 285
9.2.6 Perirhinal cortex, recognition memory, and long-term familiarity memory 286
9.3 How computations are performed in the hippocampal system 293
9.3.1 Historical development of the theory of the hippocampus 293
9.3.2 Hippocampal circuitry 296
9.3.3 Medial entorhinal cortex, spatial processing streams, and grid cells 297
9.3.4 Lateral entorhinal cortex, object processing streams, and the generation of time cells in the hippocampus 301
9.3.5 CA3 as an autoassociation memory 307
9.3.6 Dentate granule cells 325
9.3.7 CA1 cells 328
9.3.8 Backprojections to the neocortex, memory recall, and consolidation 333
9.3.9 Backprojections to the neocortex – quantitative aspects 336
9.3.10 Simulations of hippocampal operation 339
9.3.11 The learning of spatial view and place cell representations 341
9.3.12 Linking the inferior temporal visual cortex to spatial view and place cells 342
9.3.13 A scientific theory of the art of memory: scientia artis memoriae 344
9.4 Tests of the theory of hippocampal cortex operation
   9.4.1 Dentate gyrus (DG) subregion of the hippocampus
   9.4.2 CA3 subregion of the hippocampus
   9.4.3 CA1 subregion of the hippocampus

9.5 Comparison with other theories of hippocampal function

10 The parietal cortex, spatial functions, and navigation
   10.1 Introduction and overview
   10.1.1 Introduction
   10.1.2 Overview of what is computed in the parietal cortex
   10.1.3 Overview of how the computations are performed in the parietal cortex
   10.2 Precuneus and medial area 7
   10.3 Navigation: What computations are performed in the parietal and related cortex
   10.4 How navigation is performed
   10.4.1 Navigation using a hippocampal allocentric Euclidean cognitive map
   10.4.2 Navigation using an entorhinal cortex goal vector system
   10.4.3 Transforms between allocentric and egocentric representations
   10.4.4 Navigational computations using neuron types found in primates

11 The orbitofrontal cortex, amygdala, reward value, and emotion
   11.1 Introduction and overview
   11.1.1 Introduction
   11.1.2 Overview of what is computed in the orbitofrontal cortex
   11.1.3 Overview of how the computations are performed by the orbitofrontal cortex
   11.2 The topology and connections of the orbitofrontal cortex
   11.2.1 Inputs to the orbitofrontal cortex
   11.2.2 Outputs of the orbitofrontal cortex
   11.3 What is computed in the orbitofrontal cortex
   11.3.1 The orbitofrontal cortex represents reward value
   11.3.2 Neuroeconomic value is represented in the orbitofrontal cortex
   11.3.3 A representation of face and voice expression and other socially relevant stimuli in the orbitofrontal cortex
   11.3.4 Negative reward prediction error neurons in the orbitofrontal cortex
   11.3.5 The human medial orbitofrontal cortex represents rewards, and the lateral orbitofrontal cortex non-reward and punishers
   11.3.6 Decision-making in the orbitofrontal / ventromedial prefrontal cortex
   11.3.7 The ventromedial prefrontal cortex and memory
   11.3.8 The orbitofrontal cortex and emotion
   11.3.9 Emotional orbitofrontal vs rational routes to action
   11.3.10 Comparison between the functions of the orbitofrontal cortex and amygdala in emotion
   11.4 How the computations are performed in the orbitofrontal cortex
   11.4.1 Decision-making in attractor networks in the brain
   11.4.2 Analyses of reward-related decision-making mechanisms in the orbitofrontal cortex
   11.4.3 A model for reversal learning in the orbitofrontal cortex
   11.4.4 A theory and model of non-reward neural mechanisms in the orbitofrontal cortex
   11.5 Highlights: the special computational roles of the orbitofrontal cortex

12 The cingulate cortex
   12.1 Introduction to and overview of the cingulate cortex
   12.1.1 Introduction
   12.1.2 Overview of what is computed in the cingulate cortex
   12.1.3 Overview of how the computations are performed by the cingulate cortex
   12.2 Anterior Cingulate Cortex
12.2.1 Anterior cingulate cortex anatomy and connections 450
12.2.2 Anterior cingulate cortex: A framework 451
12.2.3 Pregenual anterior cingulate representations of reward value, and supracallosal anterior cingulate representations of punishers and non-reward 453
12.2.4 Anterior cingulate cortex and action-outcome representations 455
12.2.5 Anterior cingulate cortex lesion effects 455
12.2.6 Subgenual cingulate cortex 456
12.3 Mid-cingulate cortex, the cingulate motor area, and action–outcome learning 457
12.4 The posterior cingulate cortex 458
12.5 How the computations are performed by the cingulate cortex 459
12.5.1 The anterior cingulate cortex and emotion 459
12.5.2 Action–outcome learning 459
12.5.3 Connectivity of the posterior cingulate cortex with the hippocampal memory system 460
12.6 Synthesis and conclusions 461

13 The motor cortical areas 464
13.1 Introduction and overview 464
13.2 What is computed in different cortical motor-related areas 464
13.2.1 Ventral parietal and ventral premotor cortex F4 464
13.2.2 Superior parietal areas with activity related to reaching 466
13.2.3 Inferior parietal areas with activity related to grasping, and ventral premotor cortex F5 466
13.3 The mirror neuron system 466
13.4 How the computations are performed in motor cortical and related areas 467

14 The basal ganglia 468
14.1 Introduction and overview 468
14.2 Systems-level architecture of the basal ganglia 469
14.3 What computations are performed by the basal ganglia? 471
14.3.1 Effects of striatal lesions 471
14.3.2 Neuronal activity in different parts of the striatum 473
14.4 How do the basal ganglia perform their computations? 485
14.4.1 Interaction between neurons and selection of output 485
14.4.2 Convergence within the basal ganglia, useful for stimulus-response habit learning 489
14.4.3 Dopamine as a reward prediction error signal for reinforcement learning in the striatum 491
14.5 Comparison of computations for selection in the basal ganglia and cerebral cortex 494

15 Cerebellar cortex 497
15.1 Introduction 497
15.2 Architecture of the cerebellum 498
15.2.1 The connections of the parallel fibres onto the Purkinje cells 498
15.2.2 The climbing fibre input to the Purkinje cell 499
15.2.3 The mossy fibre to granule cell connectivity 500
15.3 Modifiable synapses of parallel fibres onto Purkinje cell dendrites 502
15.4 The cerebellar cortex as a perceptron 502
15.5 Highlights: differences between cerebral and cerebellar cortex microcircuitry 503

