Introduction to Biological Psychology

INTRODUCTION TO BIOLOGICAL PSYCHOLOGY

Edited by Catherine N. Hall

UNIVERSITY OF SUSSEX LIBRARY

DR ELEANOR J. DOMMETT; DR JIMENA BERNI; PROFESSOR PETE CLIFTON; HANS CROMBAG; PROFESSOR CLAIRE GIBSON; DR PALOMA MANGUELE; DR EMILIANO MERLO; DR KYRIAKI NIKOLAOU; PROFESSOR JOSE PRADOS; DR BRYAN F. SINGER; PROFESSOR HARRIET ALLEN; DR ANDREW YOUNG; DR CATHERINE LAWRENCE; AND DR CATHERINE HALL

CONTENTS

Part VI. Dysfunction of the nervous system

CREDITS

Editor

Dr Catherine Hall

Managing Editor

Dr Catrina Hey

Scientific illustrator

Dr Eliza Wolfson

Contributing authors

Professor Harriet Allen; Dr Jimena Berni; Professor Pete Clifton; Professor Hans Crombag; Dr Eleanor J. Dommett; Professor Claire Gibson; Dr Catherine N. Hall; Dr Catherine Lawrence; Dr Paloma Manguele; Dr Emiliano Merlo; Dr Kyriaki Nikolaou; Jose Prados; Dr Bryan F. Singer; Dr Andrew Young

Reviewers

Dr Maria Hadjimarkou; Dr Catherine Hall; Professor Sarah King; Dr Liat Levita

Joseph Henderson; Letitia McMullan; Magali Ostriviecki; Kira Shaw; Harry Trewhitt

Picture editor

Sally Hendergate

References editor

Jack Coull

FOREWORD

Libraries have an important role to play in supporting open initiatives, extending beyond journals and transformative agreements to include all types of scholarly outputs and their underlying infrastructure. The University of Sussex Library is committed to actively supporting innovation in open publication, and recognises that *open* has as much value for education and students as it does for research.

We are proud to present our first open textbook, the result of a pilot project run by the Library in collaboration with authors from Sussex and other UK institutions. We are hugely grateful for their contributions of time and expertise. The outcome represents an exciting change to the way the University of Sussex community creates, shares and values scholarly textbooks. These are the first steps in the development of an institutional open publishing infrastructure which offers the opportunity to invest in our scholarship, and drive a cultural shift to a form of publishing that is more equitable and sustainable.

The textbook is designed to be used as primary reading for undergraduate Psychology students studying core biological psychology modules and we invite course leaders at other

institutions to adopt the textbook for teaching. We hope that teachers and students alike will find it useful.

Jane Harvell, University Librarian and Director of Library Services
Brighton, UK
February 2023

PART I
BACKGROUND TO BIOLOGICAL PSYCHOLOGY

When we study biological psychology, we are interested in the biological processes that shape how our brains create our minds, thereby generating who we are and what we do – our sense of self and our behaviour. In this introductory section of the textbook, which consists of a single chapter, we will explore three distinct aspects of biological psychology.

We begin with a brief survey of the ways in which our understanding of the relationship between the mind and our physical body, especially the brain, has changed over time. Almost all biological psychologists now take a broadly materialistic view which assumes that the mind, once seen as a quite separate entity from the body, is simply another aspect of the physical functioning our brain.

We then explore the methods used to investigate the relationship between brain function and behaviour. Although contemporary neuroscientists have developed techniques that permit us to monitor, and potentially interfere with, brain function in ways that were unimaginable only twenty or thirty

years ago, there are still fundamental limitations which it is important to understand.

All scientific study is subject to ethical constraints. Psychology as a discipline has developed a strong research ethics code and in the final section of the chapter we explore how this is reflected in studies that use either human or non-human animals.

There is also a brief postscript which introduces three key concepts from biology (cells , inheritance and evolution) that you may find helpful.

Learning Objectives

By the end of this section you will be able to:

- briefly describe the way in which our understanding of the relationship between brain and behaviour has evolved in the last two millennia
- understand how experimental approaches to investigating the relationship between brain and behaviour can be used

- appreciate some of the limitations of techniques that use either correlational techniques or experimental manipulations
- reflect on how the broad ethical principles that underpin psychological research apply in the context of biological psychology.

1.

INTRODUCTION TO BIOLOGICAL PSYCHOLOGY

Professor Pete Clifton

Research at the interface of biology and psychology is amongst the most active in the whole of science. It seeks to answer some of the 'big' questions that have fascinated our species for thousands of years. What is the relationship between brain and mind? What does it mean to be conscious? Developments in our understanding of conditions such as depression or anxiety that can severely impact on our quality of life give hope for the development of more effective treatments.

One reason for rapid advances in biological psychology has been the technological progress that has provided tools to study brain processes in ways that were unimaginable just a few decades ago. It easy to forget that it was only 130 years ago, in the late nineteenth century, that Santiago Ramón y Cajal, Camillo Golgi and others were describing the detailed structures of nerve cells in the brain.

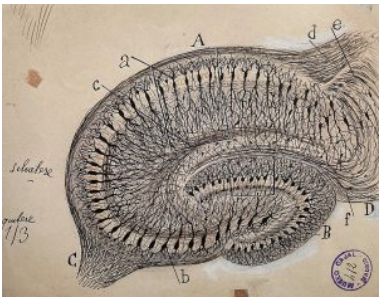

Fig 1.1. The hippocampus is a critical structure for learning and memory (see 'Exploring the brain'). This pen and ink drawing by Cajal, which emphasises pyramidal cells and their connections, probably dates from the late 1890s and shows his skill as an artist.

In addition to his skill in describing the detailed structure of the nervous system, Cajal made crucial theoretical advances and, for that reason, is often referred to as 'the father of neuroscience'. He was the first to argue clearly that transmission of the nerve impulse is in one direction only and that individual neurons communicate at specialised structures called synapses. The detailed mechanisms that underlie communication – the nervous system, the action potential, and synaptic transmission – were not elucidated until the 1950s. At about the same time, records of the firing of single cells in the nervous system, especially within those parts of the brain that receive sensory stimuli, began to suggest the way in which information was coded and processed.

Fig 1.2. This photograph of the artist is a self-portrait from about the same time.

At that time the experimental methods for investigating the contribution of particular brain structures to aspects of behaviour were crude. They essentially involved permanently destroying ('lesioning') a few square millimetres of brain tissue containing several hundred thousand nerve cells, and then examining the effects on behaviour. Since then, there have been remarkable advances in our ability to study brain mechanisms and their relationship to behaviour. It is now possible to record the activity of many cells simultaneously. There are techniques that permit modulation of the activity of groups of nerve cells with identified functions for a short period of time before allowing them to return to normal functioning. As you will read in later chapters of this book, this has made possible a much greater understanding of the workings of the brain under normal circumstances, and the ways in which its functioning may be disturbed in different disease states.

As you read through this book it may seem as though the brain is a terribly 'tidy' organ. That, at least, is the impression that you might easily get as you look at the drawings and pictures that illustrate the text. But it is worth reflecting on how it feels to encounter a soft, jelly-like, living brain in

practice. Here are a couple of sentences from the neurosurgeon Henry Marsh, describing the way in which he uses a small suction device to gradually approach, and then remove, a tumour located towards the centre of his patient's brain. This particular tumour was located in the pineal gland, a structure with a long and fascinating history in neuroscience. He writes:

> *I look down my operating microscope, feeling my way downwards through the soft white substance of the brain, searching for the tumour. The idea that my sucker is moving through thought itself, through emotion and reason, that memories, dreams and reflections should consist of jelly, is simply too strange to understand. (Do No Harm: Stories of Life, Death and Brain Surgery. Weidenfeld & Nicolson, 2014).*

Mind and brain: a historical context

Heart or brain as the basis for thought and emotion?

During much of recorded human history, there was uncertainty about whether the heart or the brain was responsible for organising our behaviour. The early debates on this topic are best recorded in the works of the Greek philosophers, because, to some extent at least, complete texts are available in a way that is not true for most other ancient

human civilisations. In the 700 years from about the 5th century BCE to the 2nd century CE, the Greeks put forward two contrasting views.

One group of philosophers, of which **Aristotle** (~350 BCE) is the best known, held that it was the chest, most likely the heart, that was crucial in organising behaviour and thought.

Fig 1.3. Plato (left) holding the Timaeus, and Aristotle, holding the Ethics

The heart was the seat of the mind, and the brain had no important role other than, perhaps, to cool the blood. In threatening or exciting situations it is changes in heart function that we become consciously aware of, while we have no conscious access to the changes in brain activation that are occurring at the same time. So it is hardly surprising that the heart was ascribed 'emotional' and 'cognitive' functions, and that it still features in our communication and language in a way that the brain does not. Apologies can be 'heartfelt'; the universal emoji for affection and love is a heart; Shakespeare has Macbeth conclude that 'False face must hide what the false heart doth

know' (*Macbeth* 1.7.82) as he resolves to murder King Duncan and take the Scottish throne for himself.

In fact, there was a grain of truth in these ideas. Experiments conducted by Sarah Garfinkel and others in the last few years have demonstrated that pressure receptors in the arteries that lead from the heart become active on each heart contraction and can influence the processing of threat-related stimuli (Garfinkel and Critchley, 2016). It isn't just the heart that can affect the way in which emotionally salient stimuli are processed by humans – abnormal stomach rhythms enhance the avoidance of visual stimuli, such as faeces, rotting meat, or sour milk, that elicit disgust (Nord et al., 2021). It is interesting that William James, whose *Principles of Psychology* (1890) is regarded as one of the foundational texts of modern psychology, put forward a similar idea in a short paper published in the 1880s (James 1884).

Another group of early Greek writers gave the brain a much more important role, of which **Hippocrates** and **Plato** (4th century BCE) and **Galen** (2nd century BCE) are best known. They developed the idea that, since the brain was physically connected by nerves to the sense organs and muscles, it was also most likely the location of the physical connection with the mind. Galen was a physician whose training was in Alexandria. He moved to Rome and frequently treated the injuries of gladiators. He accepted Hippocrates' argument about the importance of the brain in processing sensory information and generating behavioural responses. The

immediate loss of consciousness that could be produced by a head injury confirmed his view. He also performed experiments that demonstrated the function of individual nerves; he showed that cutting the laryngeal nerve, which runs from the brain to the muscles of the larynx, would stop an animal from vocalising. He rejected another earlier belief that the lungs might be the seat of thought, suggesting that they simply acted as a bellows to drive air through the larynx and produce sounds.

In subsequent centuries Galen's insights were forgotten in much of western Europe, but remained very influential in the Islamic world. **Ibn Sina** (still sometimes known as Avicenna, the Latinised version of his name) was born in 980 CE, and is generally acknowledged as the greatest of the physicians of the Persian Golden Age. He refined and corrected many of Galen's ideas. He provided a detailed description of the effects of a stroke on behaviour, and correctly surmised that strokes might either result from blockages or bursts in the circulation of blood to the brain. He studied patients suffering from a disorder similar to severe depression that, at the time, was called the 'love disorder'. When treating someone with this 'disorder' he used changes in the heart's pulse rate as a way of identifying the names of individuals that were especially significant to them. He also, as might be expected of a physician, had a detailed knowledge of the effects of plant-derived drugs such as opium, the dried latex from poppy seed heads of which the active component is morphine (Heydari et

al., 2013). He knew that it was especially valuable in treating pain, but had serious side effects including suppression of breathing and, with long term use, addiction (see Chapters 6 and 15).

Fig 1.4. A 16th-century representation of Galen, Ibn Sina and Hippocrates with Ibn Sina (Avicenna) in the centre

Mind and body

In the intellectual ferment of post-Reformation Europe, the relationship between the mind and body, especially in the context of what we can know with certainty to be true, became a subject of great controversy.

Fig 1.5. This portrait of Descartes is in the Louvre, Paris, and is ascribed to the Dutch painter Frans Hals.

René Descartes was a French philosopher whose most influential work in this area, the *Discourse on Method*, was published in 1637. He reached the conclusion that we can never be certain that our reasoning about the external world is correct, nor can we be sure that most of our own experiences are not simply dreams. However, he argued, the one thing we can be certain of is that, to use Bryan Magee's free translation, *I am consciously aware, therefore I know that I must exist:* 'Si je pense, donc je suis' in the original French, or famously '**Cogito, ergo sum'** in the later Latin translation (Magee, 1987).

This became the basis for Descartes' argument for **dualism**. However, he also recognised that mind and body had to interact in some way. In his book *De homine* [About humans] written in 1633 but not published until just after his death, he suggested that this might happen through the pineal gland, as the only non-paired structure within the brain. He also thought that muscle action might be produced through some kind of pneumatic mechanism involving the movement of 'animal spirits' from the fluid-filled ventricles within the brain

to the muscles. In this scheme, the pineal gland acted as a kind of valve between the mind and the brain. We now know that the pineal gland is in fact an **endocrine** gland that is important in regulating sleep patterns.

Descartes also described simple reflexes, such as the withdrawal of limb from heat or fire, as occurring via the spinal cord, and not involving the pineal gland (Descartes [1662], 1998). Although many of Descartes' ideas about the way in which the body functioned were incorrect, he was a materialist, in the sense that he viewed the body as a mechanism.

Fig 1.6. Drawing from Descartes' 'On Man' (1662).

Descartes' legend to this drawing (Figure 1.6) reads:

> *For example, if the fire A is close to the foot B, the small particles of fire, which as you know move very swiftly, are able to move as well the part of the skin which they touch on the foot. In this way, by pulling at the little thread cc, which you see attached there, they at the same instant open e, which is the entry for the pore d, which is where this small thread terminates; just as, by pulling one end of a cord, you ring*

> *a bell which hangs at the other end.... Now when the entry of the pore, or the little tube, de, has thus been opened, the animal spirits flow into it from the cavity F, and through it they are carried partly into the muscles which serve to pull the foot back from the fire, partly into those which serve to turn the eyes and the head to look at it, and partly into those which serve to move the hands forward and to turn the whole body for its defense. (de Homine, 1662)*

The common feature of both the heart- and early brain-centred ideas about the relationship between the body and behaviour was of a fluid-based mechanism that translated intentions into behaviour. In the brain-centred view, the fluid in the ventricles (Descartes' 'animal spirits') connected to the muscles by nerves had this function, whereas in the heart-centred account, that role was taken by the blood.

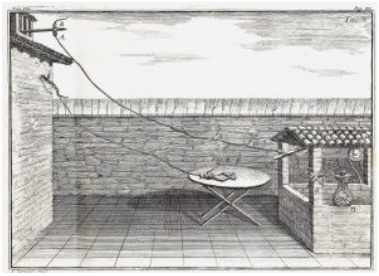

Fig 1.7. This is an early version of Galvani's frog leg experiment (1791) in which the electrical stimulus is provided by lightning during a storm. Later versions used primitive electrical generators. Mary Shelley's novel 'Frankenstein, or the Modern Prometheus' was published in 1816, and she was aware of the attempts by Giovanni Aldini, Galvani's nephew, to reanimate human corpses with electrical stimulation.

Electrical phenomena were documented as long ago as 2600 BCE in Egypt, but it was not until the eighteenth century that they became a subject of serious scientific enquiry. In 1733 Stephen Hales suggested that the mechanism envisaged by Descartes might be electrical, rather than fluid, in nature, although this idea remained very controversial.

By 1791 Luigi Galvani had confirmed electrical stimulation of a frog sciatic nerve could produce contractions in the muscles of a dissected frog leg to which it was connected. He announced that he had demonstrated 'the electric nature of animal spirits'.

In the 1850s Hermann von Helmholtz used the same frog nerve/muscle preparation to estimate that the speed of conduction in the frog sciatic nerve was about 30 metres per second. This disproved earlier suggestions that the nerve impulse might have a velocity that was as fast as, or even faster, than the speed of light! During the course of the twentieth

century, gradual progress was made in understanding exactly how information is transmitted both within and between individual nerve cells. You can read more about this in Chapters 4 and 5.

At about the same time that Galvani and Helmholtz were uncovering the mechanisms of electrical conduction in nerves, there was also great interest in the extent to which different cognitive functions might be localised within the brain. Franz Gall and Johann Spurzheim, working during the first half of the nineteenth century, suggested that small brain areas, mainly located in the cortex, were responsible for different cognitive functions. They also believed that the extent to which individuals excelled in particular areas could be discovered by carefully examining the shape of the skull. Spurzheim coined the term **phrenology** to describe this technique. The idea became very popular in the early 1800s, but subsequently fell out of favour.

Gall and Spurzheim disagreed spectacularly after they began to work and publish independently. After one book by Spurzheim appeared, Gall wrote: 'Mr. Spurzheim's work is 361 pages long, of which he has copied 246 pages from me. .... Others have already accused him of plagiarism; it is at the least very ingenious to have made a book by cutting with scissors' (Whitaker & Jarema, 2017).

In the 1860s, studies by physicians such as Paul Broca, working in Paris, began to correlate the loss of particular functions, such as language, with damage to specific areas of

the brain. They could only determine this by dissections performed after death, whereas today techniques such as computed tomography (CT) and magnetic resonance imaging (MRI) scans reveal structural changes in the living brain. Studies of this kind could indicate that a particular area was necessary for that ability but did not indicate that they were the only areas of importance. The activation of different brain areas while humans perform both simple and complex cognitive tasks including speech can be measured using functional magnetic resonance imaging (fMRI) while the participant lies in the brain scanner.

Broca also noted that loss of spoken language was almost always associated with a lesion in the left cortex and often associated with weakness in the right limbs – the first clear example of cerebral lateralisation. Although he is usually credited with that discovery, Marc Dax, another neurologist working in the 1830s, appears to have made the same observation as Broca, although his data were not published until after his death in 1865 – just a few months before Broca's own more comprehensive publication. Although it was assumed until fairly recently that brain lateralisation was uniquely human and associated with language there is now convincing evidence that it is widespread amongst vertebrates and has an ancient evolutionary origin (Vallortigara & Rogers, 2020).

Broca, like so many of the physicians and scientists discussed earlier, was a man of very wide interests. He published

comparative studies of brain structure in different vertebrates and used them to support Charles Darwin's ideas on evolution. He has been criticised for potentially racist views (Gould, 2006). He believed that different human races might represent different species and that they could be distinguished through anatomical differences in brain size and the ratios of limb measurements. Modern genetic studies demonstrate that living humans are a single species, although there is also evidence that early in our evolutionary history our species interbred with other early, and now extinct, humans including Neanderthals and Denisovans (Bergström et al., 2020).

Positivism and the study of behaviour

In the latter part of the nineteenth century, psychologists often relied on introspection for their primary data. But the French philosopher Auguste Compte who led the positivist movement, argued that the social sciences should adopt the same approach as physical scientists and rely solely on empirical observation. In the study of behaviour this meant relying on recording behaviour either in the laboratory or field. North American psychologists studying conditioning and animal learning including John Watson (of 'Little Albert' fame – see below) and Burrhus Skinner took this approach. Nikolaas Tinbergen and Konradt Lorenz used a

similar framework as they developed the discipline of ethology in Europe.

Behaviorism in North America

Watson, in an article published in *Psychological Review* in 1913, suggested that 'Psychology, as the behaviorist views it, is a purely objective, experimental branch of natural science which needs introspection as little as do the sciences of chemistry and physics' (Watson, 1913). Watson had a varied scientific career, starting with studies on the neural basis of learning. He was particularly interested in the idea that neurons had to be myelinated in order to support learning. He then spent a year carrying out field studies on sooty and noddy terns using an experimental approach to investigate nest site and egg recognition. This was followed by experimental studies on conditioning using rats, which emphasised the role of simple stimulus-response relationships in behaviour. He resisted any consideration of more complex cognitive processes because, to him, they risked returning to a dualist separation of mind and body.

Towards the end of his short scientific career, he performed the infamous 'Little Albert' experiment in which he conditioned a nine-month-old baby, Albert B., to fear a white rat (Watson & Rayner, 1920). Initially, the baby showed no fear of the rat, but the experimenters then made a loud and unexpected sound (a hammer hitting a steel bar) each time the

baby reached out towards the rat. After several pairings, Albert would cry if the rat were presented, but continued to play with wooden blocks that he was provided with in the same context. He became upset when presented with a rabbit, though to a lesser extent than with the rat. This is an experiment that would almost certainly not be approved under the ethical codes used in psychology today (see 'Ethical Issues in Biological Psychology' later in this chapter).

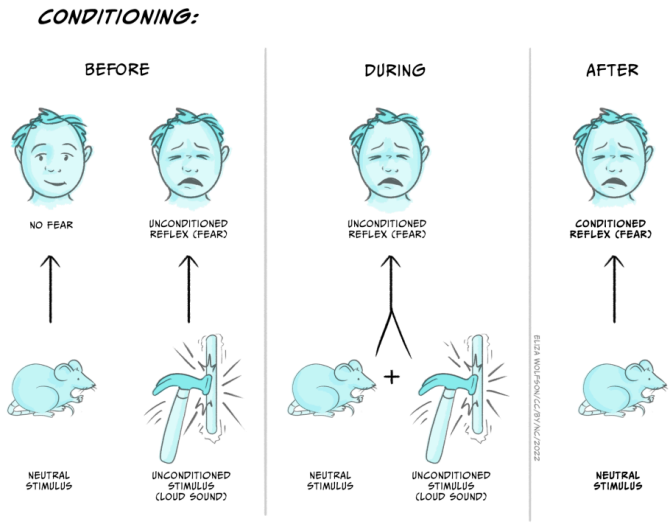

Fig 1.8 The conditioning procedure in the 'Little Albert' experiment is an example of Pavlovian conditioning.

This type of procedure is now often referred to as Pavlovian conditioning, named after the Nobel prize-winning Russian physiologist Ivan Pavlov. His experiments used dogs and

measured salivation in response to the presentation of raw meat. The dogs were conditioned by pairing the sound of a a ticking metronome (not, as often stated, a bell) with the availability of the meat. Subsequently the sound of the metronome alone was enough to elicit salivation. Although Pavlov's name is the one associated with the phenomenon, it was already well known by his time. The French physiologist Magendie described a similar observation in humans as early as 1836.

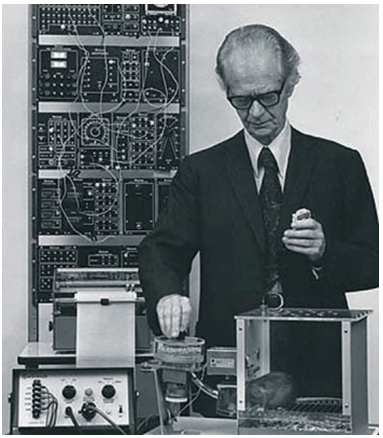

Fig 1.9. Burrhus Skinner, photographed in his Harvard laboratory with a rat, operant chamber and control equipment at the rear.

Skinner set out an even more radical approach in his book *The Behaviour of Organisms* (1938), arguing that cognitive or physiological levels of explanation are unnecessary to understand behaviour. In the Preface added to the 1966 edition, he wrote the following:

> *The simplest contingencies involve at least three terms – stimulus, response, and reinforcer – and at least one other variable (the deprivation associated with the reinforcer) is implied. This is very much more than input and output, and when all relevant variables are thus taken into account, there is no need to*

appeal to an inner apparatus, whether mental, physiological, or conceptual. The contingencies are quite enough to account for attending, remembering, learning, forgetting, generalizing, abstracting, and many other so-called cognitive processes.

Yet this approach had already been criticised. Donald Hebb, in *The Organisation of Behavior*, published in 1949, argued for a close relationship between the study of psychology and physiology. In the opening paragraphs of his book he contrasted his own approach with that of Skinner, saying:

Fig 1.10 Donald Hebb photographed in 1979 by Professor Richard Brown (Dalhousie University). Hebb held diametrically opposed views to Skinner on the importance of physiology to the study of behaviour. Hindsight is a fine thing, but it is clear that Hebb was on the right side of that argument.

> *A vigorous movement has appeared both in psychology and psychiatry to be rid of 'physiologising' that is, to stop using physiological hypotheses. This point of view has been clearly and effectively put by Skinner (1938), ...*
> *The present book is written in profound disagreement with such a program for psychology.* (Hebb, 1949, page xiv of the Introduction).

Today Hebb is best remembered for the suggestion that learning involves information about two separate events, converging on a single nerve cell, and the connections being strengthened in such a way as to support either Pavlovian or

operant conditioning. His hypothesis, which he acknowledged as having its origin in the writings of Tanzi and others some fifty years before, is often remembered by Carla Schatz' mnemonic 'Cells that fire together, wire together'. Today, that phrase is most often associated with the phenomenon of long term potentiation (LTP) which acts as an important model for the neural mechanisms involved in learning and memory.

There were other areas of psychology, especially the study of sensation and perception, where the radical approach of the behaviourists never took a full hold. Helmholtz, whose early work on neural conduction was so important, also made ground breaking contributions to the study of auditory and especially visual perception. He emphasised the importance of unconscious inferences in the way in which visual information is interpreted. All perceptual processing involves a mix of 'bottom-up' factors that derive from the sensory input and 'top-down' factors which involve our memories and experience of similar sensory input in the past. Alternative, more descriptive, terms for 'bottom-up' and 'top-down' are data driven and concept driven. You will learn much more about these processes in modules that explore cognitive psychology.

The development of ethology in Europe

Although the study of animal learning during the mid twentieth century was a dominant paradigm in biological psychology in North America, this was much less true in Europe. Here, an alternative approach evolved.