16 The prefrontal cortex 505
16.1 Introduction and overview 505
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.2 Divisions of the lateral prefrontal cortex</td>
<td>508</td>
</tr>
<tr>
<td>16.2.1 The dorsolateral prefrontal cortex</td>
<td>509</td>
</tr>
<tr>
<td>16.2.2 The caudal prefrontal cortex</td>
<td>510</td>
</tr>
<tr>
<td>16.2.3 The ventrolateral prefrontal cortex</td>
<td>512</td>
</tr>
<tr>
<td>16.3 The lateral prefrontal cortex and top-down attention</td>
<td>512</td>
</tr>
<tr>
<td>16.4 How the computations are performed in the prefrontal cortex</td>
<td>515</td>
</tr>
<tr>
<td>16.4.1 Cortical short-term memory systems and attractor networks</td>
<td>515</td>
</tr>
<tr>
<td>16.4.2 Prefrontal cortex short-term memory networks, and their relation to perceptual networks</td>
<td>517</td>
</tr>
<tr>
<td>16.4.3 Mapping from one representation to another in short-term memory</td>
<td>522</td>
</tr>
<tr>
<td>16.4.4 The mechanisms of top-down attention</td>
<td>524</td>
</tr>
<tr>
<td>16.4.5 Computational necessity for a separate, prefrontal cortex, short-term memory system</td>
<td>525</td>
</tr>
<tr>
<td>16.4.6 Synaptic modification is needed to set up but not to reuse short-term memory systems</td>
<td>525</td>
</tr>
<tr>
<td>16.4.7 Sequence memory</td>
<td>525</td>
</tr>
<tr>
<td>16.4.8 Working memory, and planning</td>
<td>526</td>
</tr>
<tr>
<td>17 Language and syntax in the brain</td>
<td>527</td>
</tr>
<tr>
<td>17.1 Introduction and overview</td>
<td>527</td>
</tr>
<tr>
<td>17.1.1 Introduction</td>
<td>527</td>
</tr>
<tr>
<td>17.1.2 Overview</td>
<td>527</td>
</tr>
<tr>
<td>17.2 What is computed in different brain systems to implement language</td>
<td>529</td>
</tr>
<tr>
<td>17.2.1 The Wernicke-Lichtheim-Geschwind hypothesis</td>
<td>529</td>
</tr>
<tr>
<td>17.2.2 The dual-stream hypothesis of speech comprehension</td>
<td>529</td>
</tr>
<tr>
<td>17.2.3 Reading requires different brain systems to hearing speech</td>
<td>530</td>
</tr>
<tr>
<td>17.2.4 Semantic representations</td>
<td>531</td>
</tr>
<tr>
<td>17.2.5 Syntactic processing</td>
<td>533</td>
</tr>
<tr>
<td>17.2.6 The parietal cortex: supramarginal and angular gyri</td>
<td>533</td>
</tr>
<tr>
<td>17.3 Hypotheses about how semantic representations are computed</td>
<td>534</td>
</tr>
<tr>
<td>17.4 A neurodynamical hypothesis about how syntax is computed</td>
<td>535</td>
</tr>
<tr>
<td>17.4.1 Binding by synchrony?</td>
<td>535</td>
</tr>
<tr>
<td>17.4.2 Syntax using a place code</td>
<td>536</td>
</tr>
<tr>
<td>17.4.3 Temporal trajectories through a state space of attractors</td>
<td>536</td>
</tr>
<tr>
<td>17.4.4 Hypotheses about the implementation of language in the cerebral cortex</td>
<td>537</td>
</tr>
<tr>
<td>17.4.5 Tests of the hypotheses – a model</td>
<td>540</td>
</tr>
<tr>
<td>17.4.6 Tests of the hypotheses – findings with the model</td>
<td>545</td>
</tr>
<tr>
<td>17.4.7 Evaluation of the hypotheses</td>
<td>548</td>
</tr>
<tr>
<td>17.4.8 Further approaches</td>
<td>552</td>
</tr>
<tr>
<td>18 Cortical attractor dynamics and connectivity, stochasticity, psychiatric disorders, and aging</td>
<td>554</td>
</tr>
<tr>
<td>18.1 Introduction and overview</td>
<td>554</td>
</tr>
<tr>
<td>18.1.1 Introduction</td>
<td>554</td>
</tr>
<tr>
<td>18.1.2 Overview</td>
<td>554</td>
</tr>
<tr>
<td>18.2 The noisy cortex</td>
<td>555</td>
</tr>
<tr>
<td>18.2.1 Reasons why the brain is inherently noisy and stochastic</td>
<td>556</td>
</tr>
<tr>
<td>18.2.2 Attractor networks, energy landscapes, and stochastic neurodynamics</td>
<td>559</td>
</tr>
<tr>
<td>18.2.3 A multistable system with noise</td>
<td>564</td>
</tr>
<tr>
<td>18.2.4 Stochastic dynamics and the stability of short-term memory</td>
<td>566</td>
</tr>
<tr>
<td>18.2.5 Stochastic dynamics in decision-making, and the evolutionary utility of probabilistic choice</td>
<td>570</td>
</tr>
<tr>
<td>18.2.6 Selection between conscious vs unconscious decision-making, and free will</td>
<td>571</td>
</tr>
<tr>
<td>18.2.7 Stochastic dynamics and creative thought</td>
<td>572</td>
</tr>
<tr>
<td>18.2.8 Stochastic dynamics and unpredictable behaviour</td>
<td>573</td>
</tr>
<tr>
<td>18.3 Attractor dynamics and schizophrenia</td>
<td>573</td>
</tr>
<tr>
<td>18.3.1 Introduction</td>
<td>573</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>18.3.2</td>
<td>A dynamical systems hypothesis of the symptoms of schizophrenia</td>
</tr>
<tr>
<td>18.3.3</td>
<td>Reduced functional connectivity of some brain regions in schizophrenia</td>
</tr>
<tr>
<td>18.3.4</td>
<td>Beyond the disconnectivity hypothesis of schizophrenia: reduced forward</td>
</tr>
<tr>
<td></td>
<td>but not backward connectivity</td>
</tr>
<tr>
<td>18.4</td>
<td>Attractor dynamics and obsessive-compulsive disorder</td>
</tr>
<tr>
<td>18.4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>18.4.2</td>
<td>A hypothesis about obsessive-compulsive disorder</td>
</tr>
<tr>
<td>18.4.3</td>
<td>Glutamate and increased depth of the basins of attraction</td>
</tr>
<tr>
<td>18.5</td>
<td>Depression and attractor dynamics</td>
</tr>
<tr>
<td>18.5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>18.5.2</td>
<td>A non-reward attractor theory of depression</td>
</tr>
<tr>
<td>18.5.3</td>
<td>The orbitofrontal cortex, and the theory of depression</td>
</tr>
<tr>
<td>18.5.4</td>
<td>Altered connectivity of the orbitofrontal cortex in depression</td>
</tr>
<tr>
<td>18.5.5</td>
<td>Activations of the orbitofrontal cortex related to depression</td>
</tr>
<tr>
<td>18.5.6</td>
<td>Implications, and possible treatments, and subtypes of depression</td>
</tr>
<tr>
<td>18.5.7</td>
<td>Mania and bipolar disorder</td>
</tr>
<tr>
<td>18.6</td>
<td>Attractor stochastic dynamics, aging, and memory</td>
</tr>
<tr>
<td>18.6.1</td>
<td>NMDA receptor hypofunction</td>
</tr>
<tr>
<td>18.6.2</td>
<td>Dopamine and norepinephrine</td>
</tr>
<tr>
<td>18.6.3</td>
<td>Impaired synaptic modification</td>
</tr>
<tr>
<td>18.6.4</td>
<td>Cholinergic function and memory</td>
</tr>
<tr>
<td>18.7</td>
<td>High blood pressure, reduced hippocampal functional connectivity, and impaired memory</td>
</tr>
<tr>
<td>18.8</td>
<td>Brain development, and structural differences in the brain</td>
</tr>
<tr>
<td>19</td>
<td>Computations by different types of brain, and by artificial neural systems</td>
</tr>
<tr>
<td>19.1</td>
<td>Introduction and overview</td>
</tr>
<tr>
<td>19.2</td>
<td>Computations that combine different computational systems in the brain to produce behaviour</td>
</tr>
<tr>
<td>19.3</td>
<td>Brain computation compared to computation on a digital computer</td>
</tr>
<tr>
<td>19.4</td>
<td>Brain computation compared with artificial deep learning networks</td>
</tr>
<tr>
<td>19.5</td>
<td>Reinforcement Learning</td>
</tr>
<tr>
<td>19.6</td>
<td>Levels of explanation, and the mind-brain problem</td>
</tr>
<tr>
<td>19.7</td>
<td>Levels of explanation, and levels of investigation</td>
</tr>
<tr>
<td>19.8</td>
<td>Brain-Inspired Intelligence</td>
</tr>
<tr>
<td>19.9</td>
<td>Brain-Inspired Medicine</td>
</tr>
<tr>
<td>19.9.1</td>
<td>Computational psychiatry and neurology</td>
</tr>
<tr>
<td>19.9.2</td>
<td>Reward systems in the brain, and their application to understanding food intake control and obesity</td>
</tr>
<tr>
<td>19.9.3</td>
<td>Multiple Routes to Action</td>
</tr>
<tr>
<td>19.10</td>
<td>Primates including humans have different brain organisation than rodents</td>
</tr>
<tr>
<td>19.10.1</td>
<td>The visual system</td>
</tr>
<tr>
<td>19.10.2</td>
<td>The taste system</td>
</tr>
<tr>
<td>19.10.3</td>
<td>The olfactory system</td>
</tr>
<tr>
<td>19.10.4</td>
<td>The somatosensory system</td>
</tr>
<tr>
<td>19.10.5</td>
<td>The auditory system</td>
</tr>
<tr>
<td>19.10.6</td>
<td>The hippocampal system and memory</td>
</tr>
<tr>
<td>19.10.7</td>
<td>The orbitofrontal cortex and amygdala</td>
</tr>
<tr>
<td>19.10.8</td>
<td>The cingulate cortex</td>
</tr>
<tr>
<td>19.10.9</td>
<td>The motor system</td>
</tr>
<tr>
<td>19.10.10</td>
<td>Language</td>
</tr>
<tr>
<td>A</td>
<td>Introduction to linear algebra for neural networks</td>
</tr>
</tbody>
</table>
A.1 Vectors
   A.1.1 The inner or dot product of two vectors 634
   A.1.2 The length of a vector 636
   A.1.3 Normalizing the length of a vector 636
   A.1.4 The angle between two vectors: the normalized dot product 636
   A.1.5 The outer product of two vectors 637
   A.1.6 Linear and non-linear systems 638
   A.1.7 Linear combinations, linear independence, and linear separability 639

A.2 Application to understanding simple neural networks 640
   A.2.1 Capability and limitations of single-layer networks 641
   A.2.2 Non-linear networks: neurons with non-linear activation functions 643
   A.2.3 Non-linear networks: neurons with non-linear activations 644

B Neuronal network models 646
   B.1 Introduction 646
   B.2 Pattern association memory
      B.2.1 Architecture and operation 647
      B.2.2 A simple model 649
      B.2.3 The vector interpretation 652
      B.2.4 Properties 653
      B.2.5 Prototype extraction, extraction of central tendency, and noise reduction 656
      B.2.6 Speed 656
      B.2.7 Local learning rule 657
      B.2.8 Implications of different types of coding for storage in pattern associators 662
   B.3 Autoassociation or attractor memory
      B.3.1 Architecture and operation 663
      B.3.2 Introduction to the analysis of the operation of autoassociation networks 665
      B.3.3 Properties 667
      B.3.4 Diluted connectivity and the storage capacity of attractor networks 674
      B.3.5 Use of autoassociation networks in the brain 685
   B.4 Competitive networks, including self-organizing maps
      B.4.1 Function 686
      B.4.2 Architecture and algorithm 687
      B.4.3 Properties 688
      B.4.4 Utility of competitive networks in information processing by the brain 693
      B.4.5 Guidance of competitive learning 694
      B.4.6 Topographic map formation 696
      B.4.7 Invariance learning by competitive networks 700
      B.4.8 Radial Basis Function networks 702
      B.4.9 Further details of the algorithms used in competitive networks 703
   B.5 Continuous attractor networks
      B.5.1 Introduction 707
      B.5.2 The generic model of a continuous attractor network 709
      B.5.3 Learning the synaptic strengths in a continuous attractor network 709
      B.5.4 The capacity of a continuous attractor network: multiple charts 711
      B.5.5 Continuous attractor models: path integration 712
      B.5.6 Stabilization of the activity packet within a continuous attractor network 715
      B.5.7 Continuous attractor networks in two or more dimensions 717
      B.5.8 Mixed continuous and discrete attractor networks 717
   B.6 Network dynamics: the integrate-and-fire approach
      B.6.1 From discrete to continuous time 718
      B.6.2 Continuous dynamics with discontinuities 720
      B.6.3 An integrate-and-fire implementation 723
      B.6.4 The speed of processing of attractor networks 725
      B.6.5 The speed of processing of a four-layer hierarchical network 727
      B.6.6 Spike response model 731
   B.7 Network dynamics: introduction to the mean-field approach 732
   B.8 Mean-field based neurodynamics 733
Contents

B.8.1 Population activity 733
B.8.2 The mean-field approach used in a model of decision-making 735
B.8.3 The model parameters used in the mean-field analyses of decision-making 737
B.8.4 A basic computational module based on biased competition 737
B.8.5 Multimodular neurodynamical architectures 739
B.9 Interacting attractor networks 742
B.10 Sequence memory implemented by adaptation in an attractor network 745
B.11 Error correction networks 745
  B.11.1 Architecture and general description 746
  B.11.2 Generic algorithm for a one-layer error correction network 746
  B.11.3 Capability and limitations of single-layer error-correcting networks 747
  B.11.4 Properties 751
B.12 Error backpropagation multilayer networks 753
  B.12.1 Introduction 753
  B.12.2 Architecture and algorithm 753
  B.12.3 Properties of multilayer networks trained by error backpropagation 756
B.13 Convolution networks 757
B.14 Contrastive Hebbian learning: the Boltzmann machine 758
B.15 Deep Belief Networks 760
B.16 Reinforcement learning
  B.16.1 Associative reward–penalty algorithm of Barto and Sutton 761
  B.16.2 Reward prediction error or delta rule learning, and classical conditioning 763
  B.16.3 Temporal Difference (TD) learning 764
B.17 Learning in the neocortex 767
B.18 Forgetting in cortical associative neural networks, and memory reconsolidation 769
B.19 Highlights 773

C Information theory, and neuronal encoding 774
C.1 Information theory 775
  C.1.1 The information conveyed by definite statements 775
  C.1.2 Information conveyed by probabilistic statements 776
  C.1.3 Information sources, information channels, and information measures 777
  C.1.4 The information carried by a neuronal response and its averages 778
  C.1.5 The information conveyed by continuous variables 781
C.2 The information carried by neuronal responses 783
  C.2.1 The limited sampling problem 783
  C.2.2 Correction procedures for limited sampling 784
  C.2.3 The information from multiple cells: decoding procedures 785
  C.2.4 Information in the correlations between cells: a decoding approach 789
  C.2.5 Information in the correlations between cells: second derivative approach 794
C.3 Information theory results 797
  C.3.1 The sparseness of the distributed encoding used by the brain 798
  C.3.2 The information from single neurons 809
  C.3.3 The information from single neurons: temporal codes versus rate codes 812
  C.3.4 The information from single neurons: the speed of information transfer 814
  C.3.5 The information from multiple cells: independence versus redundancy 825
  C.3.6 Should one neuron be as discriminative as the whole organism? 829
  C.3.7 The information from multiple cells: the effects of cross-correlations 830
  C.3.8 Conclusions on cortical neuronal encoding 834
C.4 Information theory terms – a short glossary 838
C.5 Highlights 839

D Simulation software for neuronal networks, and information analysis of neuronal encoding 840
D.1 Introduction 840
D.2 Autoassociation or attractor networks 841
  D.2.1 Running the simulation 841
  D.2.2 Exercises 843
D.3 Pattern association networks 843
  D.3.1 Running the simulation 843
  D.3.2 Exercises 845
D.4 Competitive networks and Self-Organizing Maps 846
  D.4.1 Running the simulation 846
  D.4.2 Exercises 847
D.5 Further developments 848
D.6 Matlab code for a tutorial version of VisNet 848
D.7 Matlab code for information analysis of neuronal encoding 849
D.8 Matlab code to illustrate the use of spatial view cells in navigation 849
D.9 Highlights 849

References 850

Index 924
19 Computations by different types of brain, and by artificial neural systems

19.1 Introduction and overview

In this Chapter a comparison is made between computations in the brain and computations performed in computers (Section 19.3). This is intended to be helpful to those engineers, computer scientists, AI specialists and others interested in designing new computers that emulate aspects of brain function. In fact, the whole of this book is intended to be useful for this aim, by setting out what is computed by different brain systems, and what we know about how it is computed. It is essential to know this if an emulation of brain function is to be performed, and this is important to enable this group of scientists to bring their expertise to help understand brain function more.

The comparison between computation in the brain and computation in computers in Section 19.3 is also intended to be of interest to neuroscientists to emphasize the differences in the implementations of computation in the brain and in computers.

Part of the interest here is that many computer implementations of ‘neural networks’ implement very different algorithms to those implemented in the brain. This means that there is a lot left not only to learn about the brain, but also to learn from the brain. For example, Section 19.4 specifically compares deep learning in artificial neural networks to the type of self-organising learning implemented in the brain; and Section 19.5 compares reinforcement learning in the brain (and in computers) to other learning systems in the brain.

Sections 19.6 and 19.7 raise an important issue for understanding brains (and computers): the relation between explanations at different levels (neurons, networks, the computation that is performed, etc) of how brains and computers work, which is relevant to the mind-brain problem (Rolls, 2020a). Section 19.7 considers in particular the levels of investigation that are essential in order to understand how the brain performs its computations.

Section 19.8 is on Brain-Inspired Intelligence: how we can learn from the brain to implement new applications in computers.

Section 19.9 is on Brain-Inspired Medicine: how our developing understanding of how the brain computes has many implications not only for understanding how we as humans operate in health, but also what happens in for example mental disorders (considered in Chapter 18) and in other types of behaviour, including behaviour driven by our reward systems that can result in overeating and obesity.

Finally, Section 19.10 makes it clear why the focus of this book is on computations in primate (and that very much includes human) brains, rather than on rodent (rat and mice) brains. It is because the systems-level organization of primate including human brains is quite different from that in rodents.

We start in Section 19.2 by making the point that to understand how the brain works, a way forward is to understand how its different systems operate; and how that provides the foundation for how different systems operate together to produce a particular behaviour. That is also an important way forward when aiming to emulate in a computer how the brain performs a particular function.
19.2 Computations that combine different computational systems in the brain to produce behaviour

By understanding what is computed in each brain system, how the information is represented, and how each system connects to other systems, we have a foundation for understanding how different brain systems operate together to perform behavioural functions. The ability to understand computationally what each part of the brain computes, and how the results of its computations are represented, is crucial to this aim. Studying each of these systems when the same behaviour is being performed is also key to this aim. This approach is feasible, to understand the computations each system separately, but while the same behaviour is being performed, because of the evidence that there is considerable localisation of function in the brain, as set out by Rolls (2016b).

To model how the brain implements a function such as visual object recognition, or episodic memory, the strategy can next be to combine together the computations performed by several brain systems using information of the type presented in the different Chapters of this book, and in the Appendices. When moving forward to do this, it is suggested that a twofold strategy will be useful. The first part is to utilize the information provided in this book about what is computed in different brain systems, and how it is computed. The second part towards the implementation is to take into account the principles of operation of the cerebral cortex, as set out for example in Cerebral Cortex: Principles of Operation (Rolls, 2016b). In that book I did not attempt to produce a theory of how different brain systems compute, but instead focussed on the principles of operation and computation of the cortex. Thus what is in these two books complements each other in understanding the computations performed by the brain, and together they provide a foundation for further developments in our understanding, and in applying that understanding. Important applications of these advances are to understanding brain operation in disease, and developing new machines with what might be described as ‘Brain-Inspired Intelligence’.

19.3 Brain computation compared to computation on a digital computer

An important component of the approach taken in this book is to describe what computations are performed, and then how they may be performed. This enables the different levels of analysis in neuroscience to be brought together in a multilevel causal account.

However, the types of computation that are performed by the brain, and the computational style, are very different from the type of computation performed by a digital computer, which performs specified logical / syntactic operations on exact data retrieved from memory, and then stores the exact result back in memory. To highlight some of the principles of brain computation described in this book, and to emphasize how the type of computation is very different from that performed by digital computers, the principles of computation by the brain are compared next with those of a digital computer, with a summary of some of the differences in Table 19.1.