Fig 1.11 Nikolaas [Niko] Tinbergen (right) at Walney Island, Cumbria, the site of many of his studies of gull behaviour

Konrad Lorenz and Nikolaas Tinbergen emphasised the detailed study of animal behaviour, often in a field rather than a laboratory setting. Both were fascinated by natural history when young. Lorenz was especially taken by the phenomenon of **imprinting.** It was first described in birds, such as chickens and geese, that leave the nest shortly after hatching and follow their parents. In a classic experiment he divided a clutch of newly-hatched greylag geese into two. One group was exposed to their mother, the other to him. After several days the young goslings were mixed together. When he and the mother goose

walked in different directions the group of youngsters divided into two, depending on whom they were originally imprinted. The term imprinting is now often used much more generally for learning that occurs early in life but continues to influence behaviour into adulthood. Lorenz remains a controversial figure because of his association with Nazism during World War II.

Tinbergen began his scientific career in Holland, performing experiments that revealed how insects use landmarks to locate a burrow. They were reminiscent of Watson's earlier studies with terns, though much better designed. During World War II he was held as hostage but survived and subsequently moved to Oxford University in the late 1940s. He is best remembered for an article written in 1963 on the aims of ethology, which will be the basis of the next section of this chapter.

Lorenz and Tinbergen, like Watson and Skinner, were interested in explaining behaviour in its own terms rather than exploring underlying brain or physiological mechanisms, and in this sense they were both adherents of the positivist approach championed by Auguste Compte for all of the social sciences. Tinbergen made this point very clearly in his book *The Study of Instinct* (1951). He acknowledges that the behaviour of many other animals can resemble that of a human experiencing an intense emotion (as Darwin had pointed out in *The Expression of Emotions in Man and Animals* in 1872) but went on to say that 'because subjective phenomena cannot

be observed objectively in animals, it is idle to claim or deny their existence' (Tinbergen, 1951, p.4).

A cognitive perspective in animal learning and ethology

Fig 1.12. Anthony [Tony] Dickinson – and BMW motorbike – outside the Psychological Laboratory at the University of Cambridge, where he worked post Sussex.

By the second half of the twentieth century, psychologists interested in human behaviour were becoming increasingly dissatisfied with the limitations of the behaviorist approach. Psychologists working on animal learning also realised that

some of the phenomena that could be observed in their experiments could only be explained by postulating intervening cognitive mechanisms. Tony Dickinson, in his short text *Contemporary Animal Learning Theory*, published in 1980, used the example of sensory preconditioning. Dickinson was one of the initiators of the so-called 'cognitive revolution' in animal learning.

In a standard Pavlovian procedure, rats are initially trained to press a bar for small pellets of sweetened food. Then they are exposed to an initially neutral stimulus – a light in the wall of their cage – immediately prior to receiving a mild foot shock. After several such pairings the rats are exposed to the light alone – there is no shock. Nevertheless, the rats 'freeze' (remain immobile) and pause bar-pressing for the food reward for a few seconds before returning to their normal behaviour.

The critical modification in sensory preconditioning is to expose the rats to pairings of a light with a second neutral stimulus, in this case a sound, prior to the main conditioning task. Since nothing happened after these initial pairings the rats rapidly come to ignore them. Their ongoing behaviour, in this case bar-pressing for the food pellets, was not changed. Now the rats were conditioned to associate the light and shock in the standard way. Following conditioning they were exposed either to the light, or to the tone. Rats exposed to the tone paused in a similar way to rats that were exposed to the light despite never having experienced that sound preceding a shock. So, despite little or no change in their behaviour during

those initial light-tone pairings, they clearly had learned something. Learning does not have to involve any overt behavioural change.

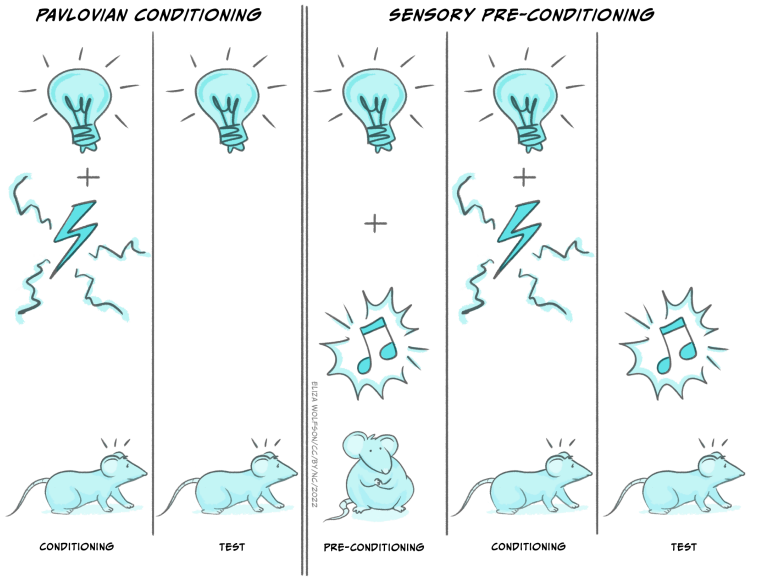

Fig 1.13 In simple Pavlovian conditioning (A) the animal is exposed to several pairings of light and shock and then to the light alone. It responds as though expecting a shock. In sensory pre-conditioning (B) the animal is pre-exposed to a pairing of light and sound then to the light/shock pairings. Despite never have been exposed to a sound/shock pairing, presentation of sound alone elicits a similar response to the light in A.

Dickinson set out the implications in the following way:

> *As we shall see, sensory preconditioning is but one of many examples of behaviourally silent learning, all of which*

> *provide difficulties for any view that equates learning with a change in behaviour. Something must change during learning and I shall argue that this change is best characterised as a modification of some internal structure. Whether or not we shall be able at some time to identify the neurophysiological substrate of these cognitive structures is an open question. It is clear, however, that we cannot do so at present.* (Dickinson, 1980, p.5)

That was 1980!

As I write this, in the 2020s, it is possible to give an optimistic answer to Dickinson's final query. Modern techniques from neuroscience can identify the brain structures and changes in small groups of neurons that support learning and memory. In one recent example, Eisuke Koya and his colleagues, who work in the School of Psychology at the University of Sussex, describe a simple learning task in which mice are exposed to a short series of clicks (the 'to-be-conditioned' stimulus) followed by an opportunity to drink a small quantity of a sucrose solution. After being exposed to such pairings over a period of a few days, they learn to approach the sucrose delivery port as soon as the clicks begin. The mice used in these experiments were genetically modified in such a way that nerve cells in the frontal cortex that were activated during the conditioning trials could be made to glow with a green fluorescence. The research established that small, stable groups of nerve cells ('neuronal ensembles') were activated from one day to the next. The researchers were also able to manipulate these cells in such as way as to produce an

abnormal change in their activity during subsequent tests in which the animals were exposed to the conditioned stimulus. Approach to the sucrose delivery port was disrupted when this was done, suggesting that the activation of these cells contributed in an important way to the learnt behaviour of the mouse (Brebner et al., 2020). Experiments of this kind are approaching the goal of identifying the neural structures and mechanisms that support learning and memory, and demonstrate how psychologists and neuroscientists can collaborate to tackle the fundamental problems of biological psychology.

Field studies of non-human primates

Researchers in the field of animal learning were not the only ones who found it difficult to explain their experimental results without invoking cognitive processes that could not be directly observed. Ethologists faced a similar challenge.

Jane Goodall began her fieldwork on chimpanzees in the early 1960s and, as she got to know and was accepted by the troop that she was studying at Gombe Stream in Tanzania, gathered evidence about their rich social and emotional lives and the use of tools that revolutionised studies of primate behaviour. To the consternation of her colleagues at Cambridge, she also gave names to the chimpanzees rather than the numbers which would supposedly lead to more objective study.

Fig 1.14. Alison Jolly (shown with ring-tailed lemur) who, with Jane Goodall and others, was among the first to recognise the way in which the complex social lives of primates might drive the evolution of cognition. She worked at Yale and then Sussex.

At about the same time, **Alison Jolly** was beginning her studies of lemur behaviour in Madagascar. She argued that the major driving force in evolution of primate cognition came from the complex demands that come from living in complex and long lasting social groups (Jolly, 1966).

Field studies carried out since then have demonstrated long-lasting social relationships in primate species such as baboons and vervet monkeys. Calls that the animals make in the context of aggressive encounters are interpreted in terms of an animal's prior knowledge of social hierarchies within their group and their prior behaviour. For example, female chacma baboons, studied by Dorothy Cheney and Robert Seyfarth, have a call, the 'reconciliatory grunt', that is given just after an aggressive encounter to indicate a peaceful conclusion between the two individuals. When the call was played to a female who had just been involved in a mutual

grooming session with another female, she behaved in a way that implied that the call must be directed at someone else and not her. However, if she had been involved in an aggressive encounter with that same female a little earlier, she behaved in a way that implied that the grunt had been directed at her (Engh et al., 2006).

Field studies of other long lived mammals, including elephants and dolphins, suggest that they also have complex social networks and sophisticated cognitive abilities.

Two features stand out at the end of this very short historical survey of ideas about the relationship between the brain and behaviour. The first is that, among contemporary psychologists and neuroscientists, there is an almost universal acceptance of some form of materialism. In other words all of the complexity in our behaviour, including such relatively less well understood areas such as consciousness, are a consequence of physical mechanisms operating in our bodies, and primarily in the nervous system. The second is that, despite a hiatus that lasted for at least the first half of the twentieth century, the study of phenomena such as emotion and consciousness are no longer seen as 'off limits' for scientific study. One challenge for modern neuroscience is to understand how the nervous system builds, uses and attaches emotional weight to internal representations of aspects of the external world.

What kinds of questions does Biological Psychology ask?

Since biological psychology is concerned with both behaviour and relevant physiological and brain mechanisms, it will often start with some interesting behavioural observations or experimental data. Once the behaviour of interest has been adequately documented, it is time to ask some questions. The ethologist Niko Tinbergen suggested that were four broad kinds of questions that might be of interest (Tinbergen 1963). The first two have a timescale within the life cycle of an individual animal and are concerned with:

(i) the underlying causes of changes in behaviour, such as brain mechanisms or hormonal changes, and

(ii) the development of behaviour, for example as an individual matures to adulthood.

These are sometimes referred to as the **proximate causes** of the behaviour. The remaining two questions are set in a much broader time frame and can be thought of as the **ultimate causes** (Bateson & Laland, 2013). They are concerned with:

(iii) the evolutionary relationships between patterns of behaviour in different species, and

(iv) the advantages of particular patterns of behaviour in the context of natural selection.

A couple of examples should illustrate how these questions differ from one another and yet still address the question of why a particular behaviour occurs in the form that it does.

Example 1: Bird song

The chaffinch (*Fringilla coelebs*) is a common European songbird. In the early Spring, adult males have a striking breeding plumage and their bills darken. At the same time the birds begin to perch in conspicuous places within a small territory that they defend from other males and sing. The females do not sing, although they use a variety of other calls. So, why do male chaffinches sing?

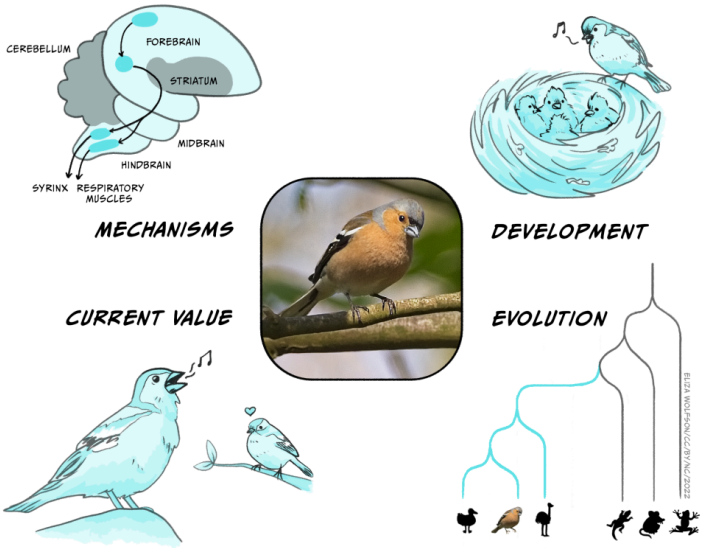

Fig 1.15. Why does a chaffinch sing in the Spring? Proximate and ultimate explanations.

Causes and mechanisms. The changes in bill colour and

onset of singing are associated with the increase in day length in the Spring. Experimental studies in other songbird species, such as those by Fenando Nottebohm in canaries, have shown that there is an increase in the secretion of testosterone at this time, and that experimental administration of this steroid hormone produces the same changes in appearance and behaviour (Nottebohm, 2002). Further studies have demonstrated that there are changes in the bird's brain at this time.

The most surprising result was a clear demonstration that new neurons are formed through cell division in the areas of the brain involved in song. In the mid 1980s, when these studies were performed, the consensus was that neurons were never added to a mature vertebrate brain. Nottebohm's work forced a re-examination of this idea. The methods that his lab used were repeated in other species, and demonstrated that the same thing could also happen in mammals, including humans. In the years since these ground-breaking studies, a good deal more has been learned about circuits within the songbird brain that support song production.

Development. You do not need to be an experienced birdwatcher to recognise the song of a chaffinch. It consists of several short trills followed a characteristic terminal flourish. This chaffinch sonagram shows the loudness (upper panel) and frequency changes in the song of a chaffinch. Alternatively, play the song as a short video.

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://openpress.sussex.ac.uk/introductiontobiologicalpsychology/?p=532#oembed-1

However, although it is difficult to mistake the song of a chaffinch for that of a different species, the song has unexpected complexity. An individual may sing several subtly different song types and there are also clear differences in the song over the different parts of their widespread distribution through Europe. This raises several questions about the way in which a young chaffinch develops its song repertoire. If a nestling is reared in an acoustically isolated environment, it develops a highly abnormal song lacking the detailed structure typical of a chaffinch. However, if a bird reared in the same way is exposed to tape-recorded song, then it accurately copies the song types and incorporates them into its repertoire. Of course there will be be a considerable gap between hearing the song as a summer fledgling, and then singing the song in the following spring, so this is another example of behaviourally silent learning. In the wild, things turn out to be be much more complex. A chaffinch may acquire some song types in the first summer, but additional ones may also be learnt from neighbours in the following spring as they set up territories. It

is clear that one early, popular idea about the development of song types is incorrect: they are not exclusively learnt from a bird's father (Riebel et al., 2015).

Evolutionary relationships. Birds, like mammals, reptiles and amphibians, are vertebrates. Although there is tremendous variety in their appearance and behaviour, there are common features, such as the presence of a backbone. There are also strong similarities in broad aspects of brain structure and functioning. Songbirds are one of several groups of living birds and there are some 5000 different species. They all have a well-developed syrinx (the rough equivalent of the human larynx) which has a complex musculature that allows the bird to sing. There is tremendous variation from one species to another, from the complex, extended song of thrushes such as the blackbird and nightingale, to the much simpler song of the chaffinch. Songbirds evolved about 45 million years ago, and birds diverged from the vertebrate line that also gave rise to mammals some 320 million years ago. Our common ancestor probably resembled a small lizard. Present day lizards can show some degree of behavioural flexibility and social learning, so there are interesting questions about the extent to which some of the more advanced cognitive abilities of birds and mammals evolved independently or build on components already present in that common ancestor.

Function or current utility. Song is energetically expensive at a time of the year when food is not at its most abundant. It can also be dangerous. There is a risk that a

sparrowhawk will appear over the top of the hedge on which the chaffinch is singing and provide the hawk with its next meal! It follows that song must also potentially enhance the biological fitness of the bird in some way. There are at least two factors at play here. A male has to attract a female to nest in his territory and mate with her. His song also advertises to other males that he holds, or is attempting to hold, that space, and may be a prelude for fighting over ownership of the best areas. There is also experimental evidence that characteristics of the song, particularly the complexity of the final trill, may be attractive to females and lead them to prefer one male over another. Although the evidence is not completely convincing for chaffinches, this idea would fit in with findings from a variety of other species. The croak of a frog, the roar of a red deer stag or the colours of a peacock's tail may act as signals about the quality of the individual making the call, and may have the advantage of being hard to falsify – so called 'honest signals'. However the precise way in which such signals evolve remains unclear (Penn & Számadó, 2020; Smith, 1991).

Example 2: Anxiety and fear

Tinbergen's general approach can be productive in thinking about any aspect of behaviour. In humans anxiety or fear is an unpleasant emotional experience that may come in many forms including panic and phobias of various kinds. The emotion of fear is often evoked by quite specific threat-related

stimuli – perhaps a snake (snake phobia) or wide open spaces (agoraphobia). In the same way as for bird song, we can ask the question 'why?' and break it down into queries about either proximate or ultimate causes.

Causes and mechanisms. The underlying physiological and brain mechanisms are well studied. They include increases in heart rate, release of hormones such as adrenalin, and activation of a specific brain network that includes the amygdala. One part of the amygdala, the central nucleus, is responsible for activating these different physiological changes in a coordinated manner (LeDoux, 2012). An understanding of these types of mechanisms has clinical relevance. Drugs that act selectively on these threat-processing circuits may have value as treatments for anxiety. Indeed benzodiazepines such as valium are still widely used in this way and are known to have especially potent effects in the amygdala. An important, but still unanswered, question is how these physiological responses relate to the conscious feeling of fear.

Development. We also know a good deal about the way in which fear may develop during an individual's lifespan. Simple conditioning may often play a role and there is also good evidence that species as varied as rodents and primates are more likely to become fearful and avoid some types of object rather than others. In social species observational learning may also be important. Studies of young rhesus monkeys illustrate these points. A rhesus infant will initially show little avoidance or fear of model snakes or flowers. However if they are allowed

to watch an edited video in which an adult rhesus appears to respond fearfully to either the flowers or a snake, they themselves develop fear responses to the snake but not to the flower (Cook & Mineka, 1990). This suggests an innate tendency to become fearful of some kinds of object, such as a snake, that can potentiate the effects of observational learning. In a similar way many rodent species will avoid odours associated with potential predators such as a fox or cat without having any previous experience of those animals. However, especially when young, those responses may be amplified if they observe an adult responding strongly to the same stimulus.

Evolutionary relationships. Comparative studies of the specific behaviour patterns associated with fear and the underlying physiological and brain mechanisms suggest that they have been conserved through vertebrate evolution. Charles Darwin, in the Expression of the Emotions in Man and Animals, provided some of the first really detailed behavioural descriptions of facial expressions associated with fear and especially emphasised their role in communication. Detailed comparisons of the neural circuitry in mammals, birds, amphibians and reptiles suggest that the amygdala, and especially its connections to the autonomic nervous system which activate the hormonal and other physiological responses to fear-evoking stimuli, are conserved through the entire vertebrate evolutionary line and must therefore have originated at least 400 million years ago. So it is not surprising

that one of the responses of a Fijian ground frog to the presence of a potential predator (a cane toad) is an increase in the stress hormone corticosterone as well as a behavioural response, in this case immobility, that reduces the likelihood of being eaten (Narayan et al, 2013). Exposure to a stressful situation in humans produces the same hormonal response, although it is cortisol, which is almost structurally identical to corticosterone, that is released.

Function or current utility. Questions of function, or current utility, can be thought about at multiple levels. It is clear that fear, or the perception of threat-related stimuli can be a powerful driver of learning. As we saw earlier in this chapter, previously neutral stimuli that predict threat or danger come to evoke the same responses as the threat itself (Pavlovian conditioning). In the natural environment such responses are likely to enhance biological fitness. However in addition to thinking about the likely function of fear systems in a rather global manner, it is also possible to analyse the individual behavioural elements that make up a a fear response. One such element, described by Darwin (Darwin, Charles, 1872) and also recognised in the later studies of Paul Ekman, is that the eyebrows are raised which results in the sclera (white) of the eye becoming much more obvious (Jack et al., 2014 includes a video example). The original function of this response may simply have been to widen the field of view but, especially in primates, it can now also serve as a way of communicating fear within a social group. It is likely that

behavioural responses frequently gain additional functions during evolution, perhaps even making the original function irrelevant. This is the reason that the term 'current utility' is often preferred to function (Bateson & Laland, 2013). If a functional hypothesis is be tested experimentally it will always be current utility that will be assessed. When a particular characteristic or feature acquires additional functions in this way they are sometimes described as exaptations rather than adaptations.

Scientific strategies in Biological Psychology

The first phase of any scientific investigation is likely to be descriptive. In biological psychology, this is a point at which the influence of an ethological approach is most obvious. It is easiest to describe how this phase proceeds by using some specific examples. Once the behaviour of interest has been clearly characterised it is often time to collect some empirical data. This will often involve either collecting behavioural and physiological data and correlating them, or taking a more experimental approach in which environmental or physiological factors are deliberately manipulated. A combination of these approaches will begin to elucidate the way in which neural processes influence behaviour and, in turn, are influenced by the consequences of that behaviour.

Describing behaviour: facial expressions and individual variation during conditioning

Many mammals, including rodents and primates, (including humans in the latter category), make distinctive facial expressions as they try out potential food sources. Humans will lick their lips as they eat something sweet and gooey. A food or drink that is unexpectedly sour (like pure lemon juice) or bitter (perhaps mature leaves of kale or some other member of the cabbage family) might elicit a **gaping** response in which the mouth opens wide and, in more extreme cases, saliva may drip out of the mouth.

These kinds of response can be observed in quite young babies. Indeed, as any parent is likely to know, they are very common as an infant transitions from breast feeding to solid foods. It may seem surprising but very similar responses can be observed in rats or mice as they drink sweet, sour or bitter solutions. It demonstrates that these are responses that are likely to have been conserved over relatively long periods of evolutionary time. They may serve a dual function. A response like gaping will help to remove something that may be toxic from the mouth – bitterness is often a signal that a plant contains harmful toxins. However it is also likely that, at least in some species, the 'current utility' of these expressions also includes a communicative function in species that feed in social groups. This would be another example of an

exaptation (i.e. an additional function that becomes adaptive later).

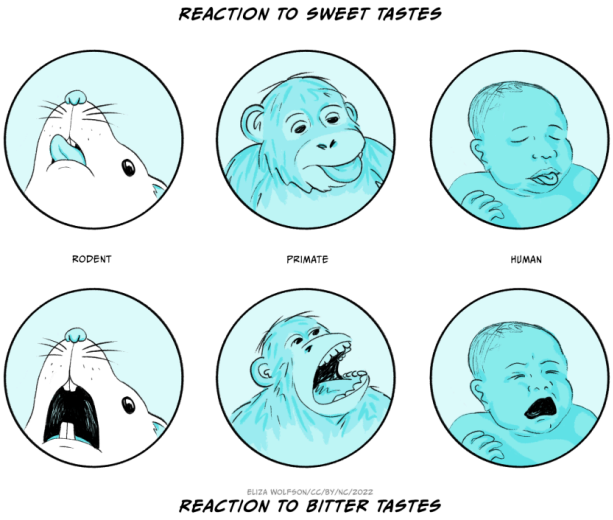

Fig 1.16. Similar facial expressions to 'nice' and 'nasty' tastes in rodents and primates.

Detailed measurement of these facial responses forms the basis of the so-called 'taste reactivity task' in which controlled amounts of solutions with different taste properties are infused into the mouth of a rat or mouse and the facial expressions quantified (Berridge, 2019). The task was initially devised to investigate the role of different brain structures in taste processing. A little later some detailed studies with a variety of different solutions revealed that the extent to which they evoked **ingestive** ('nice') and **aversive** ('nasty') responses

could, at least to some extent, vary independently. It then became clear that there were drug and brain manipulations that had no effect on the facial expressions evoked by a liked, or rewarding, sweet solution. However those same manipulations did reduce the extent to which an animal would be prepared to work (e.g. press a lever, perhaps several times) in order to gain access to that solution. The important implication was that the extent to which something is 'liked', measured by facial expressions, may depend on different factors to those that affect whether it is 'wanted', measured by effort to obtain that reward. Although this distinction first arose in the investigation of what might seem an obscure corner of biological psychology, it has, as you will read in Chapter 15, Addiction, become an important theoretical idea that helps explain some otherwise puzzling features of addiction to drugs, food and other rewarding stimuli.

One feature of animal behaviour, humans included, that seem irritating at first is that there is often substantial variation between individuals exposed to the same experimental manipulation. It can be tempting to ignore it, or to choose measures that at least minimise it. But this can be a mistake, as this example from Pavlovian conditioning demonstrates.

As a group of rats learns the association between the illumination of light and the delivery of a food pellet they retrieve the pellet more rapidly and the averaged behaviour of the group generates a smooth 'learning' curve. However, careful observation of individuals reveals several things. First,

the changes in individual animals are much more discontinuous – almost as though at some point, individual rats 'get' the task, but at different times during the training process! The behaviour of individuals also varies in other, potentially interesting, ways. Some rats approach the light when it illuminates, rearing up to investigate it, and continue to do so even when they subsequently approach the place where the pellet is delivered. Other rats move immediately to the location where the food pellet will be delivered, apparently taking no further interest in the light. The behaviour of the former group is referred to as 'sign tracking' and the latter as 'goal tracking'. It turns out that sign trackers and goal trackers differ in other interesting ways. For example, sign-tracking rats show greater impulsivity in other tasks and acquire alcohol self administration more readily. The same kinds of distinction may also show up in human behaviour and predict vulnerability to drug addiction and relapse.