Data addressing. An item of data is retrieved from the memory of a digital computer by providing the address of the data in memory, and then the data can be manipulated (moved, compared, added to the data at another address in the computer etc.) using typically a 32-bit or 64-bit binary word of data. Pointers to memory locations are thus used extensively. In contrast, in the cortex, the data are used as the access key (in for example a pattern associator, autoassociator, and competitive network), and the neurons with synaptic weights that match
Table 19.1 Brain computation vs Digital computer computation

<table>
<thead>
<tr>
<th>Brain Computation</th>
<th>Digital computer computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dot product similarity followed by a threshold vs logical functions e.g. AND, NAND, OR</td>
<td></td>
</tr>
<tr>
<td>2. 10,000 inputs per neuron vs several inputs to a logic gate</td>
<td></td>
</tr>
<tr>
<td>3. Content-addressable memory vs data accessed by address</td>
<td></td>
</tr>
<tr>
<td>4. Fault tolerant: dot product vs no inherent fault tolerance</td>
<td></td>
</tr>
<tr>
<td>5. Generalization and completion vs only exact match in the hardware</td>
<td></td>
</tr>
<tr>
<td>6. Fast: parallel update of a neuron and of neurons in a network vs serial processing with fast components</td>
<td></td>
</tr>
<tr>
<td>7. Approximate, low precision vs high precision 64 bit</td>
<td></td>
</tr>
<tr>
<td>8. Stochastic spiking noise: probabilistic computation helps originality, creativity vs noise free implemented by non-linearities</td>
<td></td>
</tr>
<tr>
<td>9. Dynamics are parallel e.g. an attractor network retrieves in 1–2 time constants of the synapses vs inherently serial</td>
<td></td>
</tr>
<tr>
<td>10. Heuristics, e.g. Invariant object recognition analyses only part of a scene vs attempt to understand a whole scene</td>
<td></td>
</tr>
<tr>
<td>11. Syntax: none inherent vs syntactical operations on data at an address to implement the computation.</td>
<td></td>
</tr>
<tr>
<td>12. The architecture adapts by learning vs fixed architecture with different software</td>
<td></td>
</tr>
<tr>
<td>13. Sparse distributed representation vs binary encoding in a computer word</td>
<td></td>
</tr>
<tr>
<td>14. Mind vs brain vs software vs hardware</td>
<td></td>
</tr>
</tbody>
</table>

the data respond. Memory in the brain is thus content-addressable. In one time constant of the synapses / cell membranes, the brain has thus found the correct output. In contrast, in a digital computer a serial search is required, in which the data at every address must be retrieved and compared in turn to the test data to discover if there is a match.

Vector similarity vs logical operations. Cortical computation including that performed by associative memories and competitive networks operates by vector similarity – the dot product of the input and of the synaptic weight vector of each neuron are produced, and the neurons with the highest dot product will be most activated (Section 1.4 and Appendix B). Even if an exact match is not found, some output is likely to result. In contrast, in a digital computer, logical operations (such as AND, OR, XOR) and exact mathematical operations (such as addition, subtraction, multiplication, and division) are computed. (There is no bitwise similarity between the binary representations of 7 (0111) and 8 (1000).) The similarity computations performed by the brain may be very useful, in enabling similarities to be seen and parallels to be drawn, and this may be an interesting aspect of human creativity, realized for example in *Finnegans’s Wake* by James Joyce in which thoughts reached by associative thinking abound. However, the lateral thinking must be controlled, to prevent bizarre similarities being found, and this is argued to be related to the symptoms of schizophrenia in Section 18.3.

Fault tolerance. Because exact computations are performed in a digital computer, there is no in-built fault tolerance or graceful degradation. If one bit of a memory has a fault, the whole memory chip must be discarded. In contrast, the brain is naturally fault tolerant, because it uses vector similarity (between its input firing rate vector and synaptic weight vectors) in its calculations, and linked to this, distributed representations. This makes the brain robust developmentally with respect to ‘missing synapses’, and robust with respect to later losing some synapses or neurons (see Appendix B).

Word length. To enable the vector similarity comparison to have high capacity (for example memory capacity) the ‘word length’ in the brain is typically long, with between 10,000 and 50,000 synapses onto every neuron being common in cortical areas. (Remember that the
leading term in the factor that determines the storage capacity of an associative memory is the number of synapses per neuron – see Sections B.2 and B.3.) In contrast, the word length in typical digital computers at 32 or 64 bits is much shorter, though with the binary and exact encoding used this allows great precision in a digital computer.

**Readability of the code.** To comment further on the encoding: in the cortex, the code must not be too compact, so that it can be read by neuronally plausible dot product decoding and can use vector similarity to generalize to similar patterns (Rolls, 2016b), as shown in Sections 1.7, B.2.4.3 and C.3.1, and throughout Appendices B and C. In contrast, the binary encoding used in a digital computer is optimally efficient, with one bit stored and retrievable for each binary memory location. However, the computer binary code cannot be read by neuronally plausible dot-product decoding.

**Precision.** The precision of the components in a digital computer is that every modifiable memory location must store one bit accurately. In contrast, it is of interest that synapses in the brain need not work with exact precision, with for example typically less that one bit per synapse being usable in associative memories (Treves and Rolls, 1991; Rolls and Treves, 1998). The precision of the encoding of information in the firing rate of a neuron is likely to be a few bits – perhaps 3 – as judged by the standard deviation and firing rate range of individual cortical neurons (Appendix C).

**The speed of computation.** This brings us to the speed of computation. In the brain, considerable information can be read in 20 ms from the firing rate of an individual neuron (e.g. 0.2 bits), leading to estimates of 10–30 bits/s for primate temporal cortex visual neurons (Rolls, Treves and Tovee, 1997b; Rolls and Tovee, 1994), and 2–3 bits/s for rat hippocampal cells (Skaggs, McNaughton, Gothard and Markus, 1993; Rolls, 2016b; Rolls and Treves, 2011) (Appendix C). Though this is very slow compared to a digital computer, the brain does have the advantage that a single neuron receives spikes from thousands of individual neurons, and computes its output from all of these inputs within a period of approximately 10–20 ms (determined largely by the time constant of the synapses) (Rolls, 2016b). Moreover, each neuron, up to at least the order of tens of neurons, conveys independent information, as described in Appendix C.

**Parallel vs serial processing.** Computation in a conventional digital computer is inherently serial, with a single central processing unit that must fetch the data from a memory address, manipulate the word of data, and store it again at a memory address. In contrast, brain computation is parallel in at least three senses.

First, an individual neuron in performing a dot product between its input firing rate vector and its synaptic weight vector does operate in an analog way to sum all the injected currents through the thousands of synapses to calculate the activation \( h_i \), and fire if a threshold is reached, in a time in the order of the synaptic time constant. To implement this on a digital computer would take \( 2C \) operations (\( C \) multiply operations, and \( C \) add operations, where \( C \) is the number of synapses per neuron – see Equation 1.1).

Second, each neuron in a single network (e.g. a small region of the cortex with of the order of hundreds of thousands of neurons) does this dot product computation in parallel, followed by interaction through the GABA inhibitory neurons, which again is fast. (It is in the order of the time constant of the synapses involved, operates in continuous time, and does not have to wait at all until the dot product operation of the pyramidal cells has been completed by all neurons given the spontaneous neuronal activity that allows some neurons to reflect their changed inputs rapidly by when the next spike occurs.) This interaction sets the threshold in
Brain computation compared to computation on a digital computer

Third, different brain areas operate in parallel. An example is that the ventral visual stream computes object representations, while simultaneously the dorsal visual stream computes (inter alia) the types of global motion described by Rolls and Stringer (2007) (Section 3.3), including for example a wheel rotating in the same direction as it traverses the visual field. Another example is that within a hierarchical system in the brain, every stage operates simultaneously, as a pipeline processor, with a good example being V1–V2–V4–IT, which can all operate simultaneously as the data are pipelined through (Chapter 2).

We could refer to the computation that takes place in different modules, that is in networks that are relatively separate in terms of the number of connections between modules relative to those within modules, such as those in the dorsal and ventral visual streams, as being parallel computation. Within a single module or network, such as the CA3 region of the hippocampus, or inferior temporal visual cortex, we could refer to the computation as being parallel distributed computation, in that the closely connected neurons in the network all contribute to the result of the computation. For example, with distributed representations in an attractor network, all the neurons interact with each other directly and through the inhibitory interneurons to retrieve and then maintain a stable pattern in short-term memory (Section B.3). In a competitive network involved in pattern categorization, all the neurons interact through the inhibitory interneurons to result in an active population of neurons that represents the best match between the input stimulus and what has been learned previously by the network, with neurons with a poor match being inhibited by neurons with a good match (Section B.4). In a more complicated scenario with closely connected interacting modules, such as the prefrontal cortex and the inferior temporal cortex during top-down attention tasks and more generally forward and backward connections between adjacent cortical areas, we might also use the term parallel distributed computation, as the bottom-up and top-down interactions may be important in how the whole dynamical system of interconnected networks settles (see examples in Sections 16.4.1, 16.3, 2.11, and Appendix B).