Although the descriptive phase is likely to begin any serious scientific investigation, once interesting observations have been made they are likely to raise the kinds of questions that were discussed in the last section. What are the brain and physiological mechanisms that underlie the behaviour? How do the behaviour patterns develop through lifespan? and so forth. In working out how to tackle such questions there are a number of potential ways forward. They typically use a combination of two general strategies.

Investigational strategies: correlational approaches, and experimental manipulations

One strategy is to observe changes in behaviour, typically in a carefully controlled test situation or using a clearly defined set of behaviour patterns when doing fieldwork, while measuring changes in physiological and brain function that are likely to be relevant. Then, using appropriate statistical techniques, the changes in behaviour and physiology can be correlated together. The second strategy is to deliberately manipulate a test situation in order to determine the extent to which an imposed change in in physiology or brain function leads to a change in behaviour, or a change in behaviours leads to a physiological change.

The issue with the first correlational strategy is that, although the results may suggest that there is some type of causal relationship between behaviour and physiology, they don't clarify the nature of the relationship. Behaviour A may cause changes in physiological or brain variable B. But equally, the physiology may influence the behaviour. Finally, it may be that there is no direct mechanistic linkage between behaviour and physiology. Instead, some third variable is independently affecting both. A further complication is that there may be important feedback loops that influence the outcome. So, although correlational studies can be very useful, they do have limitations when it comes to their interpretation. The second

strategy, involving deliberate experimental manipulation, has a better chance of determining the direction of causation. But it may raise other problems. Deliberately interfering with the functioning of a complex biological system may lead it to respond in unpredictable ways and may also be ethically problematic.

Fig 1.17. Male red deer during the Autumn rut

Let's see how this works in practice by considering the behaviour of red deer during the autumn rutting season. At this time, a successful male deer may establish a 'harem' of female deer and defend them against other males. His behaviour makes it more likely that only he will have the opportunity to mate with them, and hence that the resulting offspring would enhance the representation of his genes in the next generation. Females also gain from being in the harem by gaining some protection from harassment by other males, which provides the opportunity to feed in a more uninterrupted way. They may choose the harem of a particular male on the basis of his perceived fitness.

It would first be possible to correlate the gradual increase in aggressive behaviour during the red deer rut with the increasing blood testosterone that occurs during the autumn. It might be tempting to conclude that the increased

testosterone level *causes* the increased level of aggressive behaviour directed at other males. However there are at least two, and perhaps more, possibilities that would need to be excluded first.

First, it is possible that the changes in behaviour and increased testosterone level are triggered independently by some third factor. One alternative would be that decreasing day length acts as that trigger. The day length signal might be detected within the pineal gland (Descartes' supposed valve from the soul to the body!) and independently trigger both the hormonal and behavioural changes.

A second possibility is that the behavioural changes actually trigger the hormonal change. This isn't a completely implausible suggestion. Such effects have been documented in a number of mammals, including human (male) tennis players, where testosterone levels increase after a match that they have just won. In the same way, testosterone secretion might be sensitive to whether aggressive encounters between the male deer are won or lost.

An experimental approach can be used to overcome the difficulty in deciding what causes what. In the case of the role of testosterone in aggression, one possible strategy is to remove the source of testosterone and determine whether aggressive behaviour continues. In fact this has been common practice for millennia in managing farmed animals. An intact bull may be very aggressive but castrated male cattle (bullocks) are typically much less so. In experimental animals the specific role

of testosterone could be demonstrated by administering the hormone to a castrated animal and showing that aggressive behaviour returns to the expected level. Experiments of exactly this kind have been performed on red deer, and indicate that testosterone does indeed restore rutting behaviour when administered to a castrated male in the autumn. However the effect of the hormone on rutting behaviour is absent when the same treatment is given in the spring, although there is some increase in aggressive behaviour (Lincoln et al., 1972). So, factors like day length and hormone level interact in a more complex way than might be expected.

A similar experimental approach can be taken in relation to the contribution that particular brain structures or identified groups of neurons may make to a specific behaviour pattern. Suppose that we have already discovered, perhaps by making recordings of neural activity, that neurons in a particular structure (let's call it area 'X') become more active while an animal is feeding. How can we demonstrate that those neurons are important in actually generating the behaviour rather than, for example, just responding to consequences of eating food? In other words, how can we determine that activation of area X *causes* feeding as opposed to feeding *causing* activation of area X? If stimulation of the nerve cells within area X leads to the animal, when not hungry, beginning to feed on a highly palatable food, then you might assume that demonstrates their critical role – authors reporting such an experiment will often state that the cells in this area are 'sufficient' to generate

feeding. However this doesn't show that those same cells are always active in the many other experimental situations in which that animal might begin to feed. In the same way, suppose an animal fails to eat when the cells in that same area X are inactivated. That finding doesn't demonstrate that these cells are *always* 'necessary' for feeding to occur. For example, suppose the original test situation were eating a palatable food when already sated, the animal might still eat when those same cells were inactivated after they had been deprived of food for a few hours (Yoshihara & Yoshihara 2018). It is also possible that inactivation of those cells might interfere with other types of behaviour, suggesting that they had no unique importance in relation to feeding. This highlights the moral that demonstrating causation, and especially the direction of causation, is rarely easy!

In studies where only correlations have been measured it is tempting for those presenting the research to slip from initially saying that an association has been observed to then discussing the results in terms of a causal mechanism that hasn't been fully demonstrated. This is something to watch for when critically assessing research publications. It often happens in the discussion of results based on fieldwork when an experimental approach may be much more challenging. A combination of correlational and experimental approaches can also be taken in studying questions about the development of behaviour.

This combination of approaches can be taken in the study

of human behaviour and its relationship to particular brain or other physiological changes. However, the experimental approach can be limited by the more significant ethical concerns that may arise. Clinical data has been used since the time of Galen, Ibn Sina, Broca and others to correlate the brain damage that occurs after strokes, or other forms of brain damage, with changes in behaviour. Classic studies include the patient Tan, studied by Broca, where loss of the ability to speak was linked to damage in the left temporal cortex; and Phineas Gage, whose damage to the prefrontal cortex was associated with more widespread changes in behaviour. This last example also provides a cautionary tale. There is an unresolved dispute as to how substantial and permanent Gage's changes in behaviour actually were, despite the typical clarity of textbook presentations (Macmillan, 2000).

Ethical issues in Biological Psychology

One consequence of our growing appreciation of the potentially rich cognitive and emotional lives of animal species other than humans has been a concern about the ethical position of their use in experimental or observational studies. Psychology as a discipline has also become much more concerned to treat human participants in an ethical manner, ruling out work such as the Little Albert study that we discussed earlier.

What does it mean, to behave a morally or ethically good way? Philosophers have debated this subject since the dawn of recorded human history. One rational approach which synthesises some of the different possible approaches discussed by the psychologist Stephen Pinker is to 'only do to others what you would be happy to have done to you' (Pinker, 2021, p. 66 *et seq.*). It is rational in the sense that it combines personal self-interest in a social environment but also survives the change in perspective that comes with being the giver or receiver of a particular action. It is also the basis of the moral codes promoted by many of the world's religions. This type of philosophical approach is consistent with the Code of Human Research Ethics put forward by the British Psychology Society (*BPS Code of Human Research Ethics – The British Psychological Society*, 2021), which is based on four fundamental principles:

1. Respect for the autonomy and dignity of persons
2. Scientific value
3. Social responsibility
4. Maximising benefit and minimising harm

The code goes on to explain how these principles underpin the ways in which experiments are designed, participants treated, and results disseminated. The last three principles also apply in a relatively straightforward way to research that involves non-human animal species. Applying the first principle is more

complex and partly dependent on the moral status that humans give to non-humans.

Philosophers take a variety of positions on morality that are often contradictory. However, two important approaches are **utilitarianism**, which developed from the writings of the English philosopher Jeremy Bentham in the nineteenth century, and a **rights-based** approach. Bentham, using concepts developed by the French philosopher Helvétius a century earlier, suggested that good actions are ones which maximise happiness and minimise distress: the 'greatest felicity' principle. While he recognised that non-human animals might experience pain and distress, he also regarded that as of lesser importance than the wellbeing of humans.

In the twentieth century writers such as Peter Singer and Richard Ryder rejected this 'human-centred' approach, terming it 'speciesism', by analogy with racism. But nevertheless they accepted, with this modification, the utilitarian approach of Bentham. Within this framework some use of non-human animals may be ethically justified, though every effort must be made to enhance benefits and especially to reduce any negative impact on the lives of the animals used during the course of the studies.

In contrast, Tom Regan has rejected the view that benefits and dis-benefits could be added together using some form of utilitarian arithmetic to decide whether an action was acceptable or not. Instead, he asserted that at least some non-human animals have the same rights as a human to life and

freedom from distress. They are, to use his phrase, 'subjects of a life'. Regan discusses a number of attributes that might help in deciding whether a particular group of animals meet this criterion. Regan's approach would rule out the use of many species of animal in science, agriculture and many other contexts

The first UK law designed to protect non-human animals used in scientific research was the Cruelty to Animals Act 1876. It was replaced by the Animals (Scientific Procedures) Act 1986 in line with the EU Directive 86/609/EEC (now replaced by Directive 2010/63/EU). The legislation takes a broadly utilitarian approach to judging whether a proposed set of experiment is, or is not, ethical. In other words, there is an attempt to weigh up the potential benefits of the work in terms of advancing scientific knowledge, or the clinical treatment of human and animal disease, which is set against the dis-benefits in terms of impact on the welfare of the animals that will be used in that research. However, it incorporates elements of a more rights-based approach, in that experiments involving the use of old world apes (chimpanzees, gorillas etc) are prohibited. At present, with the exception of cephalopods, the current legislation ignores invertebrates. However there is increasing evidence sentience may be more widely distributed than previously appreciated, especially amongst decapods (crabs and lobsters), so it would not be surprising if the legal definition of a protected species was widened in the future.

At a practical level William Russell and Rex Birch suggested

(Russell & Burch 1959) that there are three important considerations when designing an experiment that might, potentially, use non-humans:

1. *Replacement* – could the use of non-human animals be replaced either by the ethical use of human participants or by the use of a non-animal technique – perhaps based on cultured cells?
2. *Reduction* – could the use of animals be minimised in such a way that the results would remain statistically valid?
3. *Refinement* – could the experimental programme be modified in a way that eliminated, or at least minimised, any pain or distress, perhaps by using positive rewards (e.g. palatable food) rather than punishment (e.g. electric shock) to motivate performance in a particular test situation?

In the UK, these practical ideas have to be addressed before experimental work is approved in the context of a broader cost-benefit analysis. The proposal has then to be considered by an ethics committee which includes lay members before it receives approval from the relevant government department.

Key Takeaways

- There was an increasing realisation from the period 200 BCE to 1700 CE that the brain was critical for overt behaviour, cognition and emotion. The contributions of Galen, Ibn Sina and Descartes are of particular note.
- During the 17th – 19th centuries it became clear that there was a considerable degree of localisation of function within the nervous system. Older fluid-based notions of information flow within the nervous system were gradually replaced within an understanding that it depended on electrical and chemical phenomena.
- During the first 70 years of the 20th century the mechanisms underlying nerve impulse conduction and synaptic transmission were clarified. However, the strong influence of the positivist movement within psychology and ethology devalued the investigation of cognition and emotion.

- In the latter part of the 20th and early part of the 21st century, there has been increasing integration of neuroscientific techniques into behavioural studies. During this period it was also accepted that aspects of cognition and emotion that cannot be directly observed are proper subjects for study in biological psychology.
- The investigation of any aspect of behaviour may involve the study of (i) causation and mechanisms, (ii) development, (iii) evolution and (iv) function or current utility. These are Tinbergen's 'four questions'.
- In the study of causation and development a mix of correlational and experimental strategies can be used. They each have their own advantages and pitfalls.
- Ethical considerations are important considerations in any study of biological psychology. They are based on the strong principle-based ethical code shared by all psychologists, and the specific principles of refinement, reduction and replacement.

Postscript: three basic concepts from biology

Cells

The first descriptions of cells were made in the middle of the seventeenth century by Antonie van Leewenhoek and Robert Hooke. Only in the middle of the nineteenth century was it accepted that all living organisms are made up from one or more cells as their basic organisational unit. In 1855 Rudolf Virchow proposed that all cells arise from pre-existing cells by cell division. Each cell contains the different types of **organelles** that allows that cell to function. Critical amongst these, at least in all multicellular organisms, is the **nucleus**, which contains the chromosomes made up of DNA which, in its detailed chemical structure, encodes all the information that the cell needs to reproduce. The cell also has **mitochondria** that provide the cell with energy and a membrane that bounds it, as well as many other types of organelle.

Cells are grouped into tissues, composed of one or more different cell types, and different tissues are combined together to form organs, such as heart, lungs and brain, formed of many different types of tissue.

Inheritance

When a cell divides, the daughter cell needs the necessary information to duplicate the functions of its parent. The plant breeding experiments of Gregor Mendel in the late 1800s, when combined with the detailed studies of Thomas Hunt Morgan on the ways in which chromosomes replicate during cell division in the early twentieth century, suggested that DNA, which made up much of their substance, must be the molecule that had that function. In the 1950s James Watson, Francis Crick, Rosalind Franklin and others showed that the detailed ordering of one of four nucleotide bases in the double helical chain that makes up DNA provided the code which could be translated into the ordering of amino acids in proteins. Proteins are fundamental to cell function – they function as enzymes, allowing simple chemical transformations to build a cell's structure, generate energy and provide mechanisms that allow communication from one cell to another within complex organisms made of potentially billions of cells. Changes in a specific base within the DNA sequence are one form of mutation (so-called **single nucleotide polymorphisms** – SNPs) that can lead to a change in the structure and functioning of a protein. Larger scale mutations may involve the loss or duplication of parts of a chromosome and are often associated with more substantial changes in body structure, functioning or behaviour.

Evolution by natural selection

The idea that one type (or species) of animal can give rise to another as result of gradual change from one generation to another recurs in the writings of Islamic, Chinese and Greek philosophers, although some Greek writers – such as Plato and his pupil Aristotle – held the opposite view, and held that the form of individual species was fixed and unchanging. But it was Charles Darwin who gave the first clear account of the role of natural selection in evolutionary change. His theory proposed three essential postulates:

- that individuals differ from one to another;
- that this variation may be inherited;
- and that some individuals, as results of this variation, leave greater numbers of offspring.

The consequence is that the characteristics of more successful individuals will become more common in the population. In this way a population, perhaps subject to different environmental pressures over the areas in which it is found, may gradually split into two, and evolve mechanisms that make cross-breeding less likely so as to preserve those inherited differences which adapt them to their differing environments.

Darwin's ideas were very controversial and opposed by many public figures and religious leaders as well as by other scientists. Ironically, Virchow – who had been on the right side

of the argument in relation to the way in which new cells arise by division – was one of Darwin's most vociferous opponents in Germany. He regarded Darwin's ideas as an attack on the moral basis of society. Outstanding scientists can be right in one area, but hopelessly wrong in others!

Plan for the remainder of this textbook

The following chapters of this text cover the broad topic of biological psychology. They begin with an account of the structure of the brain and nervous system, and the functioning of the cellular units that make it up. The discussion then moves to the way in which the nervous system processes and interprets diverse types of sensory information, and the motor mechanisms that generate an observable behavioural output. In addition to a focus on the understanding of 'normal' behaviour, you will read about the extent to which clinical conditions, such as anxiety and depression, may be accompanied by specific changes in brain function and the way in which clinically useful drugs may affect both the function of individual neurones and larger scale brain circuits. Future editions of this book will also include chapters that discuss topics such as motivation, emotion and the neural mechanisms that underpin learning and memory.

References and further reading

Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: An appreciation and an update. *Trends in Ecology & Evolution*, *28*(12), 712–718. https://doi.org/10.1016/j.tree.2013.09.013

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012. https://doi.org/10.1126/science.aay5012

Berridge, K. C. (2019). Affective valence in the brain: Modules or modes? *Nature Reviews Neuroscience*, *20*(4), 225–234. https://doi.org/10.1038/s41583-019-0122-8

Brebner, L. S., Ziminski, J. J., Margetts-Smith, G., Sieburg, M. C., Reeve, H. M., Nowotny, T., Hirrlinger, J., Heintz, T. G., Lagnado, L., Kato, S., Kobayashi, K., Ramsey, L. A., Hall, C. N., Crombag, H. S., & Koya, E. (2020). The emergence of a stable neuronal ensemble from a wider pool of activated neurons in the dorsal medial prefrontal cortex during appetitive learning in mice. *Journal of Neuroscience*, *40*(2), 395–410. https://doi.org/10.1523/JNEUROSCI.1496-19.2019

Cook, M., & Mineka, S. (1990). Selective associations in the observational conditioning of fear in rhesus monkeys. *Journal of Experimental Psychology. Animal Behavior Processes*, *16*(4), 372–389.

Darwin, Charles. (1872). *The expression of the emotions in animals and Man*.

Descartes, R. (1998). *Descartes: The world and other writings* (S. Gaukroger, Ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511605727

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge University Press.

Engh, A. L., Hoffmeier, R. R., Cheney, D. L., & Seyfarth, R. M. (2006). Who, me? Can baboons infer the target of vocalizations? *Animal Behaviour*, *71*(2), 381–387. https://doi.org/10.1016/j.anbehav.2005.05.009

Garfinkel, S. N., & Critchley, H. D. (2016). Threat and the body: How the heart supports fear processing. *Trends in Cognitive Sciences*, *20*(1), 34–46. https://doi.org/10.1016/j.tics.2015.10.005

Goodall, J. (2017, January 20). Remembering my mentor: Robert Hinde. *Jane Goodall's Good for All News*. https://news.janegoodall.org/2017/01/20/remembering-my-mentor-robert-hinde/

Gould, S. J. (2006). *The mismeasure of Man*. W. W. Norton & Company.

Hebb, D. O. (1949). *The organization of behavior*. Chapman & Hall.

Heydari, M., Hashem Hashempur, M., & Zargaran, A. (2013). Medicinal aspects of opium as described in Avicenna's Canon of Medicine. *Acta Medico-Historica Adriatica: AMHA*, *11*(1), 101–112.

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

James, William. (1884). II.—What is an emotion? *Mind*, *os-IX*(34), 188–205.

Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science*, *153*(3735), 501–506. https://doi.org/10.1126/science.153.3735.501

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, *73*(4), 653–676. https://doi.org/10.1016/j.neuron.2012.02.004

Lincoln, G. A., Guinness, F., & Short, R. V. (1972). The way in which testosterone controls the social and sexual behavior of the red deer stag (Cervus elaphus). *Hormones and Behavior*, *3*(4), 375–396. https://doi.org/10.1016/0018-506X(72)90027-X

Macmillan, M. (2000). Restoring Phineas Gage: A 150th retrospective. *Journal of the History of the Neurosciences*, *9*(1), 46–66. https://doi.org/10.1076/0964-704X(200004)9:1;1-2;FT046

Magee, B. (1987). *The great philosophers: An introduction to Western philosophy*. Oxford University Press.

Marsh, H. (2014). *Do No Harm: Stories of life, death and brain surgery*. Hachette UK.

Narayan, E. J., Cockrem, J. F., & Hero, J.-M. (2013). Sight of a predator induces a corticosterone stress response and generates fear in an amphibian. *PLOS ONE*, *8*(8), e73564. https://doi.org/10.1371/journal.pone.0073564

Nord, C. L., Dalmaijer, E. S., Armstrong, T., Baker, K., & Dalgleish, T. (2021). A causal role for gastric rhythm in human disgust avoidance. *Current Biology*, *31*(3), 629-634.e3. https://doi.org/10.1016/j.cub.2020.10.087

Nottebohm, F. (2002). Neuronal replacement in adult brain. *Brain Research Bulletin*, *57*(6), 737–749. https://doi.org/10.1016/S0361-9230(02)00750-5

Penn, D. J., & Számadó, S. (2020). The handicap principle: How an erroneous hypothesis became a scientific principle. *Biological Reviews*, *95*(1), 267–290. https://doi.org/10.1111/brv.12563

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters*. Allen Lane.

Riebel, K., Lachlan, R. F., & Slater, P. J. B. (2015). Learning and cultural transmission in chaffinch song. In M. Naguib, H. J. Brockmann, J. C. Mitani, L. W. Simmons, L. Barrett, S. Healy, & P. J. B. Slater (Eds.), *Advances in the Study of Behavior* (Vol. 47, pp. 181–227). Academic Press. https://doi.org/10.1016/bs.asb.2015.01.001

Russell, W. M. S., Burch, R. L., & Hume, C. W. (1959/1992). *The principles of humane experimental technique*. Universities Federation for Animal Welfare.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century-Crofts.

Skinner, B. F. (1988). Preface to The behavior of organisms. *Journal of the Experimental Analysis of Behavior*, *50*(2), 355–358. https://doi.org/10.1901/jeab.1988.50-355

Smith, J. M. (1991). Theories of sexual selection. *Trends in Ecology & Evolution*, *6*(5), 146–151. https://doi.org/10.1016/0169-5347(91)90055-3

Swanson, L. W., Newman, E., Araque, A., & Dubinsky, J. M. (2017). *The beautiful brain: The drawings of Santiago Ramon y Cajal*. Abrams.

The Deer Year | Isle of Rum Red Deer project. (n.d.). Retrieved 30 August 2022, from https://rumdeer.bio.ed.ac.uk/deer-year

Tinbergen, N. (1951). *The study of instinct*. Oxford University Press.

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift Für Tierpsychologie*, *20*(4), 410–433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

Vallortigara, G., & Rogers, L. J. (2020). A function for the

bicameral mind. *Cortex*, *124*, 274–285. https://doi.org/10.1016/j.cortex.2019.11.018

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177. https://doi.org/10.1037/h0074428

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, *3*(1), 1–14. https://doi.org/10.1037/h0069608

Whitaker, H., & Jarema, G. (2017). The split between Gall and Spurzheim (1813–1818). *Journal of the History of the Neurosciences*, *26*(2), 216–223. https://doi.org/10.1080/0964704X.2016.1204807

Yoshihara, M., & Yoshihara, M. (2018). 'Necessary and sufficient' in biology is not necessarily necessary – confusions and erroneous conclusions resulting from misapplied logic in the field of biology, especially neuroscience. *Journal of Neurogenetics*, *32*(2), 53–64. https://doi.org/10.1080/01677063.2018.1468443

About the author

Professor Pete Clifton
UNIVERSITY OF SUSSEX

Pete Clifton is Professor of Psychology at the University of Sussex. He was the founding Head of the School of Psychology, holding that position from June 2009 to July 2014. His research has focused on the different roles of the brain transmitter serotonin in motivation, especially feeding, and cognition. He is a Chartered Psychologist and Fellow of the British Psychological Society.

PART II
ORGANISATION OF THE NERVOUS SYSTEM

In the next few chapters, we'll explore how information is received and processed by the brain, leading to generation of behaviour. The first step in this process is to learn how the nervous system is organised, in terms of the cells and structures that it contains.

2.

EXPLORING THE BRAIN: A TOUR OF THE STRUCTURES OF THE NERVOUS SYSTEM

Dr Catherine N. Hall

Learning Objectives

By the end of this chapter, you will:

- understand the organisation and main components of the nervous system
- have a sense of how information flows

through the nervous system.

What does the brain do?

All our thoughts and actions are biological structures and processes that work together to enable us to successfully exist and interact with the world, performing behaviours that keep us fed, watered and safe. In this chapter we will learn about the different parts of the nervous system that orchestrate these behaviours.

First, however, it is worth considering, on a very general level, what our brains and the nervous system do. They take in information from the outside world, and our bodies, and work out what is the best thing to do next. They then cause changes in our bodies to enable that thing to happen, whether that's running away from a lion, catching a ball, or going to sleep.

In this chapter, we are going to explore the structures, circuits and cells of the nervous system, in order to understand broadly how information flows into, through, and from the brain. You will learn a bit about how these structures and cells generate behaviours and internal responses that allow us to successfully adapt to and interact with what's going on around

and inside us. You'll learn much more about this in following chapters of the book.

The nervous system as a computer

The nervous system is the network of **neurons** and supporting cells, termed **glia**, that do this job of detecting something, transmitting that information, integrating it with other information, and sending an instruction to other parts of our body to do something about it. In other words, our nervous system is like a computer. It takes an input, performs a computation on that input (using the programs running on that computer – these determine what computation is performed), and generates an output. In fact, every part of the nervous system does this same 'input – computation – output' job, but using different inputs, running different programs and generating different outputs. The whole nervous system might detect visual information that a lion is coming, compute that it would be a good idea to run away, and generate patterns of muscle contractions in your legs to make you run. On a microscopic scale, a single nerve cell, or neuron, might receive inputs carrying information about light falling on your retina in different locations, and integrate that information to conclude and output the information that the light falling on the retina was forming a vertical line. The program run by a given cell or structure in the nervous system is determined

by how that cell or structure is connected to other cells or structures, and the biological rules that govern those connections and how they change over time. We'll learn much more about that throughout the next few chapters of this book.

First of all though, we need to learn our way around the nervous system. This chapter gives you an introduction to the anatomy of the nervous system. It should help you understand the organisation of the nervous system as well as introduce the function of some of its major components. You'll learn much more about how these structures perform their functions in later chapters.

Parts of the nervous system

The nervous system is made up of the **central nervous system** and the **peripheral nervous system** (Figure 2.1)**.**

The central nervous system (CNS) comprises the brain and the spinal cord, while the peripheral nervous system (PNS) is the network of neurons and nerves that lie outside these two structures and connect the CNS with the rest of the body. It includes most of the cranial nerves, that connect to the brain as well as the spinal nerves that take information to and from the spinal cord. The PNS provides the input to the CNS, which computes what to do with that information, and sends outputs back to the body, via the PNS. Symmetry around the midline is a general feature of nervous system organisation.

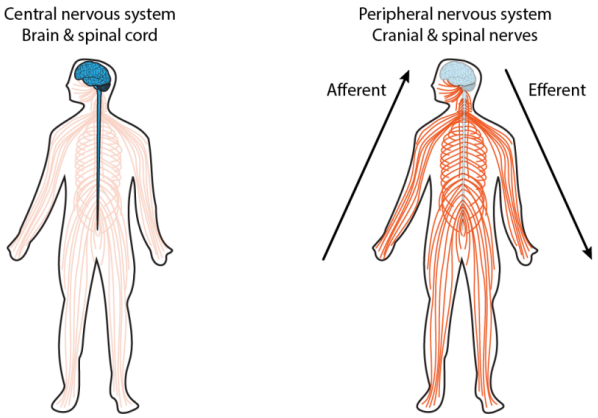

Fig 2.1. The nervous system is divided into the central nervous system, which includes the brain and spinal cord, and the peripheral nervous system, which includes the cranial and spinal nerves.

The peripheral nervous system

The PNS can be subdivided into two parts: the **somatic** and **autonomic** nervous systems. The autonomic nervous system can then be subdivided into three further divisions: the sympathetic, parasympathetic and enteric nervous systems.

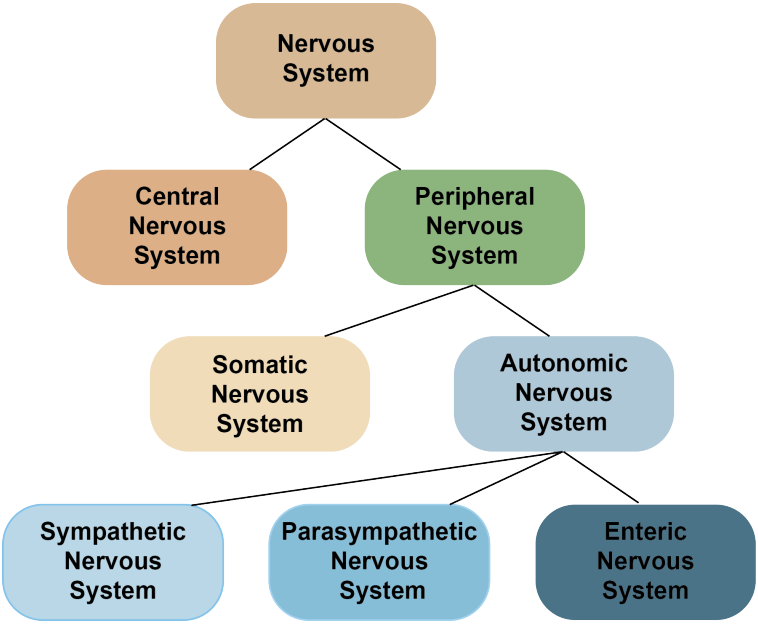

Fig 2.2. The divisions of the nervous system

The somatic nervous system

The somatic nervous system deals with interactions with the external environment: sensing the outside world via sensory neurons, and sending signals via motor neurons to control **skeletal muscles** to generate movements and behaviours that interact with that world. Many of these behaviours are voluntary, and are initiated by complex decision making processes in the brain. You hear a voice calling, you interpret the language, and turn towards the sound of your name. The somatic nervous system can also generate involuntary

movements, however, via reflexes, in which a sensory input activates a motor response without voluntary control. The simplest of these reflexes involve only a single sensory neuron activating a single motor neuron. An example is the muscle stretch reflex, in which sensory neurons detect stretch in a muscle, causing motor neurons to activate the same muscle to contract it more and counter the stretch. So if you lean to one side, stretching core postural muscles, this reflex constricts those muscles, keeping you stable. Or if someone adds a heavy weight to something you're carrying, stretching your arm muscles, they then constrict so you don't drop the load. Even these simplest reflexes involve information transfer from PNS to CNS, as the connection or synapse between these two neurons occurs in the spinal cord – part of the CNS.

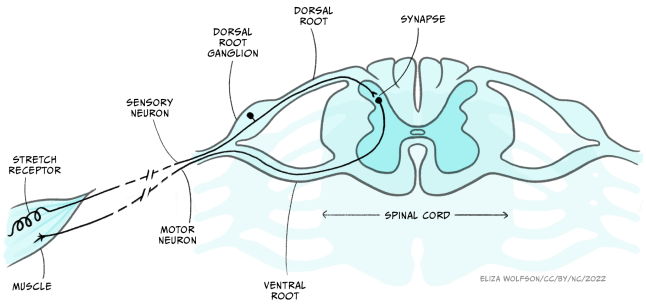

Fig 2.3. The simple spinal cord stretch reflex. Muscle stretch is detected by the sensory neuron which activates a motor neuron in the spinal cord to contract the same muscle.

Indeed, there are no neurons that exist wholly in the peripheral somatic nervous system: somatic sensory neurons synapse for the first time in the CNS, while somatic motor neurons' cell bodies are found in the CNS, with their axons leaving the CNS to innervate muscles.

These afferents (carrying sensory information *inwards* to the CNS) and efferents (carrying motor information *outwards* from the CNS) form **cranial nerves** and **spinal nerves**. (Nerves are just bundles of axons – the long projections that each neuron has to carry electrical impulses). Cranial nerves innervate the head and carry information including about smell, taste, hearing, and control of facial muscles to and from their targets directly into the **brainstem**. Spinal nerves carry information to and from the skin and skeletal muscles to the spinal cord. There are 31 pairs of spinal nerves, which carry sensory and motor information from specific parts of the body into the spinal cord. The region of skin innervated by afferents from a given spinal nerve is called a **dermatome**, while the muscles contacted by efferents from a single nerve are called a **myotome.**

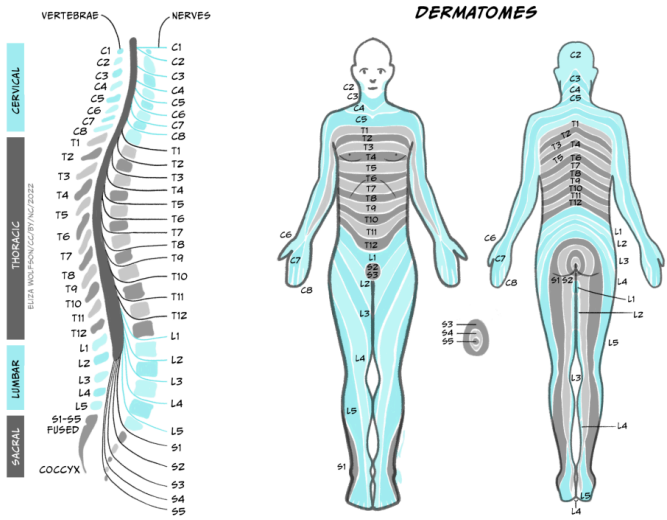

Fig 2.4. Nerves entering the spinal cord at specific segments carry sensory information about specific regions of the skin, termed dermatomes.

The sensory and motor parts of the nerve split apart at the spinal cord. Sensory afferents enter the dorsal root of the spinal cord, their cell bodies forming the dorsal root ganglion just outside the spinal cord. Motor neurons exit the spinal cord from the ventral root (Figure 2.3) before synapsing at **neuromuscular junctions** on skeletal muscle where they release acetylcholine to initiate muscle contraction (see the section Interacting with the world).

The autonomic nervous system

In contrast with the more voluntary control mediated by the somatic nervous system, the autonomic nervous system mediates interactions with the body's internal environment, for example regulating heart rate. These interactions are broadly involuntary reflexes, though modulated by the brain, and some of this regulation can be consciously done, for example people can train themselves to exert control over their heart rate. As in the somatic nervous system, sensory neurons provide information about the internal organs to the CNS, and motor neurons produce effects on the internal organs, often by modulating the tone of **smooth muscle,** for example to change blood vessel diameters. Outside the autonomic nervous system, non-neuronal pathways can also send information about the internal body state to the brain. For example, neurons in a brain region called the hypothalamus can detect increases in blood temperature, activating brain circuits that can then cause autonomic nervous system activation and increase blood vessel dilation in the skin as well as sweat gland activation.

The sympathetic, parasympathetic and enteric nervous systems

Now let's consider the different divisions of the autonomic nervous system:

The **enteric nervous system** is a large mesh of neurons which is embedded in the wall of the gastrointestinal system, from the oesophagus to the anus, and regulates motility and secretion of hormones. In humans, it contains around 500 million neurons, 0.5% of the number found in the brain and 5 times more than are found in the spinal cord. It can function without input from the brain, though can also be regulated by descending input.

The **sympathetic** and **parasympathetic** divisions of the autonomic nervous system are often thought of as the 'fight-and-flight' and 'rest-and-digest' systems, respectively, as they generate motor responses that broadly promote action or relaxation. For example, sympathetic activation increases heart rate, and increases blood flow to the brain, heart and skeletal muscles, priming the body for action. Conversely, activation of the parasympathetic nervous system reduces heart rate and blood flow to the brain, heart and skeletal muscles, instead directing blood flow to the gut and stimulating digestion.

While it is useful to think of the distinction between 'fight and flight' and 'rest and digest' functions of the two systems, the body doesn't switch in a binary manner between one or the other being active but rather the body's state depends on the balance of activity of the two systems at any one time. Furthermore, this balance is not uniform across the body, as is apparent from the need to independently regulate different organs, for example to control heart rate and bladder release.

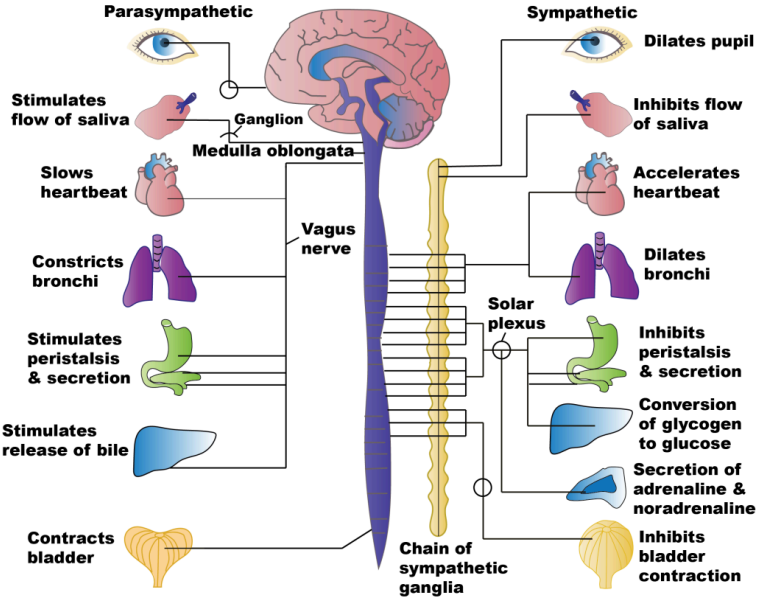

Fig 2.5. The sympathetic and parasympathetic divisions of the autonomic nervous system

The sympathetic **preganglionic neurons** leave the thoracic and lumbar spinal cord to synapse in either the **sympathetic chain ganglia** just outside the spinal cord, or in the **prevertebral ganglia** including the solar plexus or mesenteric ganglia within the abdomen. Preganglionic neurons use acetylcholine as their neurotransmitter. Postganglionic neurons – the motor neurons of the sympathetic division – use noradrenaline as their neurotransmitter, and often travel along the same nerves as the somatic nervous system [NB: Noradrenaline is called norepinephrine in the US, and adrenaline is called epinephrine].

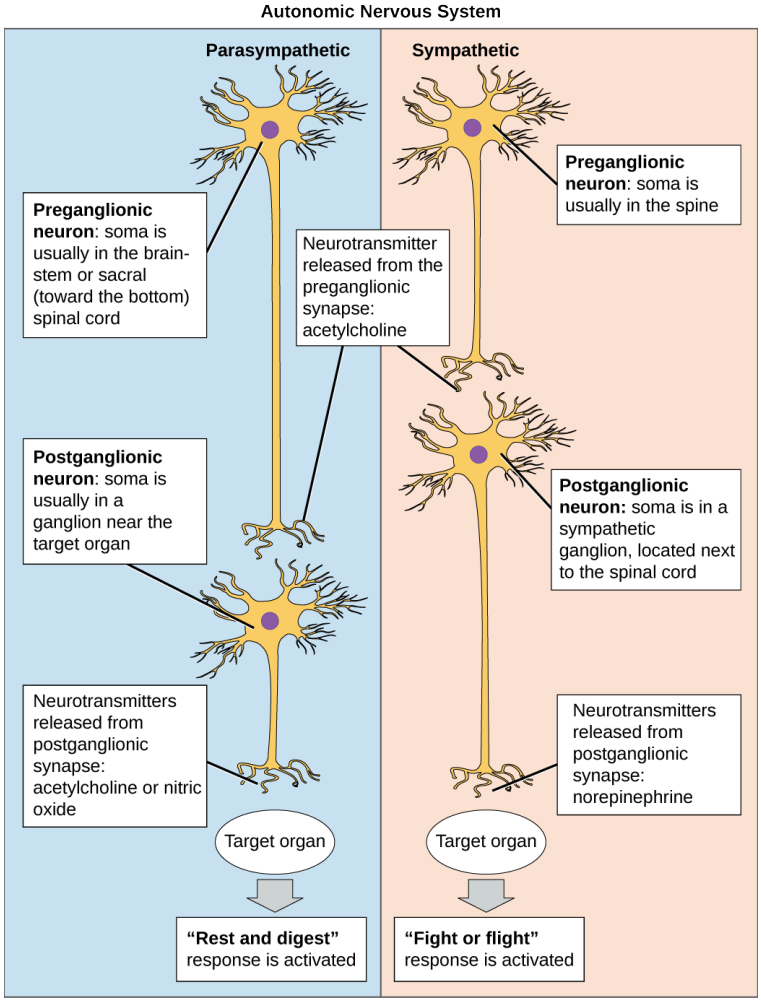

Fig 2.6. Neurotransmitters of the autonomic nervous system [Note the use of the US term norepinephrine, instead of the UK term noradrenaline]

Parasympathetic neurons leave the CNS via cranial nerves or

via sacral regions of the spinal cord. These neurons synapse in ganglia that are generally very close to the organs to be contacted, so parasympathetic preganglionic neurons are much longer than the postganglionic neurons. The vast majority of parasympathetic fibres form the vagus nerve, which innervates most of the organs in the thorax and abdomen. Both pre and post-ganglionic parasympathetic neurons use acetylcholine as a neurotransmitter.

Key Takeaways: Peripheral Nervous System

- The PNS delivers sensory information to the CNS and sends instructions from the central nervous system to control motor outputs
- The PNS is made up of the somatic and autonomic nervous systems, dealing with interactions with the external and internal environments, respectively
- The autonomic nervous system comprises the enteric nervous system in the gut and the sympathetic, and parasympathetic divisions

which have broadly opposing effects on our internal organs
- Sensory neurons synapse first in the CNS. Somatic motor neurons exit the CNS and release acetylcholine onto skeletal muscles, whereas autonomic neurons synapse onto motor neurons at ganglia outside the CNS
- Acetylcholine is the neurotransmitter released by preganglionic neurons and parasympathetic motor neurons, while noradrenaline is released by sympathetic motor neurons.

The Central Nervous System

Compass directions

The CNS comprises the brain and spinal cord. They, particularly the brain, are complex 3D structures, so before we explore them, it's useful to consider the language we can use to describe what exactly we are looking at and where different parts are located with respect to other regions.

We can look at the surface of the brain from different angles. In humans, if we look from the front, we are looking at the

anterior surface, or from the back we are looking at the **posterior** surface. If we look at the top, we are looking at the **superior** surface or from below, the **inferior** surface. These words can also be used to describe relative positions of things within the brain too (e.g. visual cortex is posterior to auditory cortex).

To confuse matters, however, the front of the brain can also be referred to as **rostral** (meaning towards the nose), the back as **caudal** (meaning towards the tail), the top as **dorsal** (towards the back) and the bottom as **ventral** (towards the stomach).

In the brain these terms don't really make sense – dorsal regions are towards the top, not towards the back of the head. They make much more sense in the spinal cord – dorsal spinal cord really is towards the back, not the top. The reason for the confusion is that humans walk upright so our brain is angled relative to the spinal cord.

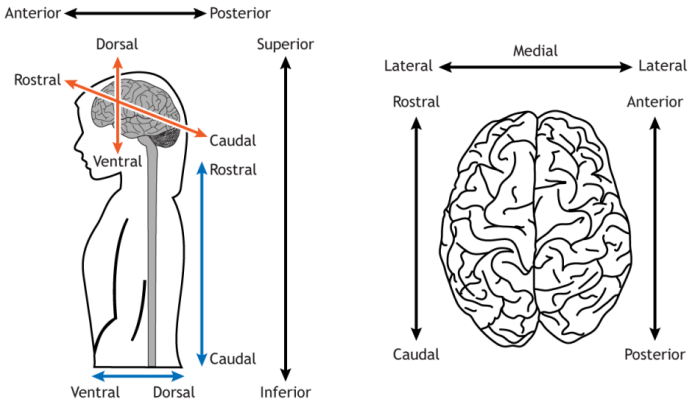

Fig 2.7. Compass directions for the nervous system

In most animals (e.g. think of mice), the brain continues in a straight line from the spinal cord, so the dorsal brain is aligned with the back of the animal. In humans, however, the top of our head points in a different direction to our back. All in all, this means we have lots of words we can use to describe whether we're looking at the front, back, top or bottom of the brain.

We don't just want to look at the brain from the outside surfaces, though, but to see inside at the many structures within. To do that we can virtually or physically slice through it, creating **sagittal**, **coronal** or **horizontal/transverse** slices. In doing so, we can notice that symmetry is a general feature of CNS organisation: the left and right halves of the brain and spinal cord are symmetrical around a midline. We can describe structures' locations with respect to this midline as being **medial** (closer to the midline) or **lateral** (closer to the

side), as well as describing their anterior-posterior/rostral-caudal and superior-inferior/dorsal-ventral dimensions.

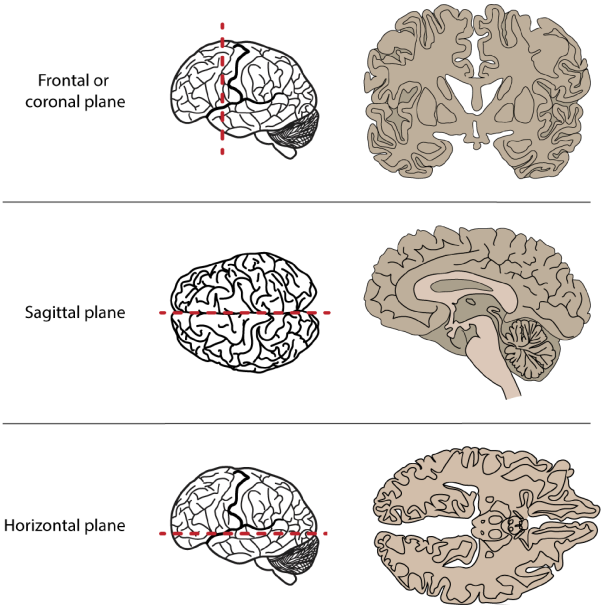

Fig 2.8. Anatomical slices allow us to visualise inside the brain

The spinal cord

The spinal cord can be divided into segments, each of which connects to a pair of sensory and motor nerves. Towards the head are 8 cervical segments, below which are 12 thoracic segments, 5 lumbar segments and 5 sacral segments (Figure 2.4).

The spinal cord carries afferent, somatosensory (touch)

information up to the brain and efferent (motor) information to the muscles of the body. It comprises grey matter (neuronal cell bodies and short range connections) around a central canal, containing cerebrospinal fluid, surrounded by a number of white matter tracts containing myelinated and unmyelinated axons forming connections to other regions. Both the grey and white matter are organised. For example, the dorsal column of the white matter contains axons from somatosensory neurons whose cell bodies are located in the dorsal root ganglia, while the lateral corticospinal tract contains axons from motor neurons from the cerebral cortex which control voluntary movement of the limbs. The grey matter is where connections between different neurons form and contains cell bodies, dendrites and synapses as well as axons. Spinal cord grey matter can be divided into three 'horns' (Figure 2.9), the dorsal horn containing neurons carrying sensory information, the lateral horn largely containing sympathetic motor neurons, and the ventral horn containing cells conveying motor information.

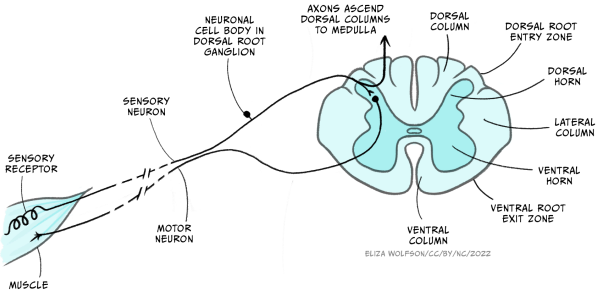

Fig 2.9. Organisation of the spinal cord. White matter is depicted in light blue and grey matter in darker blue.

The brain

The brain itself is made up of the brainstem, the cerebellum and the forebrain.

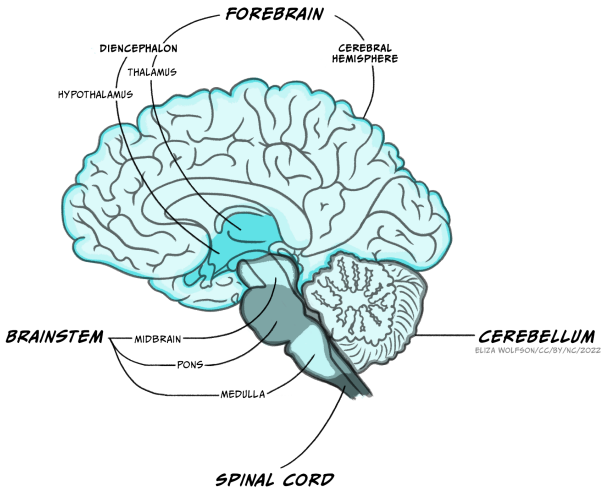

Fig 2.10. Lateral view of the brain

The brainstem

A lot of the volume of the brainstem contains white matter tracts carrying information up to the rest of the brain or down to the spinal cord, as well as to and from the cranial nerves that provide sensory and motor inputs to the face and neck. Within the medulla, the pyramids, so called for their shape, are prominent white matter bundles that carry descending motor axons to the spinal cord. On the ventral surface of the pons are the bridge-like wide mass of transverse fibres (perpendicular to

the axis of the brain stem and spinal cord), from which the pons derives its name (*pons* is Latin for bridge). These fibres connect the brainstem to the cerebellum.

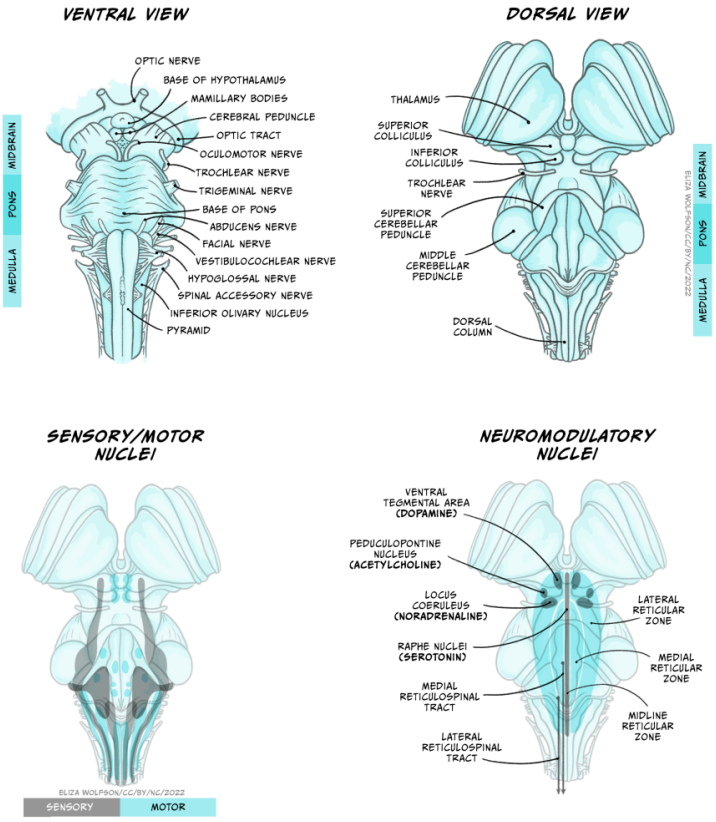

Fig 2.11. Brainstem (A) Ventral view (B) Dorsal view (C) Sensory and motor nuclei (D) Neuromodulatory and regulatory nuclei

Nestled within the numerous white matter tracts passing

through the brainstem are a number of grey matter nuclei (clusters of neuronal cell bodies). These nuclei include motor or sensory cranial nerve nuclei, containing cell bodies of neurons that project into or receive connections from the cranial nerves, as well as the dorsal column nuclei where many neurons carrying touch information from the spinal cord form their first connections. Also within the brainstem are nuclei that produce neuromodulators that are released over relatively large regions of forebrain and are involved in regulating arousal, attention, mood, movement, motivation and memory. Many of these neuromodulators are well known and will make several appearances elsewhere in this book: **Dopamine** is produced in neurons in the ventral tegmental area and substantial nigra pars compacta of the midbrain, **serotonin** from neurons in the Raphe nuclei that extend from the medulla to the midbrain, **noradrenaline** in neurons in locus coeruleus of the midbrain and the medial reticular zone of the pons, and **acetylcholine** in neurons of the pedunculopontine nucleus in the pons (as well as in the basal forebrain). Other brainstem nuclei, particularly in the medulla, regulate key functions for sustaining life, including breathing, heart rate, swallowing and consciousness. Brainstem damage can therefore be life-threatening, and can occur due to brain swelling compressing the brainstem against the skull.