Stochastic dynamics and probabilistic computation. Digital computers do not have noise to contend with as part of the computation, as they use binary logic levels, and perform exact computation. In contrast, brain computation is inherently noisy, and this gives it a non-exact, probabilistic, character. One of the sources of noise in the brain is the spiking activity of each neuron. Each neuron must transmit information by spikes, for an all-or-none spike carried along an axon ensures that the signal arrives faithfully, and is not subject to the uncertain cable transmission line losses of analog potentials. But once a neuron needs to spike, then it turns out to be important to have spontaneous activity, so that neurons do not all have to charge up from a hyperpolarized baseline whenever a new input is received. The fact that neurons are kept near threshold, with therefore some spontaneous spiking, is inherent to the rapid operation of for example autoassociative retrieval, as described in Section B.3. But keeping the neurons close to threshold, and the spiking activity received from other neurons, results in spontaneous spike trains that are approximately Poisson, that is with spikes randomly timed for a given mean firing rate. The result of the interaction of all these randomly timed inputs is that in a network of finite size (i.e. with a limited number of neurons) there will be statistical fluctuations, that influence which memory is recalled, which decision is taken, etc. as described in Section 11.4.1 and Chapter 18. Thus brain computation is inherently noisy and probabilistic, and this has many advantages, as described in Section 11.4.1, Chapter 18, and by Rolls and Deco (2010) and Rolls (2016b).

Syntax. Digital computers can perform arbitrary syntactical operations on operands, because
they use pointers to address each of the different operands required (corresponding even for example to the subject, the verb, and the object of a sentence). In contrast, as data are not accessed in the brain by pointers that can point anywhere, but instead just having neurons firing to represent a data item, a real problem arises in specifying which neurons firing represent for example the subject, the verb, and the object, and distributed representations potentially make this even more difficult. The brain thus inherently finds syntactical operations difficult (as explained in Section 17.4). We do not know how the brain implements the syntax required for language. But we do know that the firing of neurons conveys ‘meaning’ based on spatial location in the brain. For example, a neuron firing in V1 indicates that a bar or edge matching the filter characteristic of the neuron is present at a particular location in space. Another neuron in V1 encodes another feature at another position in space. A neuron in the inferior temporal visual cortex indicates (with other neurons helping to form a distributed representation) that a particular object or face is present in the visual scene. Perhaps the implementation of the syntax required for language that is implemented in the brain also utilizes the spatial location of the network in the cortex to help specify what syntactical role the representation should perform. This is a suggestion I make, as it is one way that the brain could deal with the implementation of the syntax required for language (Section 17.4).

Modifiable connectivity. The physical architecture (what is connected to what) of a digital computer is fixed. In contrast, the connectivity of the brain alters as a result of experience and learning, and indeed it is alterations in the strength of the synapses (which implement the connectivity) that underlie learning and memory. Indeed, self-organization in for example competitive networks has a strong influence on how the brain is matched to the statistics of the incoming signals from the world, and of the architecture that develops. In a digital computer, every connection must be specified. In contrast, in the brain there are far too few genes (of order 25,000) for the synaptic connections in the brain (of order $10^{15}$, given approximately $10^{11}$ neurons each with in the order of $10^{4}$ synapses) for the genes to specify every connection$^{20}$. The genes must therefore specify some much more general rules, such as that each CA3 neuron should make approximately 12,000 synapses with other CA3 neurons, and receive approximately 48 synapses from dentate granule cells (see Chapter 9). The actual connections made would then be made randomly within these constraints, and then strengthened or lost as a result of self-organization based on for example conjunctive pre- and postsynaptic activity. Some of the rules that may be specified genetically have been suggested on the basis of a comparison of the architecture of different brain areas (Rolls and Stringer, 2000; Rolls, 2016b). Moreover, it has been shown that if these rules are selected by a genetic algorithm based on the fitness of the network that self-organizes and learns based on these rules, then architectures are built that solve different computational problems in one-layer networks, including pattern association learning, autoassociation memory, and competitive learning (Rolls and Stringer, 2000; Rolls, 2016b). The architecture of the brain is thus interestingly adaptive, but guided in the long term by genetic selection of the building rules.

Logic. The learning rules that are implemented in the brain that are most widely accepted are associative, as exemplified by LTP and LTD. This, and the vector similarity operations implemented by neurons, set the stage for processes such as pattern association, autoassociation, and competitive learning to occur naturally, but not for logical operations such as XOR and NAND or arithmetic operations. Of course, the non-linearity inherent in the firing threshold

---

$^{20}$For comparison, a computer with 1 Gb of memory has approximately $10^{10}$ modifiable locations, and if it had a 100 Gb disk that would have approximately $10^{12}$ modifiable locations.
Brain computation compared to computation on a digital computer

of neurons is important in many of the properties of associative memories and competitive learning, as described in Appendix B, and indeed are how some of the non-linearities that can be seen with attention can arise (Deco and Rolls, 2005b).

**Dynamical interaction between modules.** Because the brain has populations of neurons that are simultaneously active (operating in parallel), but are interconnected, many properties arise naturally in dynamical neural systems, including the interactions that give rise to top-down attention (Sections 16.3 and 2.11), the effects of mood on memory (Rolls and Stringer, 2001b; Rolls, 2014a) etc. Because simultaneous activity of different computational nodes does not occur in digital computers, these dynamical systems properties that arise from interacting subsystems do not occur naturally, though they can be simulated.

The cortex has recurrent excitatory connections within a cortical area, and reciprocal, forward and feedback, connections between adjacent cortical areas in the hierarchy. The excitatory connections enable cortical activity to be maintained over short periods, making short-term memory an inherent property of the cortex. They also provide the autoassociative long-term memory with completion from a partial cue (given associative synaptic modifiability in these connections). However, completion on a digital computer is a difficult and serial process to identify a possible correct partial match. Another comparison is that the short-term memory property of the cortex is part of what makes the cortex a dynamical interacting system, with for example what is in short-term memory in the prefrontal cortex acting to influence memory recall, perception, and even what decision is taken, in other networks, by top-down biased competition (see Sections 16.3 and 2.11). There is a price that the brain pays for this positive feedback inherent in its recurrent cortical circuitry, which is that this circuitry is inherently unstable, and requires strong control by inhibitory interneurons to minimize the risk of epilepsy and other disorders (Chapter 18).

**Modular organization.** Brain organization is modular, with many relatively independent modules each performing a different function, whereas digital computers typically have a single central processing unit connected to memory. The cortex has many localized modules with dense connectivity within a module, and then connections to a few other modules. The reasons for the modularity of the brain are considered by Rolls (2016b).

**Hierarchical organization.** As described by Rolls (2016b), many cortical systems are organized hierarchically. A major reason for this is that this enables the connectivity to be kept within the limits of which neurons appear capable (up to 50,000 synapses per neuron), yet for global computation (such as the presence of a particular object anywhere in the visual field) to be achieved, as exemplified by VisNet, a model of invariant visual object recognition (see Fig. 2.2 and Chapter 2). Another important reason is that hierarchical organization simplifies the learning that is required at each stage and enables it to be a local operation, in contrast to backpropagation of error networks where similar problems could in principle be solved in a two-layer network (with one hidden layer), but would require training with a non-local learning rule (Appendix C) as well as potentially neurons with very large numbers of connections.

Another feature of cortical organization is that the number of areas or levels in any hierarchy in the brain is not more than 4 or 5, because each area requires 20–30 ms of computation and transmission time. (This of course contrasts with the hundreds of successive layers being explored with artificial neural networks, as described in Section B.13, which is completely biologically implausible in terms of response times.)
19.4 A comparison of brain computation with learning in artificial deep learning networks using error backpropagation

Learning in deep networks with error backpropagation (LeCun, Bengio and Hinton, 2015) is described in Section B.12. An attempt to compare the operation of hierarchical convolutional deep neural networks to the operation of the ventral visual cortical stream (Yamins and DiCarlo, 2016; Rajalingham, Issa, Bashivan, Kar, Schmidt and DiCarlo, 2018) is described in Section 2.8.3. Here the biological plausibility and advantages and disadvantages of this deep learning compared with brain computation are considered.

A first problem in terms of biological plausibility of deep networks trained by error backpropagation is that this is inherently a supervised system, with a separate teacher for each output neuron informing it how each output neuron should be responding for a given input pattern at the bottom layer of the network. There is no architecture that looks remotely like this in the cerebral cortex, with a separate teacher for each neuron. The only part of the brain where there is what might be a teacher for each neuron is in the cerebellum, where there is a single climbing fibre for each Purkinje cell, and a network with no hidden layers (Chapter 15).

But the problem is deeper than that there is no teacher for each output neuron (and the other problems that appear to make deep learning biologically implausible described below). A bigger problem for backpropagation of error networks is that someone has to know what the output should be, for every neuron, that is, backprop is a supervised learning system. The great beauty of the brain in contrast is that it is a largely self-organizing system, without a teacher telling each neuron what to do, as is made evident throughout this book, with a simple example of such a system a hierarchically organised set of competitive networks with convergence from stage to stage, as described in Chapter 2. Such self-organizing networks learn what categories to form based on the statistics of the inputs arriving from the world. They often use special adaptations to learn those statistics in a useful way, such as the trace short-term memory learning rule described in Chapter 2 that enables invariant representations to be learned because over short time periods it is likely to be the same object that is being viewed.

In essence, a deep network trained by backpropagation of error tries to achieve an optimal mapping between input patterns and output patterns of neural activity. The number of neurons in each deep layer must be carefully adjusted to ensure that the system does not become just a look-up table, yet has sufficient neurons to enable the mapping and generalization required, and the number of layers is often adjusted to be 100 or more to help to minimize overall errors. But a problem that arises is that what is being represented at each intermediate layer is very difficult to understand, and is likely to yield very little useful information to help understand the brain. There may be a lesson here:

At each stage of cortical processing, some output is always taken and directed to structures such as the striatum and basal ganglia for use in behavioral output. That is a key aspect of cortical design (Rolls, 2016b). It may reflect an interesting way in which the cortex evolves, by adding new areas each of which can produce useful output. That is not true of deep backprop networks.

An important point is that the brain does not solve a predetermined mapping problem from input to output, as in deep learning. Instead, the general architecture of the brain is specified by a relatively small number of genes that appear to specify what type of neuron connects to what other type to generate the architecture within and between cortical areas (Chapter 19 of Rolls (2016b)). The brain then builds the details of its synaptic connectivities
using self-organizing learning based on the environment in which it finds itself. This view has been developed by comparing the connectivity of different cortical areas, and then allowing genetic algorithms to select different combinations of the connectivity rules to build neural networks in computers that can successfully solve different types of computational problem for which different architectures are needed (Rolls and Stringer, 2000; Rolls, 2016b).

Another key feature of cortical design is that with its self-organizing principles rather than being taught what to compute, the cortex is able to solve complex problems with just a few layers, e.g. four, as described in Chapter 2 and by Rolls (2016b). The cortex is constrained to a few layers in the hierarchy, because each layer takes about 15 ms to compute including recurrent attractor operations, and the total compute time needs to be kept to less than 100 ms for biological utility. More layers could be tried in artificial self-organizing neural networks.

I hope that those interested in brain-inspired intelligence will look carefully at what the computational problems are that are solved by different parts of the brain, and will not only learn from our current understanding of how the brain performs its computations, but will also suggest new ways in which those computations may be performed. In this context, Cerebral Cortex: Principles of Operation (Rolls, 2016b) does set out principles of the organization of the brain many of which underlie its computational abilities and style.

There are other factors that make deep networks trained by backprop biologically unlikely. Given that the error for a hidden neuron in an error backpropagation network is calculated by propagating backwards information based on the errors of all the output neurons to which a hidden neuron is connected, and all the relevant synaptic weights, and the activations of the output neurons to define the part of the activation function on which they are operating, it is implausible to suppose that the correct information to provide the appropriate error for each hidden neuron is propagated backwards between real neurons. A hidden neuron would have to ‘know’, or receive information about, the errors of all the neurons to which it is connected, and its synaptic weights to them, and their current activations (Section B.12). If there were more than one hidden layer, this would be even more difficult.

To expand on the difficulties: first, there would have to be a mechanism in the brain for providing an appropriate error signal to each output neuron in the network. With the possible exception of the cerebellum, an architecture where a separate error signal could be provided for each output neurons is difficult to identify in the brain. Second, as noted in Chapter 9 and by Rolls (2016b), the backprojection pathways that are present in the cortex seem suited to perform recall, and this would make it difficult for them also to have the correct strength to carry the correct error signal.

A problem with the backpropagation of error approach in a biological context is thus that in order to achieve their competence, backpropagation networks use what is almost certainly a learning rule that is much more powerful than those that could be implemented biologically, and achieve their excellent performance by performing the mapping through a minimal number of hidden neurons. In contrast, real neuronal networks in the brain probably use much less powerful learning rules, in which errors are not propagated backwards, and at the same time have very large numbers of hidden neurons, without the bottleneck that helps to provide backpropagation networks with their good performance. A consequence of these differences between backpropagation and biologically plausible networks may be that the way in which biological networks solve difficult problems may be rather different from the way in which backpropagation networks find mappings. Thus the solutions found by connectionist systems may not always be excellent guides to how biologically plausible networks may perform on similar problems. Part of the challenge for future work is to discover how more biologically plausible networks than backpropagation networks can solve comparably hard problems, and then to examine the properties of these networks, as a perhaps more accurate guide to brain computation.
As stated above, it is a major challenge for brain research to discover whether there are algorithms that will solve comparably difficult problems to backpropagation, but with a local learning rule. Such algorithms may be expected to require many more hidden neurons than backpropagation networks, in that the brain does not appear to use information bottlenecks to help it solve difficult problems. The issue here is that much of the power of backpropagation algorithms arises because there is a minimal number of hidden neurons to perform the required mapping using a final one-layer delta-rule network. Useful generalization arises in such networks because with a minimal number of hidden neurons, the net sets the representation they provide to enable appropriate generalization. The danger with more hidden neurons is that the network becomes a look-up table, with one hidden neuron for every required output, and generalization when the inputs vary becomes poor. The challenge is to find a more biologically plausible type of network that operates with large numbers of neurons, and yet that still provides useful generalization. An example of such an approach is described in Chapter 2.

Limitations of current understanding of deep learning networks and how different they are from brain function have been highlighted by others (Plebe and Grasso, 2019; Sejnowski, 2020).

19.5 Reinforcement Learning

Reinforcement learning is another approach to understanding the function of some brain areas that has its origin in the machine learning literature (Sutton and Barto, 1998). Reinforcement learning is described in Section B.16, and evidence that the dopamine neurons encode a reward prediction error signal is described in Section 14.4.3 (Schultz, 2013, 2016c,b, 2017). Some evaluation about how it fits in with brain design follows.

Reinforcement learning utilizes a single reward or reinforcement signal to train a whole network. This is in contrast with backpropagation of error networks, which specify what the output should be of every output neuron for a given input pattern, which is a much more information-rich training signal. Reinforcement learning is thus in a sense simpler in terms of evolutionary design, because the single training signal can be broadcast widely in the neural network, without having to tell individual neurons how to respond, which is biologically implausible.

Reinforcement learning uses a widely broadcast reward prediction error signal to systems in the brain such as the basal ganglia. If there has been very recent neuronal activity at synapses that reflect coactivity of for example stimulus encoding and response encoding neurons, then those synaptic connections are strengthened if the dopamine input is high (see Chapter 14). In this way, stimulus-response habits may be set up in the basal ganglia. A property is that the reward is not part of what is learned, just the stimulus-response connection. So if the reward is devalued, the stimulus-response connection is still there, and the response is elicited when the stimulus occurs, even though the stimulus is no longer a goal for action. A further property is that the dopamine neurons seem to convey just a reward prediction error signal, and do not specify the type of reward, such as food, water, novelty, money etc. So lots of habits may be set up without the goal for which the habit may have been set up specified in the stimulus-response habit.

This is very different from the type of learning implemented in the orbitofrontal cortex to anterior cingulate cortex learning system to perform actions to obtain goals (with the goals the rewards specified by the orbitofrontal cortex) (Rolls, 2019c) (Chapter 12). Here there is a potentially very precise specification of the outcome, and different actions can be learned to obtain different goals, even if the actions are very similar, but the outcomes are different.
That is, a precise association is learned between a particular action and a particular outcome. Because the cingulate cortex initiates actions to obtain the goals or rewards specified by the orbitofrontal cortex (Rolls, 2019c), if the reward/goal stimulus is devalued, actions will not be performed. Moreover, the stimulus value can be reset in the orbitofrontal cortex by for example relearning or reversal of a stimulus-reward association, and that learning can be on one trial, again allowing great flexibility of action-outcome learning as implemented in the orbitofrontal cortex / anterior cingulate cortex system (Chapter 12 and Rolls (2019c)).

In comparison, with reinforcement learning and its very general reward prediction error signal, the learning tends to be slow, so that regular associations between a stimulus and a response can be evaluated, and stamped in by the dopamine modulator. If the system is not slow, then the reinforcement learning system risks setting up incorrect stimulus-response associations, because of chance stimulus and response co-occurrence, or because of noise-related effects in the brain.

The orbitofrontal cortex reward evaluation system, followed by an action-outcome system (Rolls, 2019c), thus appears to be a much more powerful learning system than reinforcement learning using a general reward prediction error signal. The reinforcement learning may have been appropriate at early stages of evolution as it does not require a memory of the response that has just been made apart from a simple eligibility signal for synapses that have recently been active. The striatum and globus pallidus do not implement a short-term memory that could be used for a longer term ‘eligibility trace’, because the basal ganglia have no recurrent collateral connections to support a short-term memory (Chapter 14). On the other hand, the cingulate cortex as it is neocortex has a highly developed set of recurrent collaterals, and may therefore be much better suited to learning, when the actions may be followed by a delay before the outcome (reward) arrives (Chapter 12 and Rolls (2019c)).

A comparison of the computations involved in reversal learning helps to show why reinforcement learning as implemented in the basal ganglia using dopamine for stimulus–response habit learning is much less efficient than the orbitofrontal cortex – cingulate cortex action–outcome learning system. The learning considered is a visual discrimination task in which a choice of one visual stimulus leads to reward, and of the other to punishment / loss. When this occurs for the dopamine system, the reward prediction errors change, but new stimulus-response connections need to be learned. This may even mean a return to trial and error learning, where many different responses may be tried until a response is found that when following the stimulus leads to a positive reward prediction error. This then takes some time to train the system, for the reasons given above. In contrast, for the orbitofrontal cortex – cingulate cortex action–outcome system, when the reward association of the visual stimuli reverses, on the very next trial after an error, the orbitofrontal cortex expected value representations reverse. Accordingly, the next time a visual stimulus is shown, the correct expected value is sent to the cingulate cortex, which then performs the same action that it has already learned for that expected value input received from the orbitofrontal cortex. (This is described in Chapter 11.) So part of the power of the orbitofrontal cortex – cingulate cortex system is that it is a two stage learning process, and the visual stimulus to reward association process, a stimulus–stimulus learning process implemented in the orbitofrontal cortex, can reverse in one trial, with no new action learning needed.