Cerebellum

Fig 2.12. The cerebellum

The cerebellum, or ‘little brain’, lies inferior to the occipital and temporal lobes of the cerebral cortex, and posterior to the pons.

Fig 2.13. A horizontal slice of the cerebellum showing the layered, folded structure. From Sobotta's *Textbook and Atlas of Human Anatomy*, 1908

Its cells are organised in clear layers – that is, it has a laminar structure – with a distinct connectivity that has made it a very interesting structure for neuroscientists to study. It receives inputs via 'mossy fibres' from nuclei in the **pons,** which in turn receive information from wide areas of the cerebral cortex, containing sensory and other information. These mossy fibres then synapse onto **granule cells** – small neurons which are packed together to form the granule cell layer. The human brain contains 50 billion granule cells – about 3/4 of the total number of neurons in the brain. The axons of these granule cells rise vertically from the granule cell layer into the molecular layer, where they split in two and send axons in opposite directions, forming a T shape. The axons of

different granule cells are aligned parallel with each other, and are termed **parallel fibres**.

These parallel fibres synapse onto the highly branched, flat dendritic trees of **Purkinje cells**, the output cell of the cerebellum which sends connections to the deep cerebellar nuclei from which information is sent to the thalamus and onto the cerebral cortex. Purkinje cells also receive synaptic input from 'climbing fibres', axons of cells that originate in the medulla and carry information from across the brain, particularly information about ongoing motor processes.

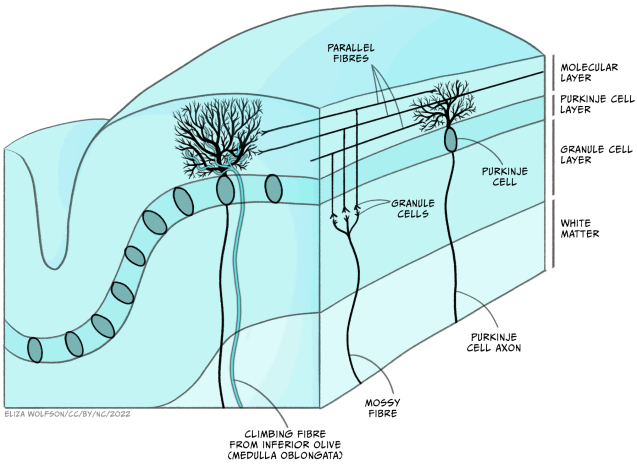

Fig 2.14. Cellular organisation of the cerebellum

Neuroscientists have been able to interrogate and understand how this highly organised circuitry mediates some of the key

functions of the **cerebellum**, allowing us detailed insights into how the cerebellum performs some of its key functions. The cerebellum is important for bringing together diverse sensory information and using this to guide motor behaviours, making it important for balance and motor learning, such as learning to ride a bicycle, or your fingers learning to play a new tune on the piano. However, while the sensory-motor functions of the cerebellum are the best understood, it also receives many different sorts of information from across the brain. Functional imaging studies and other experiments have demonstrated cerebellar involvement in processes as diverse as language comprehension, autobiographical memory and attention.

Forebrain

The forebrain comprises the diencephalon and, surrounding it the cerebrum – two cerebral hemispheres containing an outer layer- the cerebral cortex and subcortical structures such as the hippocampus, basal ganglia and amygdala.The two cerebral hemispheres are connected to each other via the **corpus callosum**, a very large white matter tract, with many other white matter tracts connecting different parts of the forebrain to each other.

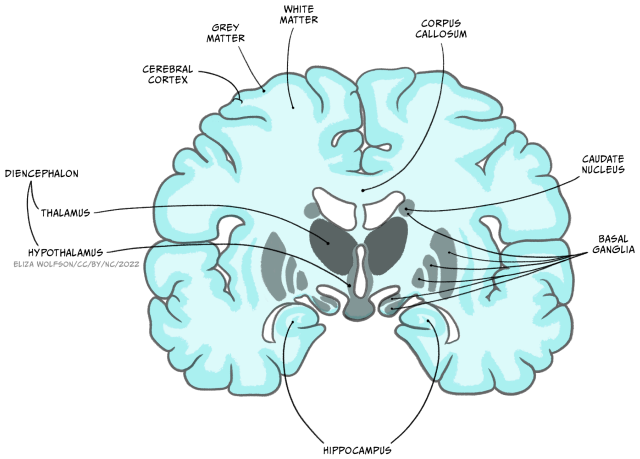

Fig 2.15. Coronal section of the forebrain showing its major structures

Diencephalon

Extending from the midbrain, the diencephalon's major components are the thalamus and the hypothalamus.

The **thalamus** is an information hub, relaying ascending and descending information from widespread brain areas. It is organised into functionally specialised nuclei which process information of certain modalities. For example the dorsal lateral geniculate nucleus of the thalamus receives visual information from the optic nerve and sends projections to primary visual cortex, while the medial geniculate nucleus

receives auditory information from the inferior colliculus and projects to auditory cortex. Rather than simply relaying information in one direction, however, a key feature of thalamic processing is that nuclei also receive descending information from the cortex, forming circuits termed **thalamocortical loops** (or **corticothalamic loops**). These thalamocortical loops are not limited to sensory processing, but from higher order areas and motor areas as well, and can include additional structures in their circuitry.

For example, the anterior thalamus plays an important role in memory, receiving information from the hippocampus, mamillary bodies of the hypothalamus as well as the cerebral cortex, and projecting to cingulate cortex. Thalamocortical loops that incorporate the striatum and other nuclei of the basal ganglia are also important for motor control and motivated behaviour, via the ventrolateral, mediodorsal and anterior thalamic nuclei. These loops, and others like them we will hear about in other circuits, demonstrate that the 'input-computation-output' function of the nervous system is not simply in one direction – instead, outputs from a given region often feed back to structures providing inputs to that region, as well as sending outputs to 'upstream' brain areas.

The hypothalamus is located below the thalamus, above the pituitary gland. It consists of around 22 nuclei and is highly connected to the brainstem, the amygdala and the hippocampus. The hypothalamus is involved in regulation of many homeostatic processes, such as the control of eating and

drinking, temperature regulation and circadian rhythms, as well as emotion and memory processing and sexual behaviour. Some hypothalamic nuclei are sexually dimorphic, being structurally and functionally different in males and females. The hypothalamus can effect changes on the body's physiology both by its projections via the brainstem to the autonomic nervous system, and by regulating hormone release via its connections with the adjacent pituitary gland. It is also involved in motivated behaviours such as defensive freezing or flight behaviours.

Cerebral cortex

Richly folded in humans, to maximise its surface area, the cerebral cortex is the outermost layer of the forebrain.

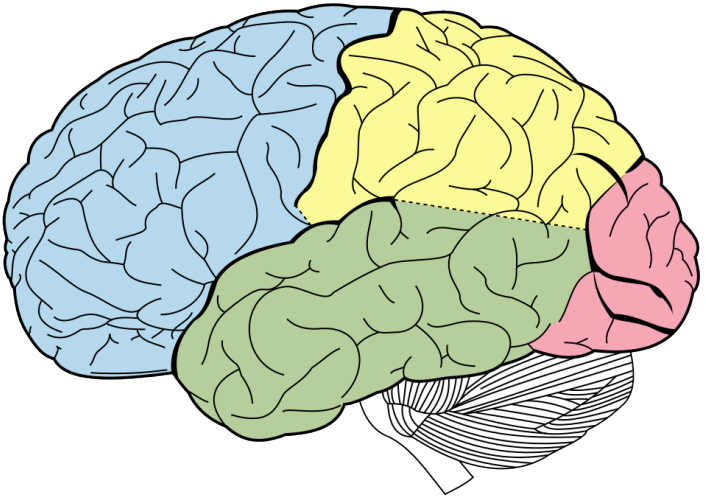

Figure 2.16. The lobes of the cerebral cortex (frontal: blue; parietal: yellow; occipital: red; temporal: green)

These folds form characteristic **sulci** (grooves) and **gyri** (ridges), the largest of which separate the cerebral cortex into 4 lobes, the frontal, temporal, parietal and occipital lobes. Most cerebral cortex is neocortex (new cortex) with 6 layers of neurons, containing different densities and types of neurons. Layer 1 has very few cell bodies and mostly contains the tips of dendrites and axons. Layers 2 and 3 contain cell bodies of neurons that receive and send projections to nearby cortical regions. Ascending inputs to the cerebral cortex from the thalamus arrive in layer 4, while cells that send descending projections to other brain areas are found in layers 5 and 6. These layers are different thicknesses across the cortex, varying

with the function of that area. Sensory cortices receive lots of afferent inputs so have a thick layer 4, while motor regions send a lot of efferents to downstream regions, so have thick layers 5 and 6. There is more connectivity vertically through these layers than horizontally, so neurons in the same vertical 'column' of cortex tend to have the same response properties (i.e. they are all activated by the same sort of stimulus).

By studying the **cytoarchitecture**, or organisation of the cell layers across the cortex, in 1909 a German anatomist called Korbinian Brodmann divided the cerebral cortex into 52 different areas, now called **Brodmann areas**. Many of these have subsequently been subdivided into smaller regions. The different cellular organisation of these regions indicates differences in the circuitry and information processing within the region. Indeed many Brodmann areas have been shown to correspond with different functional specialisations. Area 17 is primary visual cortex, for example, and area 4 is primary motor cortex. Generally the functions of different cortical regions can be categorised as being **sensory**, **motor** or **associative**. Sensory information first enters primary sensory cortices, with further processing occurring in secondary sensory areas. Where multimodal information is processed within an area (e.g. both auditory and visual), that area is considered association cortex. These areas are important for a multitude of functions from understanding and generating language to spatial processing, abstract thinking, planning and memory. Conversely, primary motor cortex contains motor neurons

that send their axons to the spinal cord to execute voluntary movements, while secondary or premotor areas project to primary motor cortex and help select or coordinate movements.

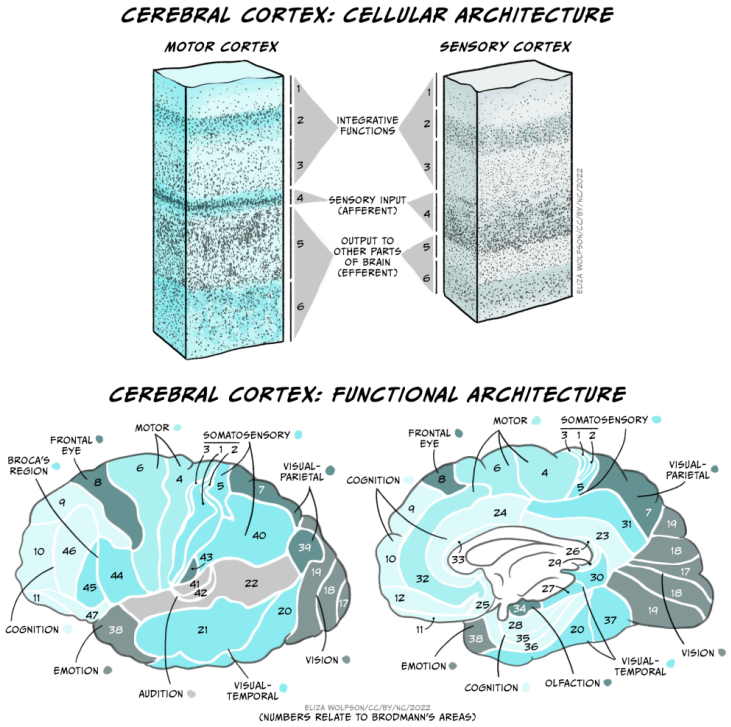

Fig 2.17. Cytoarchitecture of the cerebral cortex and Brodmann's areas

These functional areas as defined by cytoarchitecture can be further functionally subdivided. For example, primary sensory

cortices are topographically organised: adjacent parts of the skin are represented by adjacent bits of somatosensory cortex – **somatotopic organisation**, and adjacent regions of the retina are represented by adjacent parts of primary visual cortex – **retinotopic organisation**. These representations can be even further subdivided into columns processing different stimuli (orientation of visual stimuli, for example).

While most brain areas are structurally symmetrical, there is lateralisation of some functions that are subserved by the cerebral cortex. Firstly, as mentioned above, sensory processing generally occurs in the opposite side of the body from where that sensory input is received (i.e. left somatosensory cortex processes stimuli applied to the right side of the body). Secondly, lesion studies reveal lateralisation in the function of information streams through each side of the brain. Lesions to the left hemisphere visual processing streams result in deficits in perceiving fine details, while lesions to the right hemisphere impair perception of the wider field of view, or 'big picture'. Finally, language production and comprehension are typically localised to the left hemisphere, particularly in right-handed people.

Neurons in the cerebral cortex project to a multitude of different brain areas as well as to the spinal cord, but the most common projection target for cortical neurons are other cortical neurons, and most of these projections are to nearby neurons. 80% of intracortical projections occur to neurons in the same area, while most connections between areas are also

to nearby areas. Only 5% of intracortical connections are long range, to more distant cortical regions or across the corpus callous (**transcallosal**), between the two hemispheres. These connections contribute to the formation of several large scale brain networks that contribute to different aspects of perceptual and cognitive function (see later).

Basal ganglia

The **basal ganglia** are a group of subcortical nuclei (i.e. they lie beneath the cerebral cortex). They comprise the **dorsal striatum**, which is made up of the caudate nucleus and the putamen, the ventral striatum, or nucleus accumbens and the external and internal global pallidus. Two other components of the basal ganglia circuitry are actually not in the cerebrum: the subthalamic nucleus is within the diencephalon and the substantial nigra is in the midbrain.

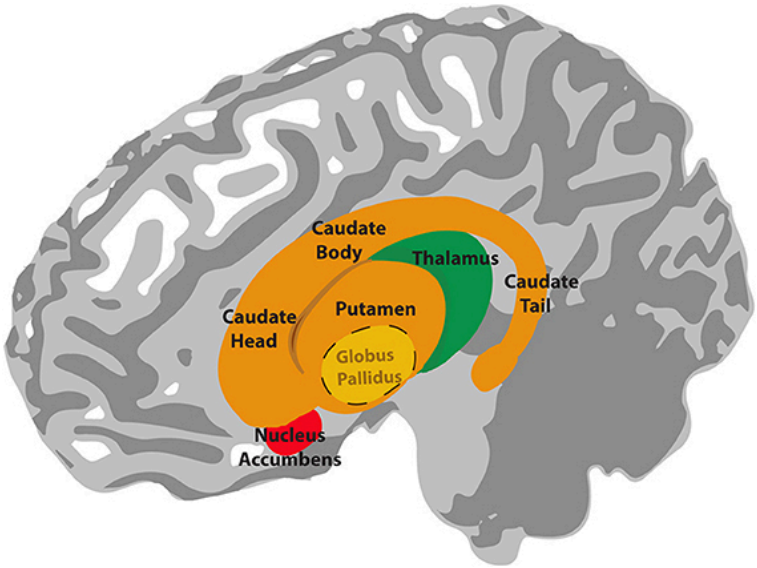

Fig 2.18. The basal ganglia

Information flows from widespread areas of the cerebral cortex, through the basal ganglia and thalamus, and back to cerebral cortex. These cortico-basal ganglia-thalamo-cortical loops are important for selecting motor actions, e.g. starting and stopping behaviours, and for aspects of motivated behaviour i.e. selecting actions based on whether they are likely to result in something good or bad happening to the individual. They include excitatory and inhibitory pathways, the balance of which is important for inhibiting or initiating motor outputs (see Interacting with the world).

Disruptions to this circuitry can cause an imbalance between these facilitatory and inhibitory effects, as is seen in a number of neurological conditions including Parkinson's

disease and Huntingdon's disease, as well as schizophrenia, Tourette's syndrome, obsessive-compulsive disorder and addiction (see Dysfunctions of the nervous system).

Hippocampus

The hippocampus is an important structure involved in episodic memory, spatial processing and contextual learning.

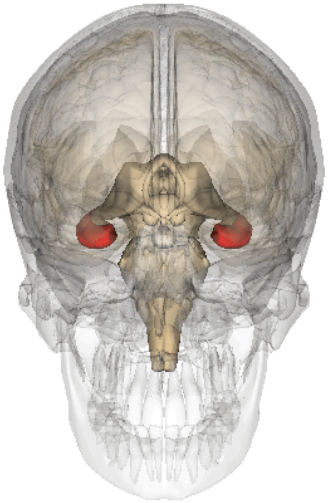

Fig. 2.19. The hippocampus

It has a distinctly different laminar structure to the cerebral cortex. It is allocortex, with fewer cell layers than neocortex, with a densely packed layer of pyramidal cell bodies, above and below which are layers with only dendritic and axonal processes and much sparser inhibitory neurons. It is formed of two interlinked U-shaped folds, the dentate gyrus and the hippocampus 'proper', curved together into an elegant 3D shape, like a seahorse (hippocampus) or a ram's horn (Cornu ammonis) from which the hippocampal subfields CA1, CA2 and CA3 are named. The hippocampus receives most of its inputs from cortical or subcortical regions via the **entorhinal cortex** and most of its outputs are sent to cortical or subcortical regions via the **subiculum**. The fornix, another important output pathway, also connects the hippocampus, via CA3, to the mammillary bodies of the diencephalon.

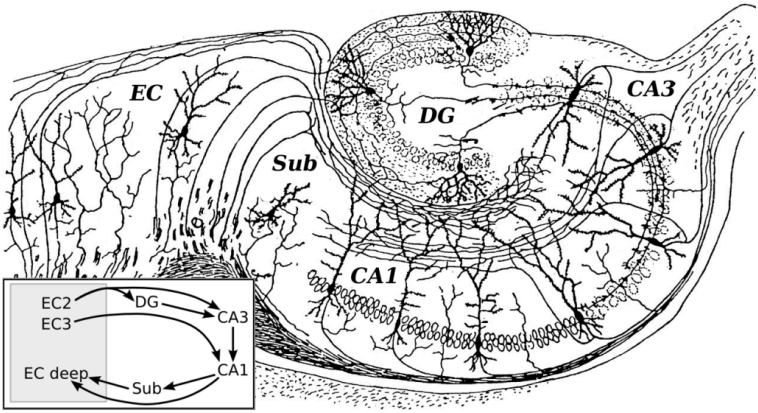

Fig. 2.20. Hippocampal circuitry as drawn by Ramon y Cajal in 1911. EC: Entorhinal cortex; Sub: subiculum; DG: dentate gyrus

Like the cerebellum, the hippocampus has a well-characterised neuronal circuitry that has provided useful insights into how it fulfils its function. From the entorhinal cortex, inputs via the perforant path synapse on granule cells in the dentate gyrus, which themselves send mossy fibres to CA3. CA3 pyramidal neurons send Schaffer collaterals (axon branches) to CA2 (which is small) and CA1, as well as recurrent collaterals – axon branches that synapse back onto CA3 cells. Changes, during learning, in the strength of connections in these subregions are thought to be important for particular aspects of memory formation and retrieval, particularly the ability to remember more of an event or stimulus when exposed to only part of it (pattern completion), and to remember events or stimuli as distinct from each other (pattern separation).

Damage to the hippocampus produces memory deficits and occurs early in Alzheimer's disease. Selective hippocampal damage and associated amnesia can also occur when the brain is deprived of oxygen, for example during birth. Because of its recurrent collateral connectivity, whereby excitatory neurons can excite themselves, the hippocampus is also a common focus of epileptic activity, where neuronal circuits are over activated, producing seizures.

Amygdala

The amygdala, named for its almond shape, sits adjacent to the hippocampus beneath the cerebral cortex within the temporal lobe (see Fig. 2.19).

It is important for processing of emotions and for the impact of emotions on learning and has been shown to be particularly involved in fear learning. It is made up of a complex of different nuclei, including the basolateral, corticomedial and centromedial nuclei. It receives inputs from wide regions of sensory and prefrontal cortex, then hippocampus and visceral information from brainstem nuclei, making it able to integrate information about the state of the body with contextual information. Its efferents go to cerebral cortex, particularly prefrontal and cingulate cortex, hippocampus, ventral striatum, thalamus and hypothalamus. This connectivity allows it to produce emotional responses appropriate to a given context; for example, when a stimulus

appears that is associated with punishment, a fear response can be produced by altering hormone release via the hypothalamus, triggering freezing behaviours and activation of the sympathetic nervous system via the brainstem.
People with lesions to the amygdala display a reduction in emotional behaviour and a placidness or 'flatness of affect', and show reduced learning about emotional or frightening stimuli or situations.

Key Takeaways: Central Nervous System

- The CNS is made up of the brain and spinal cord
- We can describe where we are in the CNS using the compass directions lateral & medial, anterior & posterior, superior & inferior, dorsal & ventral and rostral & caudal
- The spinal cord takes information to and from the brain to the periphery, as well as performing some information processing in its central grey matter

- The brainstem contains lots of white matter tracts and nuclei containing motor and sensory information and neurons that deal with automatic regulatory functions such as control of heart rate and breathing
- The cerebellum is a laminar structure with a distinct circuitry that supports sensory-motor learning
- The thalamus is an information hub that transfers information to the brain from the periphery and vice versa as well as participating in lots of thalamocortical loops with widespread cortical regions
- The hypothalamus contains lots of nuclei that have important roles in control of homeostasis such as thirst and appetite regulation and is also involved in memory
- The cerebral cortex has 6 layers which are different thicknesses in different functional regions. The cerebral cortex connects to subcortical regions but most often neurons connect to other cortical neurons
- The basal ganglia are subcortical nuclei that are important for initiating and selecting motor behaviours and for motivated

behaviour

- The hippocampus is an important structure for memory and has a distinct circuitry that supports its role in forming associations
- The amygdala is intimately connected with the cortex, hippocampus and brainstem and is important for emotional learning and behaviours.

Non-neuronal brain structures

In addition to these various brain regions, a number of other structures of the brain's anatomy are important to appreciate in order to understand the overall physiology of this organ.

Ventricles

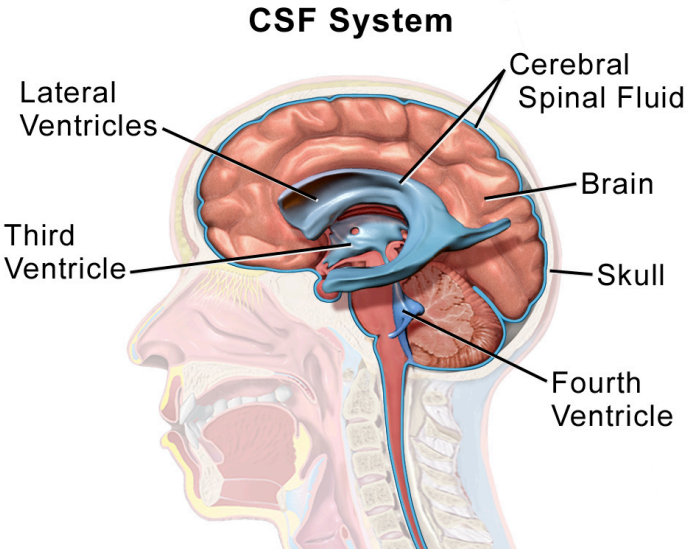

Fig 2.21. The ventricles and CSF

Cerebrospinal fluid and meninges

Within the brain and spinal cord are a series of connected spaces containing cerebrospinal fluid (CSF). The central canal of the spinal cord extends into the brainstem before expanding in the region of the pons to form the fourth ventricle. At the top of the fourth ventricle another narrow channel, the cerebral aqueduct, connects to another broader chamber, the

third ventricle, at the level of the diencephalon. From the third ventricle another two narrow channels connect to the large lateral ventricles that extend deep into the cerebrum. The CSF that fills the canals and ventricles is made by ependymal cells that line the ventricles in a specialised membrane structure called the choroid plexus. These ependymal cells surround capillaries of the vascular system and produce CSF by filtering the blood. CSF is similar in content to blood plasma, being mostly water but containing ions and glucose (click here for a table with more information), though it has less protein content than plasma.

CSF flows from the fourth ventricle into the space between the membranes or **meninges** that cover the brain. There are three meninges; the **pia** is a delicate membrane that directly covers the brain and spinal cord. Above the pia is a fluid-filled space, then the **arachnoid** membrane, so called for its web-like strands that connect to the pia through this subarachnoid space. Above the arachnoid is the tough outer membrane – the **dura** – which supports large blood vessels that drain cerebral blood towards the heart. CSF flows from the ventricles into the subarachnoid space then circulates the brain, before draining into veins in the dural sinuses. CSF functions as a shock absorber for the brain, cushioning the brain from damage during knocks to the head, as well as clearing metabolic waste products from the brain into the blood.