Additionally, as described above, the cingulate cortex is likely to be able to remember actions for longer times until the outcome becomes known when new actions need to be learned to obtain rewards; and can match actions to particular outcomes (such as food, water, money etc), whereas the reinforcement learning system uses just a single general reward prediction error. It also seems likely that the cingulate cortex has a repertoire of previously learned actions, which can rapidly be associated with new different outcomes by rapid associative learning between actions and outcomes (Rolls, 2019c).
For these reasons, I argue that the orbitofrontal cortex—cingulate cortex action–outcome learning system is a much more powerful way that the brain uses to deal with reward-related learning than the reinforcement learning stimulus-response dopamine reward prediction error habit system. Consistent with this, it is damage to the orbitofrontal cortex that has a great influence on our emotional and reward-related behaviour and their disorders including depression, as described in Chapter 11 and elsewhere (Rolls, 2018a, 2019e), whereas the effects of damage to the dopamine system produce major motor problems including Parkinson's disease. In this context, we should remember that stimulus-response learning is a simple type of motor learning.

I have also argued that reward prediction error neurons do not themselves relate to emotion, because emotions can still occur to rewarding or punishing stimuli even when there is no new learning and no reward prediction error (Rolls, 2018a, 2019e) (Chapter 11). That is, emotions are states elicited by rewards or punishers, not by reward prediction errors (Rolls, 2014a, 2018a, 2019e) (Chapter 11).

I thus argue that reinforcement learning in the brain despite great interest in it (Schultz, 2013, 2016c,b, 2017; O'Doherty et al., 2017) is a rather specialized brain system, and that other processes, described in Chapters 11 and 12 are involved in much reward and goal-related behaviour, and in emotion.

The reinforcement learning algorithm can be applied in deep networks (Botvinick et al., 2019), that is, in a network with one or more hidden layers (see Appendix B). Indeed, artificial neural networks of this type have become popular in Artificial Intelligence, and can perform well in Atari video games, Go, and Capture the Flag (Botvinick et al., 2019). However, such systems do require a great deal of training, so produce slow learning. Variations have been tried, such as incorporation of a type of episodic memory about rewards received in previous similar situations, and that can speed up the learning (Botvinick et al., 2019). However, if reinforcement learning in the brain is used mainly in the basal ganglia, then the type of problem to which it is applied is at least usually rather simple, stimulus-response habit learning (see Chapter 14). On the other hand, use of a hippocampal episodic memory system to remember the outcome when a previous situation occurred might be a solution that could be used that would not require reinforcement learning.

19.6 Levels of explanation, and the mind-brain problem

We can now understand brain processing from the level of ion channels in neurons, through neuronal biophysics, to neuronal firing, through the computations performed by populations of neurons, and how their activity is reflected by functional neuroimaging, to behavioural and cognitive effects (Rolls, 2016b; Rolls and Deco, 2010; Rolls, 2014a). Activity at any one level can be used to understand activity at the next. This raises the philosophical issue of how we should consider causality with these different levels (Rolls, 2020a, 2012e). Does the brain cause effects in the mind, or do events at the mental, mind, level influence brain activity?

What is the relation between the mind and the brain? This is the mind–brain or mind–body problem. Do mental, mind, events cause brain events? Do brain events cause mental effects? What can we learn from the relation between software and hardware in a computer about mind–brain interactions and how causality operates? Neuroscience shows that there is a close relation between mind and matter (captured by the following inverted saying: ‘Never matter, no mind’).

My view (Rolls, 2016b, 2020a) is that the relationship between mental events and neurophysiological events is similar to the relationship between the program running in a computer
Levels of explanation, and the mind-brain problem

and the hardware of the computer. Does the program (the software loaded onto the computer usually written in a high-level language such as C or Matlab) ‘cause’ the logic gates (TTL, transistor-transistor logic) of the hardware to move to the next state? And does this hardware state change ‘cause’ the program to move to its next step or state?

I propose that one way to think about this is that when we are looking at different levels of what is overall the operation of a system, causality can usefully be understood as operating within levels (causing one step of the program to move to the next; or the neurons to move from one state to another), but not between levels (e.g. software to hardware and vice versa). That is, if the events at the different levels of explanation are occurring simultaneously, without a time delay, then my view is that we should not think of causality as operating between levels, but just that what happens at a higher level may be an emergent property of what happens at a lower level. This is the solution I propose to this aspect of the mind-brain problem (Rolls, 2020a).

Following this thinking, when one step of a process at one level of explanation moves to the next step in time, we can speak of causality that would meet the criteria for Granger causality where one time series, including the time series being considered, can be used to predict the next step in time (Section 16.3) (Granger, 1969; Ge, Feng, Grabenhorst and Rolls, 2012; Bressler and Seth, 2011). In contrast, when we consider the relationship between processes described at different levels of explanation, such as the relation between a step in the hardware in a computer and a step in the software, then these processes may occur simultaneously, and be inextricably linked with each other, and just be different ways of describing the same process, so that temporal (Granger) causality does not apply to this relation between levels, but only within levels. The whole processing can now be specified from the mechanistic level of neuronal firings, etc. up through the computational level to the cognitive and behavioural level.

Thus my view of the mind–brain issue is that we are considering the process as a mechanism with different levels of explanation. As described above, we can now understand brain processing from the level of ion channels in neurons, through neuronal biophysics, to neuronal firing, through the computations performed by populations of neurons, to behavioural and cognitive effects, and even perhaps to the phenomenal (feeling) aspects of consciousness (Rolls, 2020a, 2016b). The whole processing is now specified from the mechanistic level of neuronal firings, etc. up through the computational level to the cognitive and behavioural level. Sometimes the cognitive effects seem remarkable, for example the recall of a whole memory from a part of it, and we describe this as an ‘emergent property’, but once understood from the mechanistic level upwards, the functions implemented are elegant and wonderful, but understandable and not magical or poorly understood (Rolls, 2016b, 2012e).

The point I make is that however one thinks about causality in such a mechanistic system with interesting ‘emergent’ computational properties, the system is now well-defined, is no longer mysterious or magical, and we have now from a combination of neuroscience and analyses of the type used in theoretical physics a clear understanding of the properties of neural systems and how cognition emerges from neural mechanisms. There are of course particular problems that remain to be resolved with this approach, such as that of how language is implemented in the brain, but my point is that this mechanistic approach, supported by parsimony, appears to be capable of leading us to a full understanding of brain function, cognition, and behaviour. However, the property of phenomenal consciousness is a big step for an ‘emergent property’, and that hard problem is therefore considered elsewhere (Rolls, 2020a, 2016b).
19.7 Levels of explanation, and levels of investigation

It will be evident from the approach taken in this book that to understand how the brain works, we need to understand what is computed in each brain area, as well as how it is computed. It should be evident that the level at which information is exchanged between the computing elements of the brain is the firing of single neurons, with the evidence set out by Rolls (2016b).

It is important to investigate this at the level of single neurons, and populations of single neurons, because neurons often have relatively independent responses as is made clear in Appendix C, so that it is only possible to know what is represented in a brain area by utilising evidence from neuronal firing.

Operating at a higher level, such as the signals measured in functional neuroimaging that reflect the activity of thousands of neurons, does not lend itself well to making computational models of how the brain works, because one does not know exactly what is represented at each successive stage. Thus methods such as fMRI may be useful for obtaining some evidence on the functions of different brain areas, but do need to be complemented where possible by evidence on the firing of neurons (which of course is often not possible in humans).

The architecture of the cerebral cortex is also considerably localised, with among the evolutionary reasons that this design minimises the connection length between neurons, and also the genetic information needed to specify cortical design, as set out in Cerebral Cortex (Rolls, 2016b). Those are among the reasons why there are hierarchically organised networks in the brain, with the major connections between adjacent cortical areas in the hierarchy. The level of granularity here corresponds to for example V1, V2, V4, posterior inferior temporal cortex, and anterior inferior temporal cortex, with frequently cytoarchitectural difference between such areas leading to classifications at about the level of Brodmann areas. There is also a finer level of granularity, specified by the radial extents of cortical pyramidal cell dendrites and their axons, which have a high density with a radius of 2–3 mm, and provide a basis for local cortical attractor networks (Rolls, 2016b).

The level of analysis just described, in which the different computational roles of each brain area are addressed using methods such as comparing neuronal activity in different stages of the processing, and double dissociation of the effects of brain lesions or inactivation of neuronal populations, sits uncomfortably with brain networks identified with some neuroimaging approaches in which correlations between the BOLD signals of many brain areas are found, leading to identification of a few brain networks (Ryali, Supekar, Chen, Kochalka, Cai, Nicholas, Padmanabhan and Menon, 2016). If these reflect just descriptions of correlations, that is one thing, but when simple functions are attributed to such networks, then that may be an oversimplification of exactly what is computed in each brain region, which is the different approach taken here.