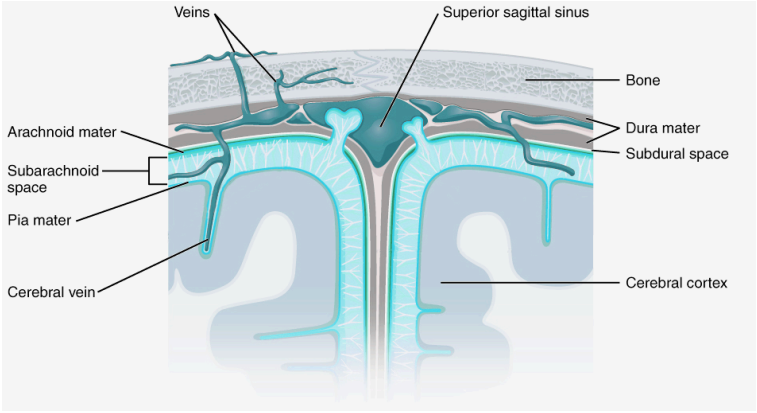

Fig. 2.22. The brain's meninges

Vasculature

The brain is an energetically demanding organ. It is only 2% of the body's mass, but uses 20% of its energy when the body is resting (i.e when muscles are not active). It relies on a constant supply of oxygen and glucose in the blood to sustain neurons, with disruption of blood flow to the brain leading to a loss of consciousness within 10 seconds. To deliver a constant flow of oxygenated blood, the brain has a complex and tightly regulated vasculature that directs blood to the most active brain regions. Four arteries feed the brain with oxygenated blood, forming a circle – the **circle of Willis** – that ensures that a reduction of blood flow to one artery can be compensated for by redistribution of flow from the others. Several large arteries branch off the circle of Willis to perfuse different regions of the brain. Branches of these major arteries

form smaller and smaller arteries and arterioles that pass through the subarachnoid space before diving into the brain, before branching yet further to form a dense capillary network.

There are a mind-blowing 1 to 2 metres of capillaries in every cubic millimetre of brain tissue. These capillaries are less than 10 microns in diameter, though, so they only take up about 2% of the brain volume. This means that, in cerebral cortex, each neuron is only around 10-20 microns from its nearest capillary. This dense vascular network can therefore supply oxygen and glucose very close to active neurons.

Specialised mechanisms allow the blood supply to be fine-tuned to brain regions that need it most. Cells called smooth muscle cells or pericytes in blood vessel walls can dilate or constrict to regulate blood flowing through the vessel. Active neurons and astrocytes produce molecules that dilate smooth muscle cells and pericytes on local arterioles and capillaries, increasing blood flow to these regions of increased activity. In fact this increase in blood flow usually supplies more oxygen than is needed, so that blood oxygen levels increase in active brain regions. This increase in blood oxygen gives rise to the **BOLD** (blood oxygen level dependent) signal that can be detected using magnetic resonance imaging and is often used as a surrogate for neuronal activity in experiments studying the function of different brain regions.

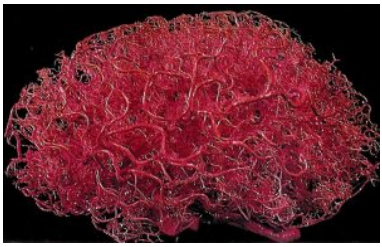

Fig 2.23a. A cast of the brain's vasculature

Another specialised feature of the brain's vascular system is the blood-brain barrier (BBB). The endothelial cells lining blood vessels in the brain are very tightly joined together and express relatively few transporter proteins that allow molecules to be transported from the blood into the brain. This means that it is harder for molecules and cells to access the brain from the blood, protecting the brain from circulating toxins or immune cells.

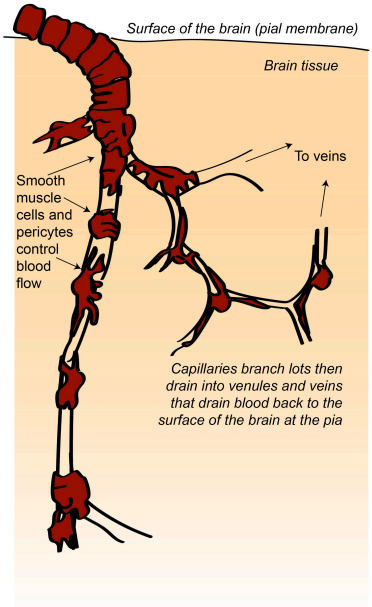

Fig 2.23b. Arterioles on the surface of the brain dive into the brain and branch many times, forming a dense capillary network, before draining to venues that take blood back to the surface of the brain

The brain's vasculature can be impacted in several neurological conditions, leading to alterations in brain function. In ischaemic or haemorrhage stroke, there is a reduction in the blood supplied to the brain due to either a blockage or leakage in blood vessels feeding the brain. This reduces the oxygen available to the brain region fed by the lesioned vessel, damaging the neurons in that region and causing corresponding functional deficits.

In other diseases there is a less severe decrease in blood flow. In Alzheimer's disease there is a decrease in brain blood flow many years before symptoms develop. It is not yet known what links this decrease in blood flow has with the development of Alzheimer's disease, but a decreased ability to clear toxic proteins from the brain, an increase in BBB permeability, or a chronic lack of oxygen supply may all be factors.

Key Takeaways: Non-neuronal brain structures

- The ventricles are cavities within the brain that contain CSF
- The brain is surrounded by 3 layers of meninges
- CSF flows between the ventricles and the subarachnoid space within the meninges
- The brain's blood vessels are specialised, to

direct blood flow to active brain regions and to regulate the degree to which molecules and cells in the blood can access the brain.

Overall in this chapter, you have learnt general principles about information flow in the brain, as well as some of the major neuronal and non-neuronal structures that mediate this transfer of information. Next we will explore the signalling and non-signalling cells of the nervous system to understand how they support information flow and computation.

References

Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Mateos-Pérez, J. M., Evans, A. C., & The Alzheimer's Disease Neuroimaging Initiative. (2016). Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. *Nature Communications 7*, 11934. https://doi.org/10.1038/ncomms11934

Kwon, Diana. The mysterious, multifaceted cerebellum. (2020). *Knowable Magazine*. https://doi.org/10.1146/knowable-093020-2

About the author

Dr Catherine Hall
UNIVERSITY OF SUSSEX
https://twitter.com/cathnaledi

Dr Catherine Hall is a member of the Sussex Neuroscience Steering Committee, the University Senate, convenes the core first year module *Psychobiology,* and lectures on topics relating to basic neuroscience, neurovascular function and dementia.

3.

UNDER THE MICROSCOPE: CELLS OF THE NERVOUS SYSTEM

Dr Catherine N. Hall

Learning Objectives

By the end of this chapter, you will be aware of:

- the main cells in the nervous system
- what these cells look like
- what their functions are.

In each region of the central and peripheral nervous systems

are specialised cells that perform or support the fundamental function of the nervous system – the detection of information about the world, integration with information about the internal body state and past experience, and the generation of an appropriate behaviour. These cells can broadly be classified as either neurons or glia.

Neurons are the cells that perform the signalling and information processing. They detect inputs, integrate information, and send signals to other cells, be they other neurons (forming neuronal circuits) or non-neural cells (such as muscles or endocrine cells), to produce a behavioural or physiological effect. **Glia** or glial cells play numerous supporting roles for the neurons. This support was originally thought to be structural – the term 'glia' is derived from the Greek for 'glue' – but is now appreciated to be highly complex, involving dynamic communication between glia, neurons and other cell types, and is able to modulate neuronal communication. In addition to neurons and glia, neural tissue contains a large number of vascular cells. As we saw in Chapter 2, Exploring the brain, brain tissue is densely vascularised in order to provide sufficient metabolites for energy-hungry neurons to function correctly.

Neurons

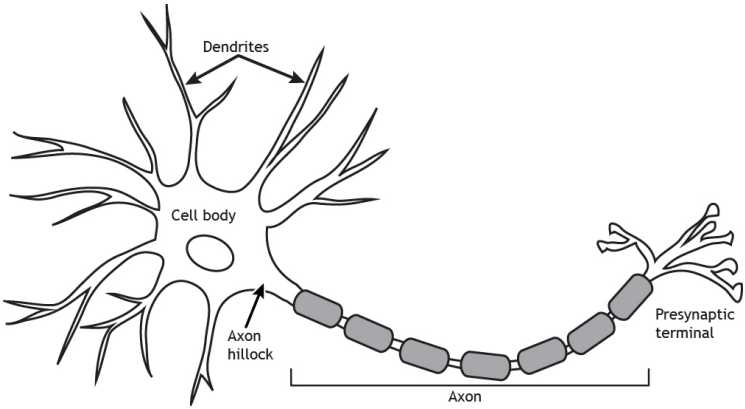

Fig 2.24. A typical neuron. Dendrites branch out from the cell body, where the nucleus is located. The axon hillock is located where the cell body transitions into the axon. The axon begins at the axon hillock and ends at the presynaptic terminal, which can branch into multiple terminals.

The basic form of a neuron is shown in Fig 2.24. It has a cell body, or **soma**, with branching processes called **dendrites** and a thin process called an **axon**. The axon can also be branched, forming **axon collaterals**. We talked in the last chapter (Exploring the brain) about the general function of the brain being to take in information, perform a computation on that information to work out what to do next, then to produce an output. As mentioned above, this same 'input-computation-output' function is also performed by individual neurons.

The dendrites, or dendritic tree, are where most inputs to

the cell are received. These inputs are integrated across the dendritic tree and soma before the cell 'decides' whether the inputs are strong enough to trigger an electrical output signal down the axon (the **action potential**, see Chapter 5: Neuronal transmission). The site of this decision is the top of the axon furthest from the terminals, termed the **axon initial segment**. The action potential travels along the axon to the axon terminal. The axon terminal is very close to, but not touching, a dendrite of another cell. This tiny gap is specialised for passing messages between two cells and is called a **synapse**. At the synapse, action potentials cause release of a chemical messenger – **a neurotransmitter** – which transmits the signal across the gap to the next cell in the circuit.

Neuron morphology affects computation

All neurons have this basic morphology, but nevertheless come in a multitude of shapes and sizes.

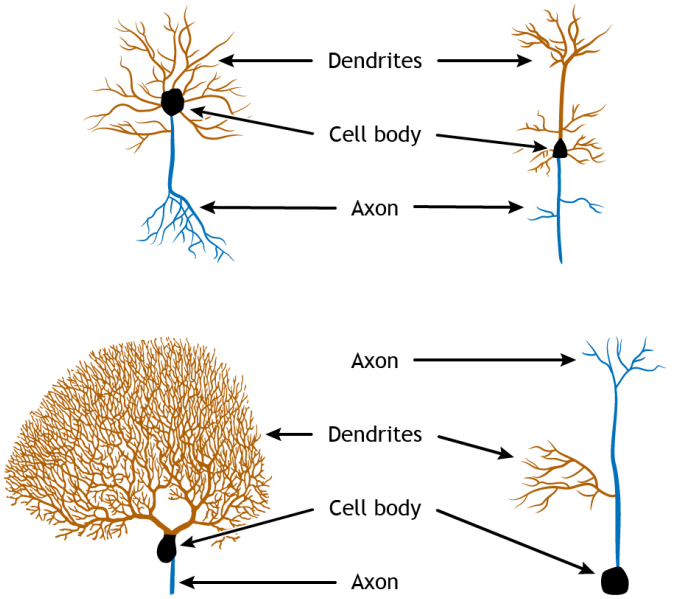

Fig 2.25. Neuron structure is variable, but the main components of cell body (shown in black), dendrites (shown in brown), and axon (shown in blue) are common among all neurons.

Most neurons are multipolar neurons, with a branched dendritic tree and a single axon. Some neurons, particularly sensory neurons (e.g. in the retina), are bipolar having a single dendrite coming out from one end of the soma, and a single axon from the other (though these may be branched near their ends). Pseudounipolar neurons have a single process, classified as an axon, which receives inputs at one end and releases transmitter at the other end. These different shapes and sizes alter how neurons perform computations.

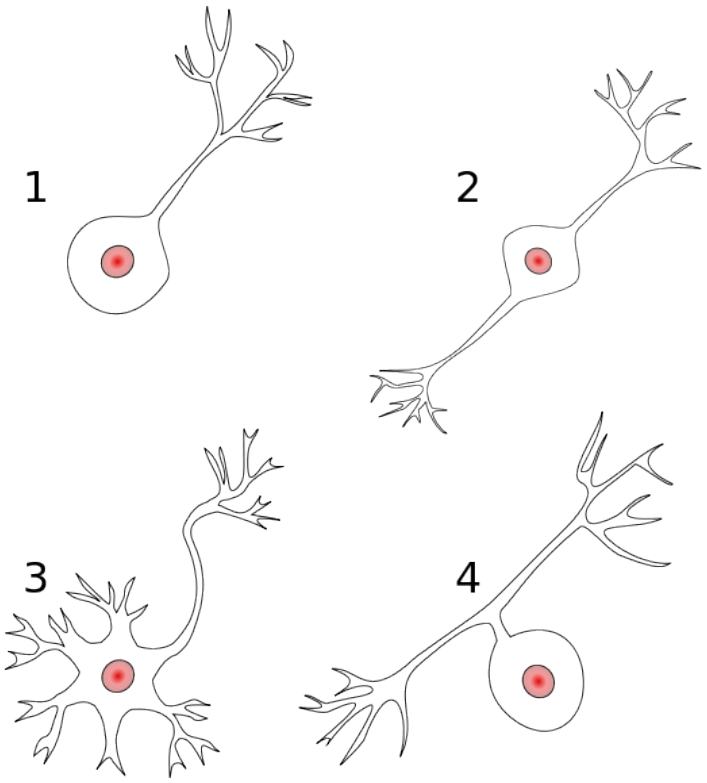

Fig 2.26. Different categories of neurons: 1: Unipolar neuron 2: Bipolar neuron 3: Multipolar neuron 4: Pseudounipolar neuron

Because the job of a neuron is to add up all its inputs and decide whether to fire an output action potential, the number of these inputs and where they are located affects how this summation happens.

For example, in the cerebellum (Figure 2.12), the Purkinje

cells receive inputs from granule cell axons and axons from deep cerebellar nuclei in the pons, called climbing fibres.

The climbing fibres are very branched and make lots of connections (synapses) to each Purkinje cell, while the granule cells' axons, called parallel fibres, are simple and form only a single synapse to each Purkinje cell. This means that the connection between a single granule cell and a Purkinje cell is weaker than the connection between the climbing fibre neuron and the Purkinje cell.

Another way that neuron morphology can affect the computations it undertakes can be seen if we zoom in on the dendrites (Figure 2.27).

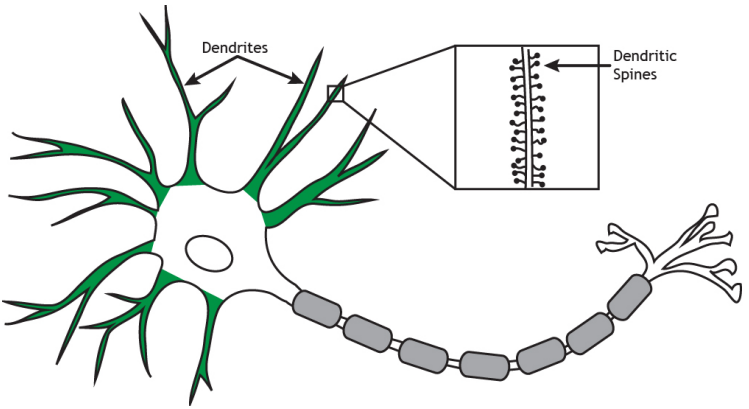

Fig 2.27. Dendrites branch out from the soma. Their function is to receive information from other neurons. Some dendrites have small protrusions called spines that are important for communicating with other neurons.

Some dendrites are covered with small protrusions, called dendritic spines, whereas others are smooth. Synapses can form on to the spine, or onto the neck of the spine, and this means that some inputs can 'gate' the effect of other inputs, altering their impact on the neuron.

Different classes of neurons

There are many different ways in which neurons can be classified and subdivided, depending on what aspect of neuronal function is being focussed on. As we have seen above, we can define neurons by their morphology, and morphology can also be used to further classify the multitude of multipolar neurons: For example, pyramidal cells have a characteristic pyramidal shaped soma, long dendrite pointing upwards (an apical dendrite), tufty basal dendrites, and an axon that often forms several collaterals. Purkinje cells of the cerebellum have a round soma, a flat highly branched dendritic tree at the top of the soma and a single long axon. Granule cells have a small cell body, a simple dendritic tree and an axon that splits in two. Chandelier cells have a highly branched axon arbour that forms distinctive 'candle-like' connections with the initial segments of lots of axons.

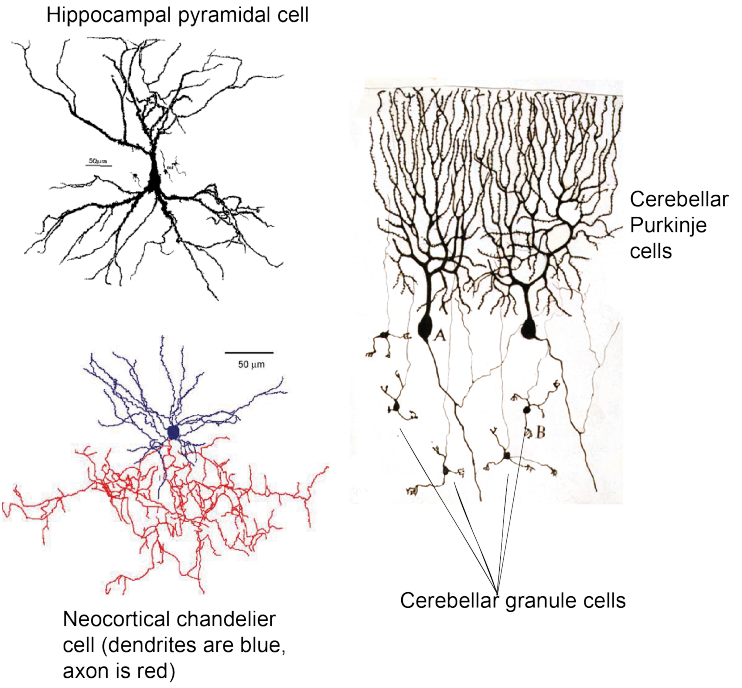

Fig 2.28. Examples of different types of multipolar neuron

We can also define neurons by their effect on other neurons, being **excitatory** or **inhibitory**, depending on whether they make the neurons they connect to more or less likely to fire an action potential (more of this in Chapter 5: Neuronal transmission).

Of the examples given above, pyramidal cells and granule cells are excitatory and Purkinje cells and chandelier cells are inhibitory. Neurons can also be classified by the type of neurotransmitter they release: glutamatergic neurons release glutamate, GABAergic neurons release GABA (gamma

aminobutyric acid), dopaminergic neurons release dopamine, and so on. As we will see later, these categories broadly overlap: glutamatergic neurons are excitatory, because glutamate excites cells and GABAergic neurons are inhibitory, because GABA inhibits cells. However other neurotransmitters such as dopamine can have different downstream effects depending on what proteins that cell expresses at the synapse.

Neurons can also be classified based on their connectivity and role in a circuit, but this can get complicated! Neurons that project a long way to a different brain region are termed **principal neurons**, while those that project locally are termed **interneurons.** Principal neurons are often excitatory, but not always (for example Purkinje cells output information from the cerebellum and are inhibitory). However, in some brain areas it is hard to decide whether a cell should be termed an interneuron or not. Is it helpful to call neocortical pyramidal cells that project to far cortical areas 'principal cells' but very similar cells that project to the next cortical column 'interneurons'? Are cerebellar granule cells interneurons because they project within the cerebellum, though they project to a distinct cell layer? Instead, the term 'interneuron' is only commonly used for inhibitory cells, referred to as inhibitory interneurons. A chandelier cell is an example of an inhibitory interneuron. Excitatory cells in local circuits are instead usually referred to by other features, e.g. location and morphology (e.g. a Layer 5 neocortical pyramidal cell as

distinct from a Layer 2/3 pyramidal cell in the example given above).

Glia

There are five main types of glial cells.

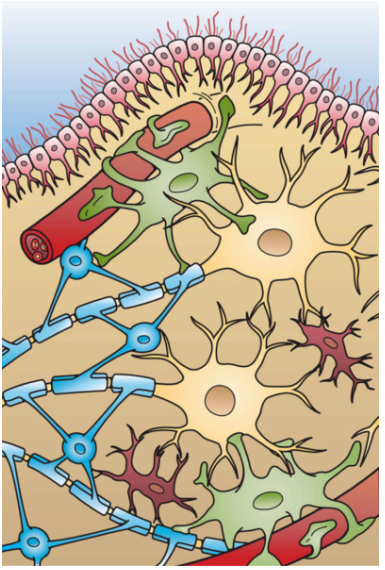

Fig 2.29. Glial cells: Astrocytes (green), oligodendrocytes (blue), microglia (maroon) and ependymal cells (pink). Blood vessels are shown in red.**Astrocytes**, thus termed because of their star-like morphology, have many, many fine processes that encircle synapses – each human astrocyte can contact up to 2 million synapses. These processes not only provide a physical support to neuronal connections but also play lots of active roles to support neuronal function and communication. For example, astrocytes are an important for removing neurotransmitter

from synapses, taking it up across their membranes to 'reset' synapses after synaptic transmission, and can also regulate the levels of ions in the extracellular space.

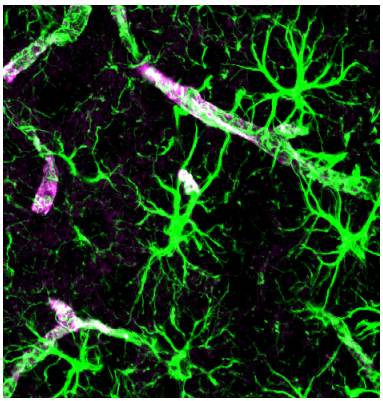

Fig 2.30. Astrocytes (green) wrap around blood vessels (magenta) as well as synapses.

Astrocytes can also release many substances onto neurons and other cells (e.g. ATP, lactate, glucose), modulating their activity and providing metabolic support. Specialised astrocyte processes, termed 'end feet' communicate with local blood vessels, altering local blood flow and taking up glucose from the blood. These end feet surround blood vessels, forming part of the BBB, and others extend to the surface of the brain, forming a thin layer just under the pia. This barrier of astrocyte end feet is termed the glia limitans, and stops (or regulates) molecules and cells from entering or leaving the nervous tissue. Astrocytes also react to damage to brain tissue, becoming 'activated' and expressing different molecules when they are exposed to infection. They can form a scar around sites of damage. While this can be helpful, it also causes problems if cells remain activated for a long time, and

the scar tissue that forms can stop neurons from making new connections through.

Oligodendrocytes and Schwann Cells perform similar roles in the CNS and PNS respectively. Both cells wrap layers of a fatty substance called **myelin** around neuronal axons. Oligodendrocytes do so by sending multiple myelinating processes to nearby axons, whereas Schwann cells in the PNS each have only one myelinating process. These layers of myelin insulate axons allowing action potentials to be conducted more quickly and robustly (see Chapter 5: Neuronal transmission).

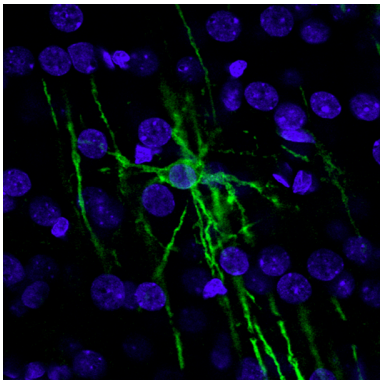

Fig 2.31. Oligodendrocyte (green) with cells' nuclei labelled in blue.

Being so closely associated with axons, oligodendrocytes and Schwann cells also provide support to axons by releasing some molecules and taking up others, regulating the extracellular environment around axons in a similar manner to the role of astrocytes at synapses. In multiple sclerosis, the body's immune cells seem to attack oligodendrocytes, leading to demyelination of axons. This impairs signalling in the axons that were ensheathed by the

damaged cells, causing a variety of neurological problems, depending on which axons are affected and what signals they were carrying.

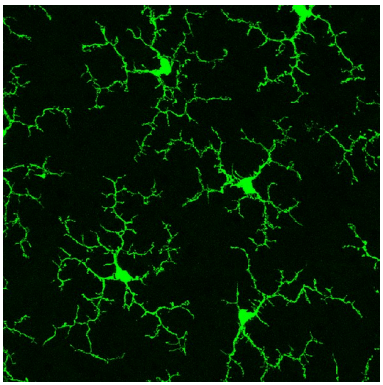

Fig 2.32. Microglia

Microglia are small cells which play an important role in repairing damaged brain tissue. Unlike the rest of the body, immune cells cannot readily enter the brain from the blood because they are prevented by the BBB. Instead, microglia act as the brain's resident immune cell. Their processes constantly extend and retract, surveying the brain for signs of damage or infection. When they find a site of damage, they 'activate' and migrate to this region, forming a barrier between healthy and damaged tissue and removing debris of dying cells. Microglia are also often associated with synapses and blood vessels, and increasingly appreciated to have important roles not just in controlling damage, but in shaping normal brain function as well, regulating synaptic transmission and signalling to blood vessels. Like astrocytes, while microglial responses to damage are usually thought to be beneficial, when they are activated for a long period of time they can themselves be harmful. This may happen in disease such as Alzheimer's, and after a stroke.

The last type of glial cell is the **ependymal cell**, which we discussed briefly in the last chapter. These cells line the ventricles and produce CSF.

Vascular cells

We heard in the previous chapter that the brain contains a dense vascular network to provide a constant, tightly regulated supply of energy (mainly oxygen and glucose) to the brain. The main cells that make up the blood vessels are endothelial cells, which form the vessel wall next to the blood, and vascular mural cells: smooth muscle cells on larger vessels (arteries and arterioles) and pericytes on smaller vessels (capillaries), which wrap around endothelial cells (Figure 2.33).

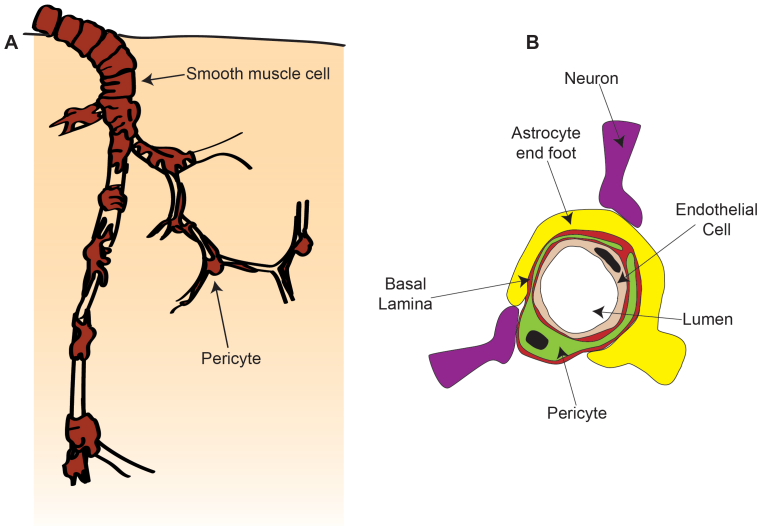

Fig 2.33. Cell types of the cerebrovasculature: A) Pericytes and smooth muscle cells; B) The neurovascular unit comprises the cells of the blood vessel and astrocytes and neurons in the brain that wrap around the vessels.

As we heard earlier, **endothelial cells** form tight junctions with adjacent endothelial cells and with pericytes, forming the BBB. A major function of endothelial cells is to regulate entry of cells and molecules across the BBB, as well as to clear waste molecules from the brain into the blood. They regulate the entry of small molecules by expressing different transporter proteins which allow certain molecules to cross the BBB into the brain. To allow immune cells into the brain, endothelial cells express proteins that stick to immune cells in the blood which then crawl between or through the endothelial cells.

Another function of endothelial cells is control of blood

flow. Endothelial cells can respond to signals in the blood or the brain to produce molecules that contract or dilate smooth muscle cells or pericytes, altering the diameter of blood vessels and changing blood flow.

Smooth muscle cells are ring-shaped cells that wrap around the vessel, while **pericytes** have a distinct cell body and processes that extend along and around the blood vessel. In addition to responding to signals from endothelial cells, these smooth muscle cells and pericytes can also constrict and dilate in response to signals from neurons and astrocytes, changing blood flow to match alterations in neuronal activity. Pericytes are also important for stabilising newly formed-blood vessels and work together with endothelial cells to control the BBB.

The brain is full!

All these cells with their complex structures are often shown in cartoons with lots of space between them. In reality, however, the different cells' processes are closely intertwined and crammed together, taking up almost all the available space. We can see this experimentally by looking at 3D reconstructions from electron micrographs – images of sequential slivers of a very small bit of tissue taken with an electron microscope.

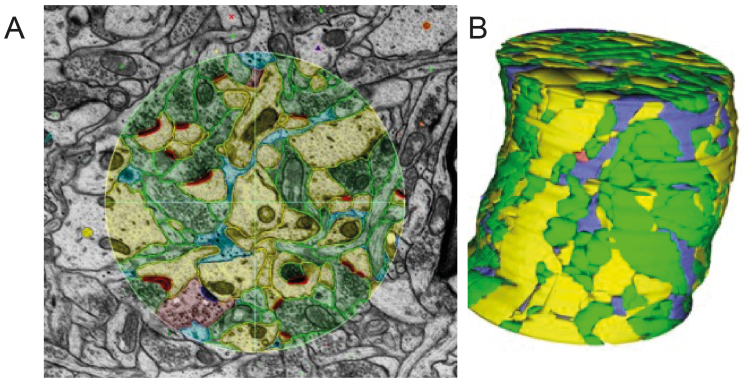

Fig 2.34. Electron micrograph of rat brain, segmented to show different structures (A), and reconstructed in a 3D volume (B). The different structures fill the space. Spiny dendrites (yellow); excitatory axons (green); an inhibitory axon (pink), synapses (red), and astroglia (light blue)

In these image sequences, structures can be labelled and traced in each image, and then assembled to get a reconstruction of the different cells in a small volume of tissue. From such images we can see in superb detail how different structures connect, e.g. which dendrites an axon contacts. However, tracking and reconstructing every process in even a small volume is very computationally expensive, and it's not yet possible to do this for even a whole cortical column, never mind a whole brain or brain region.

Key Takeaways: Cells of the nervous system

Neurons all have a **soma**, **axon** and **dendrites** but come in lots of shapes and sizes, and produce different neurotransmitter molecules, meaning they can perform lots of different computations.

There are 5 different types of **glia** in the brain (astrocytes, oligodendrocytes, Schwann cells, microglia and ependymal cells), which have many roles, including:

- controlling the extracellular environment for neurons (astrocytes, oligodendrocytes, Schwann cells, microglia)
- providing physical and metabolic support (astrocytes, oligodendrocytes, Schwann cells)
- insulating axons to allow fast neuronal transmission (oligodendrocytes, Schwann cells)
- detecting and combatting infection and tissue damage (microglia, astrocytes)

- contacting and communicating with blood vessels (astrocytes, microglia)
- producing CSF (ependymal cells)

Endothelial cells, smooth muscle cells and pericytes form a dense network of blood vessels and control the brain's energy supply, as well as what cells and molecules can go between the blood and the brain tissue.

About the author

Dr Catherine Hall
UNIVERSITY OF SUSSEX
https://twitter.com/cathnaledi

Dr Catherine Hall is a member of the Sussex Neuroscience Steering Committee, the University Senate, convenes the core first year module *Psychobiology,* and lectures on topics relating to basic neuroscience, neurovascular function and dementia.

PART III
NEURONAL COMMUNICATION

Having learnt how the nervous system is made up of cells and structures that receive, process and pass on information to select and generate behaviours, we are now going to learn how neurons actually perform this key function, by interrogating the mechanisms by which they receive, integrate and generate signals in order to communicate with each other.

4.

ELECTROPHYSIOLOGY: ELECTRICAL SIGNALLING IN THE BODY

Dr Catherine N. Hall

Learning objectives

By the end of this chapter, you will:

- understand common electrical terms and how they relate to electrical signalling by neurons
- understand the ionic basis of the membrane potential.

Key electricity concepts

Neurons signal electrically.

They receive inputs from other cells, sum up all these inputs, and generate an electrical impulse, called an action potential, which they send along their axon. Neurons are not the only cells that to use electricity to function. Muscle cells also use electrical signals to constrict and dilate.

We will be going into some detail to understand how neurons are able to use electricity to signal in this manner, but first it's worth going over some key electricity concepts.

- Electrical **currents** are flows of charged particles. In an electrical circuit in a torch, where a battery powers a lamp (Figure 3.1a), the charged particles are negatively charged electrons flowing in a wire. In your body the charged particles are **ions** – such as the sodium ion, Na^+.
- Charged particles flow because they are repelled by similar charges and attracted by opposite charges, i.e. positively charged particles attract negatively charged particles, while negatively charged particles repel other negatively charged particles, and positively charged particles repel other positively charged particles.
- Charged particles only flow if they can pass through the substance that they are in. Electrons can only flow around an electrical circuit when the circuit is complete.

If the circuit is broken by opening a switch, because the electrons can't easily pass through air, they can't flow any more and the torch lamp will go off. The ability of a material to let electricity flow through it is termed **conductance**. The inverse of conductance is **resistance** – a measure of how much a material resists the flow of electricity.

- **Voltage** is a measure of how much potential there is for charged particles to flow and is a measure of stored electrical energy. This electrical potential is analogous to storing water high up in a water tower. Because of gravity, the water has lots of potential to flow, but it cannot do this until a tap is opened. When a tap is turned on water flows out of the pipe (Figure 3.1b). A battery works like a water tower to store electrical energy. Batteries have a positive and a negative pole. When a circuit is connected, electrons are repelled from the negative pole towards the positive pole of the battery.

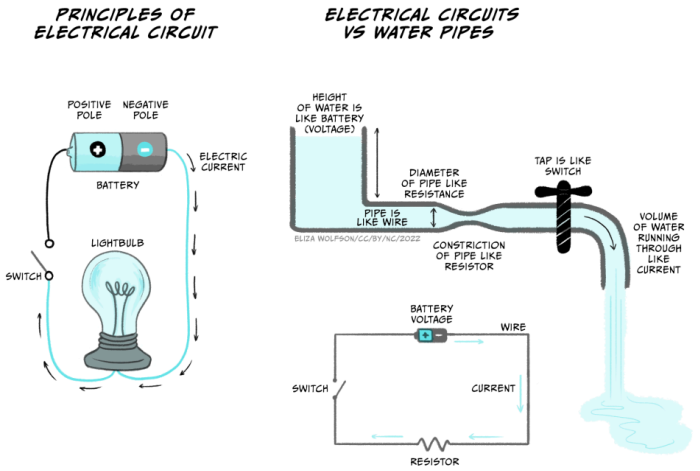

Fig 3.1. Electricity. A) A battery powers a lamp by providing a source of negatively charged electrons which flow from the negative pole of the battery towards the positive pole, through a wire. When the circuit is broken by opening the switch, electricity cannot flow. B) Electrical currents are analogous to water flowing through pipes. Water stored in a water tower has the potential to flow when a tap is opened. The rate of water flow will depend on the stored potential (how high the tower is) as well as how wide the pipes are (how much resistance to flow there is).

The current flowing in a circuit is related to the voltage across the circuit and the conductance or resistance of the wires making up the circuit, according to **Ohm's Law**.

Ohm's Law

Current is proportional to voltage and conductance, and inversely proportional to resistance:

Current = Voltage x Conductance

OR

Current = Voltage / Resistance

You can come back and look at Ohm's law later when you start thinking about currents flowing in neurons.

As you can see, if resistance goes up, and the voltage stays the same, the current (flow of charged particles) will decrease. Conversely when the resistance goes down, the current will increase. In the water pipe analogy (Figure 3.1b), high resistance is like having narrow pipes. If the hole in the middle of the pipe is tiny, you won't get very much water squirting out, never mind how big the water pressure is, but if you make the hole bigger (increasing the conductance or reducing the resistance), a lot of water will flow out of the hole. As the water

tank empties, however, the water pressure (voltage) decreases, and the water flow (current) will reduce.

Nerves conduct electricity more slowly than wires

Electrical currents in the body are not exactly the same as electrical currents in a wire.

In 1849, Hermann von Helmholtz measured the speed that electricity flows in a frog's sciatic (leg) nerve, by stimulating it electrically at one end and measuring the electrical signal at the other end. He found that the speed that electricity flowed (or was 'conducted') down a nerve was 30-40 m/s, around a million times slower than electricity travels through a wire.

So why is electrical signalling in nerves so much slower than in wires? In a wire, electrons (small negatively charged particles) travel along the wire, and they can do this very quickly in materials like metals that conduct electricity well. In nerves, however, the charged particles are **ions**, not electrons. They are positively (or sometimes negatively) charged particles that are much bigger than electrons, and they don't move down the nerve like electrons do. Instead, during a nerve impulse – termed an **action potential** – positively charged ions move into the neuronal axon from the outside. When positive ions move into the cell, the inside of the cell becomes more positive.

This little bit of the axon becoming more positive triggers

positive ion movements into the next little bit of the axon, which also becomes positive, triggering the ion movements across the next bit of axon, and so on, like a Mexican wave of a positive potential flowing along the nerve. On balance there has still been an electrical signal that's moved from one end of the axon to the other, but it has got there more slowly than if electrons had just travelled along the wire.

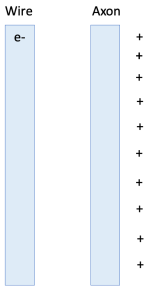

Fig 3.2. Electrical signalling in wires vs. nerves

A short history of electrophysiology

The importance of electricity in animating our bodies – a step, in a way, towards generating behaviours – was discovered in the late 18th century by Lucia and Luigi Galvani.

Fig 3.3a. Lucia Galvani

In a laboratory in their home, the couple discovered that electricity applied to a frog's leg made the muscle twitch. The frog's leg muscle also twitched when it was connected to the nerve with a material that conducts electricity. They concluded that an 'animal electricity' is generated by the body to contract muscles.

Fig 3.3b. Luigi Galvani

The study of how electricity is generated and used by the body is now termed **electrophysiology**. Animal electricity was further studied and made (in)famous by the Galvanis' nephew, Giovanni Aldini, who performed public demonstrations of animal electricity on the bodies of executed prisoners as well as oxen's heads. Tales of these demonstrations of 'Galvanism' inspired the young Mary Shelley to write *Frankenstein*, in which the monster is animated using electricity.

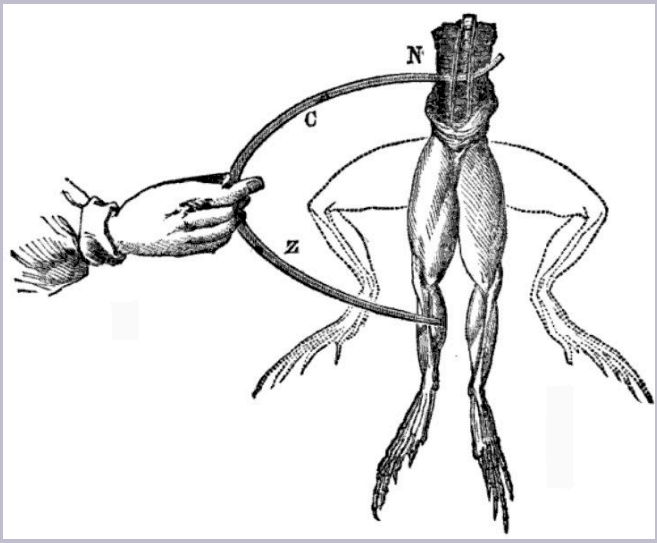

Fig 3.4. The Galvanis' experiment: the frog's legs twitch upwards when the electrode contacts the spinal cord

In the mid-nineteenth century, with the development of tools to measure electrical currents, the German physiologist Emil Du-Bois Reymond was able to measure the change in current that occurs in nerves and muscles when activated – what we now term the 'action potential' – while Hermann von Helmholtz was able to measure the speed of conduction of electrical transmission down a nerve.

Further technological developments allowed Julius Bernstein, who had worked with both Du-Bois

Reymond and Helmholtz, to record the time course of the action potential for the first time in 1868. He showed that the action potential was about 1 ms in duration and that, at its peak, the voltage rises above zero. Bernstein also measured the resting membrane potential as being around -60 mV, building on ideas developed by Walther Nernst, who proposed that the resting membrane potential is set by the potassium conductance of the membrane. Charles Ernest Overton added to this the concept that sodium and potassium exchange is critical for the excitability of cells.

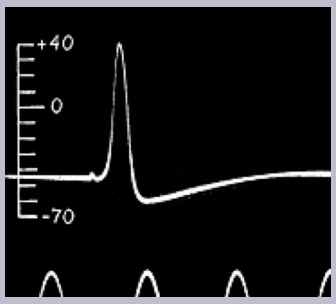

Fig 3.5. The action potential recorded in squid giant axon by Hodgkin and Huxley in 1939

The ionic basis of the action potential was fully elucidated between 1939 and 1952 by Alan Hodgkin and Andrew Huxley, who used the squid giant axon to make the first intracellular recordings of the action potential. They developed the use of the voltage clamp, which uses a feedback

amplifier to hold a cell's voltage at a set level. The feedback amplifier does this by detecting small changes in voltage and injecting current to reverse these changes so that the voltage across the membrane does not change. This injected current is opposite to that flowing across a cell's membrane – if positive charge is flowing into the cell, it depolarises the cell (makes it more positive) and the amplifier will inject negative charge to counter this depolarisation. Conversely positive charge leaving the cell would make the cell would become more negative (hyperpolarised) so the amplifier will inject positive charge to counter the outward positive current and keep the voltage across the membrane constant. Therefore, the amount of current injected by the amplifier can be used to work out what currents are actually flowing across the cell's membrane. Using voltage clamp, Hodgkin and Huxley were able to dissect the inward and outward currents and subsequently mathematically model the properties of sodium and potassium influx to accurately reproduce the action potential. These models were subsequently found to match the gating properties of voltage gated sodium and potassium channels. You'll learn all about this in the next chapter.

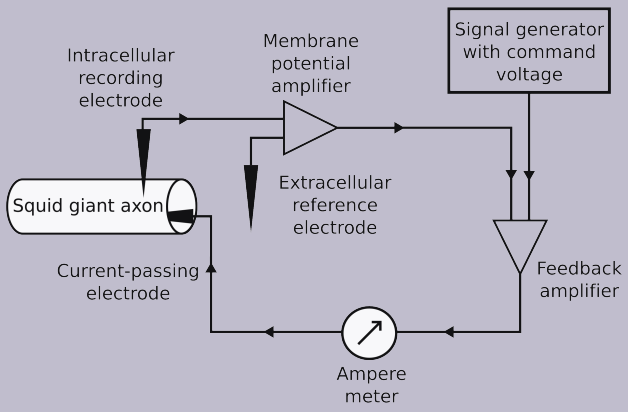

Fig 3.6. The voltage clamp operates by negative feedback. The membrane potential amplifier measures membrane voltage and sends output to the feedback amplifier; this subtracts the membrane voltage from the command voltage, which it receives from the signal generator. This signal is amplified and output is sent into the axon via the current-passing electrode.

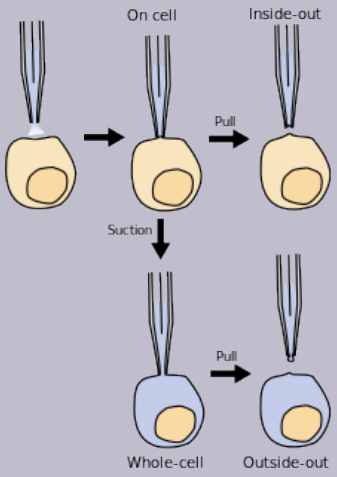

Fig 3.7a. Patch clamp electrophysiology. Different patch clamp methods

The development of patch clamping by Erwin Neher and Bert Sakmann in the 1970s and early 1980s enabled recording of very small current changes, including from single ion channels. Patch clamping involves using a glass microelectrode with a very small tip, that can be placed against a cell membrane. Applying suction tightly seals the electrode tip onto the cell so that current can only flow from the electrode across the attached membrane, reducing noise and allowing the properties of single ion channels to be studied. If the electrode is pulled away from the cell, a little patch of membrane remains on the electrode, forming an inside out patch. Different drugs can then be applied to the bath to see how they change the activity of the ion channels in this tiny patch of membrane. Alternatively, when the electrode is attached to the cell, the membrane patch attached to the electrode

can be ruptured by applying increased suction. This 'whole cell' configuration allows membrane currents from the whole cell to be studied. Pulling the electrode away from the cell at this point can form an 'outside-out' patch. These different adaptations of patch clamp electrophysiology are key tools in the study of electrical properties of neuronal signalling today.

Fig 3.7b. Patch clamp electrophysiology. Single channel currents measured across a small patch of membrane

Overall, electrophysiology has generated a wealth of knowledge about how electrical signals are integrated and generated by neurons, how different ion channels contribute to these signals, and how ion **channelopathies** (dysfunction of ion channels) contribute to disease. For example, Dravet's Syndrome is a severe familial epilepsy that is caused by mutations in the SCN1A gene. This gene encodes a sodium channel that is mostly found in inhibitory interneurons. Because the mutation stops sodium

channels working as well, interneurons are not as able to fire action potentials to inhibit excitatory neurons which become overactive, causing seizures.

How do cells such as neurons signal electrically?

Cells signal electrically by controlling how ions cross their membranes, changing the voltage across the cell membrane. This voltage change across the cell membrane is the electrical signal. The most common ions that move across the cell membrane to cause this voltage change are sodium ions (Na^+), potassium ions (K^+), chloride ions (Cl^-) and calcium ions (Ca^{2+}). These ions carry different charges – sodium and potassium ions each have a single positive charge, chloride has a single negative charge, and calcium ions carry two positive charges. Positively charged ions are called **cations**, while negatively charged ions are called **anions**. The amount of charge carried by an ion is its **valence**.

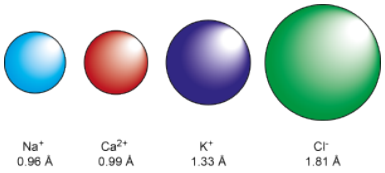

Fig 3.8. Ions are different sizes.

In addition to carrying different charges, ions are different sizes.

Some ion flow, or flux, happens at rest, and other flux happens during signalling.

In this section, we'll consider what's going on at rest, and in the next section, examine what happens to make an electrical signal.

The plasma membrane around a cell is made of a phospholipid bilayer

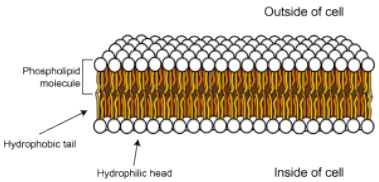

Fig 3.9. Phospholipid bilayer (two parallel layers of phospholipid molecules)

Cells are surrounded by a plasma membrane that keeps the inside separate from the outside. This membrane is made of molecules called **phospholipids**. Phospholipids have three main parts: two fatty tails that are **hydrophobic** (meaning they 'fear water') and a head that is **hydrophilic** (meaning it 'loves water'). Water molecules are slightly charged, with positively charged and negatively charged zones (they are **dipoles**). This means that other particles that are charged are attracted to

them, whereas particles with no charge are repelled by them. The phospholipid head carries a negatively charged phosphate group, so is attracted to water, while the uncharged fatty tails are repelled by water, but will happily mix with other uncharged tails. This means that the phospholipid molecules line up to form a **bilayer** (two parallel layers of phospholipid molecules) with their fatty tails next to each other on the inside of the membrane, and the hydrophilic heads lined up facing the watery inside and outside of the cell.

Small molecules such as oxygen and carbon dioxide can diffuse across the membrane, but because the inside of the membrane is uncharged and hydrophobic, water and other charged particles can't cross it. This means the inside of the cell is kept separate from the outside and the intracellular fluid, or cytosol, inside the cell can have a different constitution than the fluid outside the cell – the extracellular fluid.

Components of intracellular vs extracellular fluid

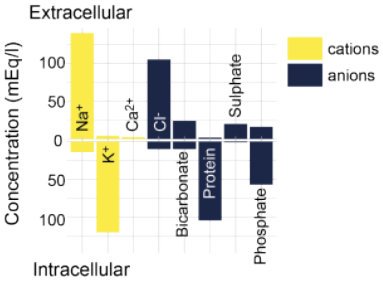

Fig 3.10. Constituents of extra- and intracellular fluid

Intracellular (ICF) and extracellular fluids (ECF) are made up of different substances (Figure 3.10). Both are mostly water, but the concentration of ions and other substances is very different. Of particular note, there is a higher concentration of potassium ions inside the cell compared to outside the cell (~130 mM inside vs. ~4 mM outside), and a high concentration of sodium ions outside the cell compared to the inside of the cell (~145 mM outside vs. ~15 mM inside). There are also more chloride and calcium ions outside the cell than inside the cell. ICF also contains more protein and a higher concentration of organic anions than ECF.

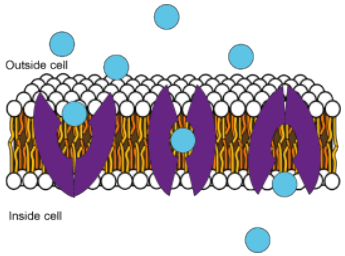

Fig 3.11. Transporter proteins shuttle molecules across the membrane

Ion channels and transporters allow substances to cross the plasma membrane

If the cell membrane was just made up of the phospholipid bilayer and nothing else, then no ions would ever be able to cross the membrane, and no electrical signalling would be possible. However, lots of proteins are embedded in the lipid membrane. Some of these are transporter proteins that can shuttle specific molecules across the membrane (Figure 3.11). For example, glucose is brought into the cell via glucose transporters.

As well as transporters, ion channels are also proteins that are embedded in the plasma membrane (Figure 3.12).

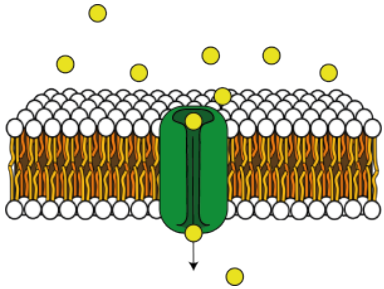

Fig 3.12. Ion channels form a pore in the membrane through which certain ions can pass to cross the membrane.

These proteins form a pore in their centre which essentially makes a hole in the membrane. They can be open all the time (leak channels) or opened by different triggers, such as voltage changes (voltage-gated ion channels) or binding of different molecules (ligand-gated ion channels). Many of these ion channels are selective, i.e. they only let certain ions through. Examples of selective ion channels include potassium leak channels or voltage-gated sodium, potassium or calcium channels. This ion selectivity means that cells can control ion fluxes across their membranes by opening certain ion channels.