For example, these large-scale networks identifiable with neuroimaging correlation-based measures have been described as follows: “The Saliency Network (SN) (Menon and Uddin, 2010) is a limbic-paralimbic network anchored in the anterior insula and dorsal anterior cingulate cortex with prominent subcortical nodes in affective and reward processing regions including the amygdala and ventral striatum. The Saliency Network plays an important role in orienting attention to behaviorally and emotionally salient and rewarding stimuli and facilitating goal-directed behavior. The fronto-parietal Central Executive Network (CEN) is anchored in the dorsolateral prefrontal cortex and supramarginal gyrus and is critical for actively maintaining and manipulating information in working memory. The Default Mode Network (DMN) is anchored in the posterior cingulate cortex, medial prefrontal cortex, medial temporal lobe, and angular gyrus and is involved in self-referential mental activity and
autobiographical memory” (Ryali, Supekar, Chen, Kochalka, Cai, Nicholas, Padmanabhan and Menon, 2016).

To take the Saliency Network (Menon and Uddin, 2010) as an example, it is clear that the dorsal anterior insula is the primary taste cortex, and that there is an area ventral to it that is probably visceral/autonomic cortex (Rolls, 2016c; Critchley and Harrison, 2013; Hassanzadeh et al., 2018), and is shown in the ventral parts of the anterior insula in Fig. 2 of Baylis, Rolls and Baylis (1994). The autonomic cortex during salient and emotional events is likely to produce autonomic output (alterations of heart rate and variability, sweating, and skin conductance), and far from ‘altering attention to behaviorally and emotionally salient and reward stimuli’, it is likely to receive outputs about such stimuli from the orbitofrontal cortex, which has strong connections to it, so that the anterior insular cortex provides a pathway to autonomic output (Rolls, 2016c, 2019e). The taste insula of primates is not involved in taste reward according to the experimental evidence (Rolls, 2016c, 2019e) (Chapter 4). And much of the rest of the insula is involved in somatosensory processing (Chapter 6). And the anterior cingulate cortex receives reward and punishment information from the orbitofrontal cortex, and uses this reward outcome information for action-outcome learning (Rolls, 2019c) (Chapter 12).

So it appears that a lot is lost if simple functions are ascribed to these large-scale human cortical networks such as the saliency network, and such concepts may not have much explanatory value. Instead the goal that I advocate is to identify what computations are performed by each cortical area. Each area will of course transmit information to other specialised cortical areas. And together in their complementary ways they will produce a particular computation of a brain system.

19.8 Brain-Inspired Intelligence

One of the aims of this book is to provide for neuroscientists a computational approach to brain function. This is I believe very important, for it takes us beyond metaphors and analogies and word-level descriptions about how the brain works, to well-defined computational hypotheses that allow testing at many levels of explanation, and indeed for the relations between the levels to be understood, as set out in Section 19.6.

Another key aim is to set out for those interested in machine intelligence, what is computed in different brain systems, so that those who wish to learn from or emulate brain function know what each part computes. That is an essential prerequisite for advances to be made, which might not only try to emulate brain function, but to be inspired by it, with possible new ways of solving similar computational problems.

This book also sets out many current hypotheses about how the brain solves these computational problems, and this is likely to be an area in which many advances can be made. Indeed, advances in our understanding of how the brain computes are likely to be facilitated by very large-scale simulations of parts of the brain, to test whether the computational hypotheses scale up appropriately, as well as enabling the machine implementations to be on the same, or a larger, scale than in the real brain. This should thus provide for fruitful advances to be made in both directions, from understanding the brain to machine intelligence; and from large scale simulation on large computers of well formulated computational hypotheses about how the brain computes. But all of this rests on evidence on what is computed in different brain regions, and how it is computed.

Advances in this area are made possible by the way in which the brain is designed compared to a digital computer, as summarized in Section 19.3 and in Cerebral Cortex: Principles of Operation (Rolls, 2016b). Key aspects that help us to understand what and how it computes
include the following. Brain functions are considerably localized, due to the importance of minimizing the length of neural connections, and of the simplicity that this offers in the genetic specification of the brain used while it develops, which often relies on proximity rules (Rolls, 2016b). Another key aspect is that the information is transmitted from the basic computing elements, the neurons, by spike trains with one output for each neuron, so that we can listen in to what is being transmitted from one brain area to another, using the methods described in Appendix C. Further, by making more advances in understanding how genes specify the architecture of the brain (Rolls and Stringer, 2000; Rolls, 2016b), it will be possible in future to use genetic algorithms to design the architecture of brain-like computing devices, which will enable the development of new types of computational solution to problems. Indeed, we were enormously surprised by the solutions found by neural architectures built by genetic algorithms, which frequently found ways to ‘cheat’, to operate in ways that we had not expected, in order to succeed at the fitness function that we had set (Rolls and Stringer, 2000). And this was when all that the genetic algorithm had to work with were the building blocks of how genes might specify differences between neural architectures, in contrast to being allowed to explore new specification rules.

It seems as if we have reasonable approaches now, since the start in about 1970, of an understanding of what is computed in many brain systems in the macaque brain, and much of that and the complementary information about the human brain is what has been described in this book. We also have at least the start to approaches about how different brain systems compute. The area of brain function about which we know computationally the least is how language, including syntax and also semantics, are implemented in the brain, and how they operate together. That is an area in which at the computational level our understanding is still very incomplete, and where major advances are needed.

### 19.9 Brain-Inspired Medicine

There are many applications showing how advances in understanding brain function and computation have led to advances relevant to medicine. Some are highlighted in the following sections. Some others include the use of deep brain stimulation for Parkinson’s disease (Little and Brown, 2020), the use of antiphase brain stimulation to cancel for example tremor, and the identification that there are two subtypes of Parkinson’s disease, based on an analysis of functional connectivity (Wang, Cheng, Rolls, Dai, Gong, Du, Zhang, Wang, Liu, Wang, Brown and Feng, 2020b).

#### 19.9.1 Computational psychiatry and neurology

As shown in Chapter 18, a number of mental disorders can be understood at least partly in terms of stability of different brain systems. This may be understood in the broader context that stability of a cortical system with positive feedback between its computing elements, the neurons, is a difficult issue to manage. This is compounded by the hypothesis that natural selection may be operating on parameters that influence the stability of different brain systems, in order to search for advantages of different options in different environments or situations.

For example, one strategy might be to make the system very sensitive to non-reward, as this may be adaptive in a number of obvious ways; yet too much sensitivity might lead to depression. Having a smaller degree of sensitivity to non-reward might also be a useful strategy in some circumstances, for it would result in continued trying, even in the face of adversity. This could result in genetic variation in sensitivity to, for example, non-reward, being maintained in the population.
The computational accounts being developed for a number of mental disorders have implications for medical treatment. For example, in depression approaches might be tried of ways to reduce the activity or connectivity of the non-reward lateral orbitofrontal cortex, and to increase the sensitivity or connectivity of the reward-related medial orbitofrontal cortex. In this context, it is interesting that modern antidepressant drugs do reduce the connectivity of the lateral orbitofrontal cortex back down towards levels in controls; but do not restore the reduced connectivity of the reward-related medial orbitofrontal cortex in depression. That sets a new target for the development of further treatments for depression aimed to facilitate the operation of the reward-related medial orbitofrontal cortex. Similarly, the understanding of brain mechanisms of emotion described in this book suggests that a key brain system in emotion is the orbitofrontal cortex; and that points the way to possible brain stimulation of the orbitofrontal cortex as a new brain area to explore for the possible relief of otherwise untreatable depression.

However, the computational approach is helpful in another way in understanding and potentially treating some mental disorders, and these treatments might be at the purely behavioural level. For example, the modern computational approach to understanding possible causes of depression indicates that non-reward inputs in which no action is possible might lead to feelings of sadness and depression (Fig. 11.20). In this situation, advice at the behavioural level might be to avoid as much as is reasonable stimuli that are associated with non-reward. In the case of the loss of a loved one, this might lead for example to attempts to divert the mind to other more rewarding thoughts, and divert activities from being constantly focussed on the loss. These types of approach and many others to depression are described further in my book *The Brain, Emotion, and Depression* (Rolls, 2018a).

A similar point can be made about attention deficit hyperactivity disorder (ADHD). We have found that the functional connectivity of the visual thalamic nucleus, the lateral geniculate, with early visual cortical areas is higher in ADHD than in controls (Rolls et al., 2020a). An implication is that this may make such individuals more easily distracted and dominated by visual stimuli in the environment, and indeed we have shown that people with ADHD do display very much more screen use (i.e. use of mobile phones, computers, computer games, television, etc) (Du et al., 2020a). There are implications from this neurocomputational approach for treatment at the purely behavioural level, for by understanding this sensitivity of some parts of the brain, advice might include limiting screen use to a reasonable level, and providing a quiet atmosphere for work without environmental distractions. There are opposite differences in schizophrenia, with decreased connectivity of thalamic with visual cortex connectivity (as well as overconnectivity of backprojections from areas concerned with the self and internal thoughts such as the precuneus), as described in Section 18.3 and by Rolls et al. (2020a), and that may also have implications for treatment.

In addition, understanding with computational methods the connectivity of the brain can help to identify subtypes of a disorder. For example, subtypes of depression have been found with and without anxiety as a comorbid condition (Drysdale et al., 2017), and this may be helpful for treatment, as it may be that the subgroups with anxiety are particularly helped by Transcranial Magnetic Stimulation of the lateral orbitofrontal cortex (Feffer et al., 2018).

### 19.9.2 Reward systems in the brain, and their application to understanding food intake control and obesity

What is computed in reward systems in the brain, and how it is computed, are described in Chapters 4–7, 11, 12, and 14. In this Section, we consider some applications of this understanding.

Different individuals have different sensitivities to rewards, non-reward, etc, and indeed