The resting membrane potential

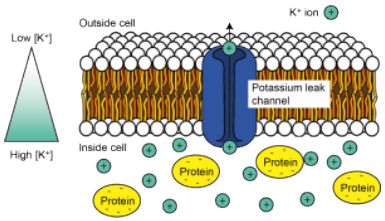

Fig 3.13. Potassium ions are at a higher concentration inside the cell, but some can move out of the cell, down their concentration gradient, through potassium leak channels.

At rest, in the absence of any neuronal signalling activity, it turns out a certain type of ion channel – **potassium leak channels** – are open. This means that, at rest, potassium ions (K^+) can leak out of the cell. Because of the K^+ concentration gradient across the cell – i.e. because there are more K^+ ions inside the cell – as they wiggle and jiggle and randomly move about, some ions will find these holes in the membrane and pass through them to exit the cell (Figure 3.13).

Once some positively charged K^+ ions have left the cell, however, that leaves an imbalance of positive and negative charges on the inside of the cell. The inside of the cell is now more negatively charged compared to the outside of the cell. There is now a voltage, or potential difference across the cell (Figure 3.14), which we could think of as an electrical gradient.

But K+ ions are positively charged, so they are attracted to negative charges and repelled by positive charges. So once there is a potential difference or electrical gradient across the cell's membrane, the K^+ ions are repelled by the positive charge outside of the cell, and attracted to the negative charge inside of the cell. For K^+ ions, the electrical gradient therefore works in the opposite direction to the concentration gradient. The concentration gradient of K^+, with high concentrations inside the cell and low concentrations outside the cell, tends to make K^+ ions leave the cell, while the electrical gradient tends to make K^+ ions enter the cell.

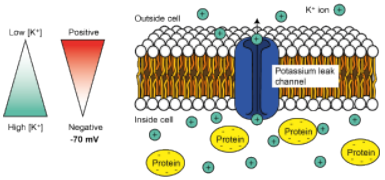

Fig 3.14. Positively charged potassium leaving the cell sets up an electrochemical gradient. The inside of the cell is negative with respect to the outside of the cell, stopping more potassium ions from leaving.

We call the combination of the effect of the electrical and the concentration gradient an **electrochemical gradient**. This movement of K^+ ions out of the cell through leak channels is the main driver of the **resting membrane potential** of the cell – the voltage difference across its membrane at rest – which is around -70 mV in neurons.

Equilibrium potentials

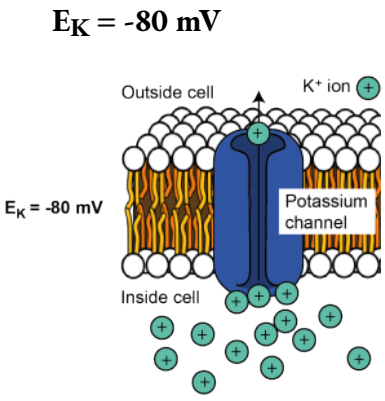

Fig 3.15. Equilibrium potential for potassium

K^+ ions will leave the cell down the concentration gradient until the electrical gradient is so negative that K+ ions are stopped from leaving. At this point K^+ is in equilibrium – the number of ions leaving because of the concentration gradient is the same as the number entering due to the electrical gradient, so there is no net movement of K^+ across the membrane. The voltage difference across the cell at which this equilibrium is reached is called the **equilibrium potential** for a given ion. It is dictated by the concentration difference across the membrane and the charge of the ion. We can consider different ions and how their electrochemical gradients shape the equilibrium potential for each.

As we saw above, because K^+ is positively charged and is at a higher concentration inside the cell, it tends to leave the cell when channels permeable to K^+ are opened in the cell membrane. Positive K^+ ions leaving the cell make the cell's membrane potential (the electrical gradient or voltage

difference across the cell's membrane) more negative. The membrane becomes more and more negative until it reaches the equilibrium potential for K^+ when there is no longer any net flux (flow) of K^+. Therefore the equilibrium potential for K^+ is negative. For most cells, it is around -80 mV. This is often written as E_K (or the electrical potential for K^+) = -80 mV.

E_{Na} = +62 mV

There is more sodium (Na^+) outside the cell than inside the cell, so if ion channels that are permeable to Na^+ open in the membrane, sodium will tend to enter down its concentration gradient. Na^+ is positively charged, so initially it is attracted to the negative potential on the inside of the cell. Na^+ entry makes the inside of the cell more positive, though, until enough Na^+ has entered to make the inside of the cell so positive that it repels further Na^+ entry. – i.e. it reaches equilibrium. This happens at around + 62 mV. Therefore the equilibrium potential for Na^+ (E_{Na}; the electrical potential across the cell where there is no net flux of Na^+ ions) is +62 mV.

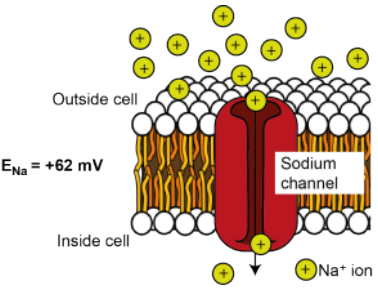

Fig 3.16. Equilibrium potential for sodium

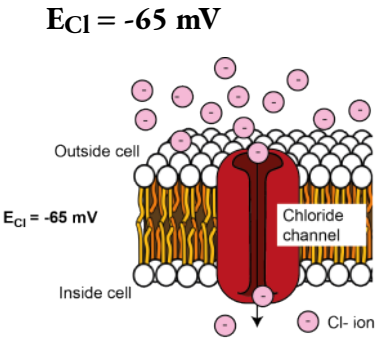

Fig 3.17. Equilibrium potential for chloride

There is more chloride (Cl-) outside the cell than inside the cell, so when ion channels that are permeable to Cl- open in the membrane, Cl- tends to enter the cell. As Cl- is negatively charged, its entry makes the cell's membrane potential more negative, until it reaches equilibrium, being sufficiently negative to repel further Cl- entry. This happens at around -65 mV, so E_{Cl} = -65 mV.

The Nernst Equation

We can mathematically calculate the equilibrium potential for different ions using the **Nernst Equation.**

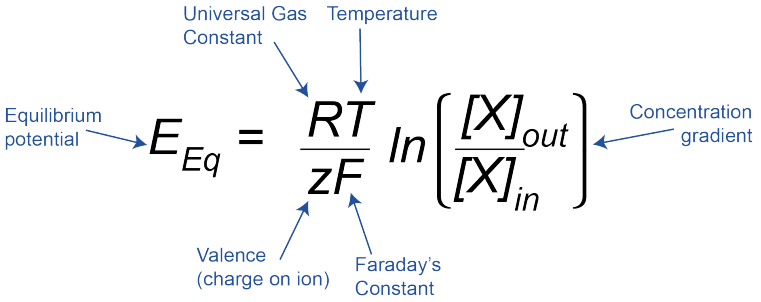

Fig 3.18. The Nernst Equation

This equation might look complicated, but if we break it down we can see that it just relates the concentration gradient across the membrane of an ion X ($[X]_{out}/[X]_{in}$, where [X] is the concentration of the ion of interest) and the charge or valence on that ion (z) to the equilibrium potential (E_{Eq}). R and F are just constants, and T is the temperature, which is constant inside the body, so we can ignore R, F, and T here, as they will always stay the same.

Logarithms and the Nernst Equation

To understand the Nernst Equation fully, we also need to understand what 'ln' means. This is an instruction, which means 'take the natural log of the number inside the brackets'. (In this case, that number is the ratio of the outside and inside concentrations). The log of a number is the power to which a base number has to be raised to equal the original number. The base number can be anything. In the first example below, we are using base 10, but in the case of the natural logarithm (ln) it is a specific mathematical constant called 'e' or **Euler's constant**. It's roughly 2.71828.

Example 1: Logarithms to the base 10:

y is the power to which 10 must be raised to equal x, so y is the log to the base 10 ($\log_{10}$) of x.

$$y = \log_{10}(x)$$

$$10^{y} = x$$

Example 2: Logarithms to the base e (natural logarithms):

In the equations below, y is the power to which e must be raised to equal x, so y is the natural log (ln, also written $\log_e$) of x.

$$y = \ln(x)$$

$$e^y = x$$

More about logarithms and powers

Raising a number to a positive power makes it bigger, whereas raising a number to a negative power makes it smaller (x^{-1} means the same as 1/x; x^{-2} means the same as $1/x^2$). Conversely, the log of a number between 0 and 1 is negative, and the log of a number over 1 is positive. For example (using 10 as the base, not e, to make the sums clearer):

$$10^2 = 100;\ \log_{10}(100) = 2$$

$$10^{-2} = 1/100 = 0.01;\ \log_{10}(0.01) = -2$$

You can find more background on powers and logarithms here. We can now look at K^+, Na^+ and Cl^- and see how their equilibrium constants come out of this equation, even just by broadly considering the

charge of the ion and whether there is more of an ion on the inside or outside of the cell.

E_K = -80 mV

There is more K^+ inside the cell than outside the cell, so $([K^+]_{out}/[K^+]_{in}) < 1$. The natural log of <1 is negative. We then need to multiply this by the charge, which is +1 for K^+. Therefore the equilibrium potential, E_K, is negative.

E_{Na} = +62 mV

There is more Na^+ outside the cell than inside the cell, so $([Na^+]_{out}/[Na^+]_{in}) >1$. The natural log of numbers greater than 1 is positive, and this is multiplied by the charge of +1 for Na^+. Therefore the equilibrium potential, E_{Na}, is positive.

E_{Cl} = -65 mV

There is more Cl^- outside the cell than inside the cell, so $([Cl^-]_{out}/[Cl^-]_{in}) >1$. The natural log of numbers greater than 1 is positive, but this is then multiplied by the charge of -1 for Cl^-. The equilibrium potential, E_{Cl}, is therefore negative.

The membrane potential

We have discussed that when the cell is at rest, potassium leak channels are open and this drives the resting membrane potential to be negative, at around -70 mV. But we also just saw the equilibrium potential for potassium is -80 mV. If the resting membrane potential is set by potassium flux through leak channels, why is the resting membrane potential not the same as E_K? The answer is that at rest the membrane is actually also a tiny bit permeable to Na^+ as a small number of sodium channels are also open. This pushes the resting membrane potential a tiny bit away from the equilibrium potential for potassium towards the equilibrium potential for sodium. The resting membrane potential is closest to E_K as the membrane is most permeable to K^+ ions (more K^+ channels are open), but is a bit more positive than E_K because of the small amount of permeability at rest to Na^+ ions.

In fact, at any point during neuronal signalling or at rest, the membrane potential is set by the electrochemical gradients to different ions and the relative permeability of the membrane to these ions. Cells control their membrane potentials by opening and closing ion channels in the membrane to alter the permeability to different ions, which then flow down their electrochemical gradients into or out of the cell. When sodium channels open, for example, the permeability to Na^+ increases and Na^+ ions enter the cell, driving the membrane potential to more positive potentials towards the equilibrium potential

for Na^+, E_{Na}. When sodium channels close, the permeability to Na+ decreases again and, as the membrane is now more permeable to K^+ than Na^+, the membrane potential will again become more negative, returning to the resting membrane potential. There are lots of types of ion channels that are selective for different ions and have different gating properties, i.e. they are opened and closed by different stimuli, such as changes in membrane voltage, and binding of specific molecules. We'll discuss these more in later chapters.

The Goldman-Hodgkin-Katz equation

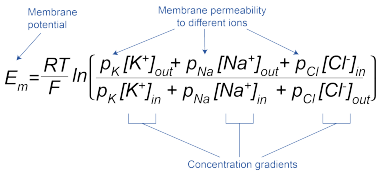

Fig 3.19. The Goldman-Hodgkin-Katz equation

The Goldman-Hodgkin-Katz equation allows the membrane potential of the cell (E_m) to be calculated from the permeabilities of the membrane to different ions (pK, pNa and pCl for K^+, Na^+ and Cl^-, respectively) and their concentration gradients (Figure 3.19). As for the Nernst equation (see Box, Figure 3.18), R and F are constants, and T is temperature, so RT/F can be considered unchanging. During neuronal signalling, the permeability of the membrane to different ions changes, and the membrane potential is

weighted in favour of the equilibrium potential of the ion with the greatest permeability at that moment. Note that the Cl^- concentration gradient is expressed in reverse compared to K^+ and Na^+, to account for the fact that, being negatively charged, it is oppositely charged to K^+ and Na^+.

The sodium-potassium ATPase

We have seen above that during rest and neuronal signalling, ions flow through ion channels down their electrochemical gradients, altering the membrane potential. This flow of ions down their electrochemical gradients does not require any energy. The membrane potential is controlled by changing the permeability of the membrane to different ions, and not by changing the concentration gradients between the inside and outside of the cell. Very few ions need to flow to change the membrane potential of a cell, which means that the concentrations of ions inside and outside the cell do not change very much over the short term. However, because the membrane potential does not sit at the equilibrium potential for any ion, even at rest, there is a net K^+ current or flux out of the cell and, a net Na^+ flux into the cell.

Over the longer term, however, these fluxes would dissipate the ionic concentration gradients if cells did not have a mechanism to continually pump ions back to where they came from. The pump that does this really important job is the sodium-potassium pump, or the Na^+/K^+ ATPase. This is a

protein that sits in the plasma membrane and pumps sodium out of the cell and potassium back into the cell. Because this pumping occurs against the ions' electrochemical gradients, it requires energy in the form of ATP to pump the ions back and maintain their concentration gradients. The sodium-potassium ATPase removes a phosphate group from ATP, to form ADP, releasing some energy, which changes the shape (or conformation) of the Na^+/K^+ ATPase enabling it to move 3 Na^+ ions out of the cell and 2 K^+ ions into the cell for every ATP molecule used. Because 3 Na+ ions are removed for every 2 K^+ ions brought into the cell, the Na^+/K^+ ATPase is **electrogenic**, causing a net export of positive charge. This contributes a little bit to the negative resting membrane potential, but by far the strongest effect the Na^+/K^+ ATPase has on the resting membrane potential is to maintain the potassium electrochemical gradient, so that the equilibrium potential for potassium is maintained. Because even at rest there are ion fluxes, the Na^+/K^+ ATPase is always at work, but it's activity is increased when neurons are signalling and so more ions need to be pumped back.

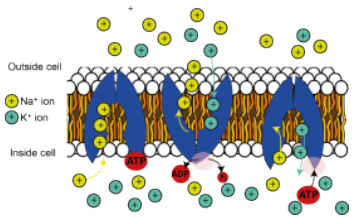

Figure 3.20.
Sodium-potassium ATPase. Across one cycle of activity, 3 Na+ ions are transported out of the cell, 1 ATP is hydrolysed to ADP and phosphate (Pi) and 2 K+ ions are transported into the cell.

Maintaining ion concentration gradients is so important for sustaining neuronal activity that the Na^+/K^+ ATPase is the single most energy-consuming process in the brain, consuming over half of all the energy it uses. As the brain is a very energetically expensive organ, using 20% of the body's energy at rest, despite comprising only 2% of the body's mass, the Na^+/K^+ ATPase alone uses over 10% of the energy used by the whole body – quite staggering given there are over 20,000 different types of proteins in our bodies at any one time!

Key Takeaways

- Electrical signalling in neurons (and other cells) works because they have ion channels that allow specific ions to flow across neuronal membranes and change the membrane potential of the cell.
- The membrane potential of the cell is determined by the concentration gradient of ions across its membrane, and the permeability of its membrane to those ions.
- Ions flow down their electrochemical gradients, which doesn't need any energy, but energy in the form of ATP must be used up to fuel the Na^+/K^+ ATPase which pumps ions back up their electrochemical gradients to maintain their concentration gradients across the membrane.

About the author

Dr Catherine Hall
UNIVERSITY OF SUSSEX
https://twitter.com/cathnaledi

Dr Catherine Hall is a member of the Sussex Neuroscience Steering Committee, the University Senate, convenes the core first year module *Psychobiology,* and lectures on topics relating to basic neuroscience, neurovascular function and dementia.

5.

NEURONAL TRANSMISSION

Dr Catherine N. Hall

Learning Objectives

By the end of this chapter, you will understand:

- that neurons signal electrically within each cell and chemically between cells
- the ionic basis of the action potential and how it is conducted
- the processes involved in synaptic transmission
- how neurons integrate information at synapses.

In the last chapter, we learnt about electrical signalling in the brain and how electrochemical gradients and ion channels allow neurons to set their membrane potential. In this chapter we will learn how these processes generate the signals within and between neurons that form the basis for the information processing in the brain.

Signals are transmitted electrically within neurons and chemically between neurons, at synapses. Electrical signals within neurons take the form of action potentials and synaptic potentials. We can talk of electrical signals in cells as producing a positive change in the membrane potential – termed **depolarisation**, or a negative change in the membrane potential, termed **hyperpolarisation**.

Action potentials

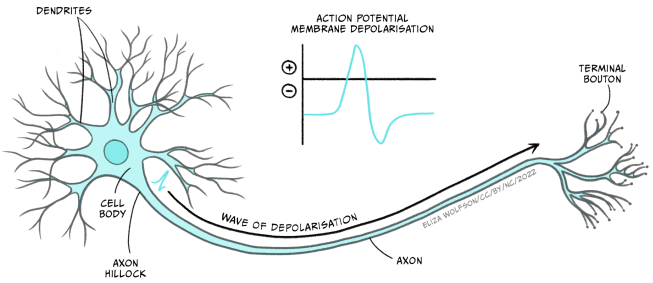

Fig 3.21. An action potential is a transient voltage change that spreads from the axon hillock to the axon terminals.

An action potential is a brief electrical signal that is conducted from the axon hillock where the neuron's soma joins the axon, along the axon to the axon terminals. It can be measured from electrodes placed in or near a neuron connected to a voltmeter (Figure 3.21). This electrical signal is a rapid, localised change in the membrane voltage which transiently changes from the negative resting membrane potential to a positive membrane potential. A positive shift in the membrane potential like this is termed **depolarisation**. The membrane then rapidly (within 1 ms) becomes negative again – it **repolarises** – and then shifts even more negative, becoming **hyperpolarised** before returning to the resting membrane potential less than 5 ms after it first depolarised (Figure 3.22). This transient voltage

change then spreads like a wave down the axon with a conduction velocity of between 1 and 100 m/s.

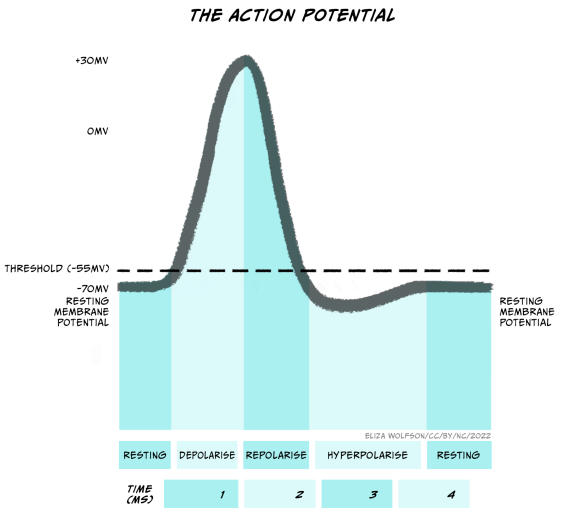

Fig 3.22. Membrane potential changes during an action potential.

The action potential is caused by opening and closing of voltage-gated ion channels

What is happening within the axon to cause these changes in membrane voltage?

As discussed above, the way in which neurons generally alter

their membrane potentials is by changing their membrane permeability to different ions by opening and closing ion channels, and that is exactly what is happening during the action potential. The ion channels that open and close to form the action potential are **voltage-gated ion channels**. As their name suggests, these channels open or close depending on the voltage across the membrane. There are many different types of voltage-gated ion channels, which differ in their thresholds for activation – the voltages at which they open and close – as well as their selectivity for ions. When they open, ions flow down their electrochemical gradients towards their equilibrium potentials.

Voltage-gated sodium and potassium channels open to depolarise then hyperpolarise the membrane

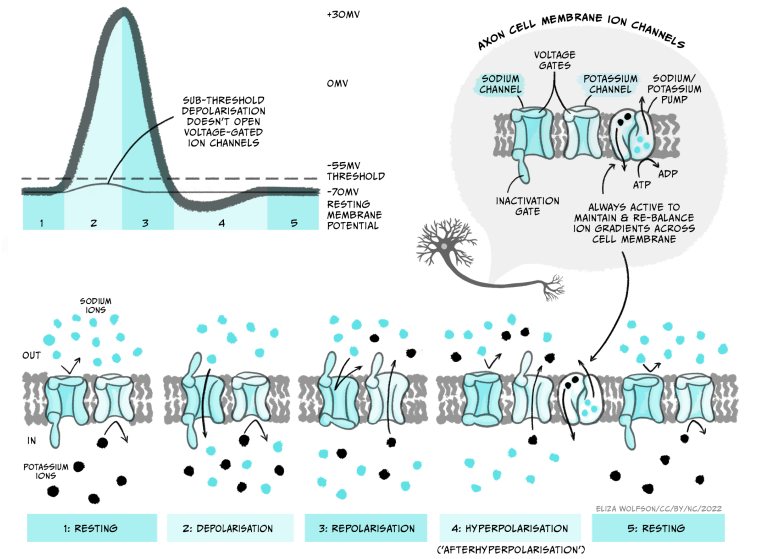

Fig 3.23 The action potential is caused by the opening and closing of voltage gated ion channels.

The upstroke (when the voltage depolarises rapidly) of the action potential (Figure 3.23) is caused by the opening of voltage-gated sodium channels that have a threshold for opening of -55 mV. When the membrane of the neuron depolarises to -55 mV, these voltage-gated sodium channels start to open. Sodium ions flood into the cell, depolarising the membrane and opening even more sodium channels, causing a very rapid depolarisation of the membrane. This feedforward activation of sodium channels makes the action potential an **all or nothing event** (it either happens, or it does not). If the threshold is reached, sodium channels open, accelerating

depolarisation happens and an action potential occurs (or 'fires'). If the threshold is not reached, sodium channels do not open and no action potential will fire. Furthermore, the action potential is always the same size and is not graded by the size of the incoming depolarisation.

If the sodium channels stayed open, then the membrane potential would stabilise at the equilibrium potential for sodium (E_{Na}), at +62 mV, but instead the voltage reaches only around +40 mV before hyperpolarising again, so the membrane is depolarised for less than 1 ms. The depolarisation is so brief for two reasons: firstly, the voltage-gated sodium channels rapidly inactivate, closing the channel and preventing further Na+ influx to the cell. Secondly, a second type of voltage-gated channel activates: the voltage-gated potassium channel. Some of these voltage-gated potassium channels activate at the same threshold as the sodium channels but more slowly, and others activate at a more positive voltage (around +30 mV). Both these factors mean that opening of voltage-gated potassium channels is delayed relative to the Na^+ influx. When channels open, however, K^+ leaves the cell, causing the membrane to become more negative, or hyperpolarised, producing the falling phase or downward stroke of the action potential (Fig. 3.23).

Increased K+ permeability causes an afterhyperpolarisation

Many voltage-gated potassium channels switch off quite slowly after the membrane potential falls below their threshold voltage. This means that after the membrane potential has repolarised, reaching the resting membrane potential, there are still some voltage-gated potassium channels open, in addition to the potassium leak channels that are always open. Because the membrane is now more permeable to K^+ than at rest, the membrane potential hyperpolarises below the resting membrane potential, getting even nearer to the equilibrium potential for K^+, E_K. This hyperpolarised phase is termed the **afterhyperpolarisation**. Then as the voltage-gated potassium channels close, the permeability of the membrane for potassium returns to normal and the membrane potential depolarises slightly back to the resting membrane potential.

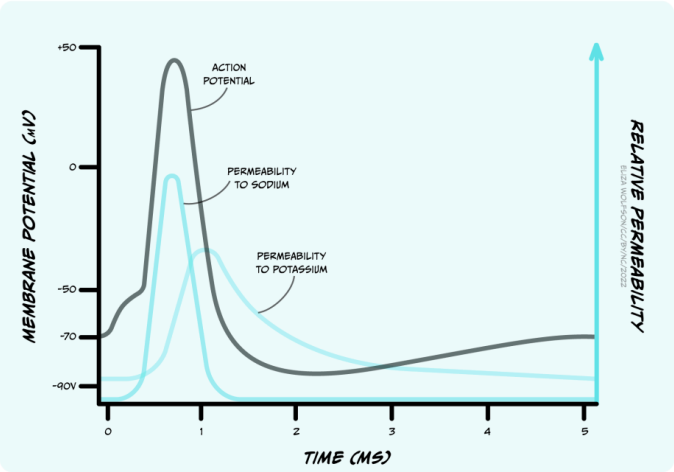

Fig 3.24. Permeability of the membrane to sodium and potassium during the action potential

Sodium channel inactivation causes the refractory period for action potential firing

The opening and closing of voltage-gated sodium and potassium channels at different threshold voltages and inactivation of sodium channels occur because gates in the proteins move to open and close the pore region in the centre of the channel that allows ions to flow across the membrane (Figure 3.24). At the resting membrane potential, voltage-gated sodium and potassium channels both have a

conformation or shape that means part of the protein blocks the ion channel's pore (i.e. it is like there is a closed gate blocking the pore). When the threshold voltage is reached, the shape of the ion channel proteins change slightly so that this gate opens to let ions through. This gate opens quickly in voltage-gated sodium channels but more slowly, or at more depolarised potentials in voltage-gated potassium channels, so during the rising phase of the action potential only the sodium channel gates are open. After a very short time, however, an inactivation gate on the intracellular side of the voltage-gated sodium channel swings shut, blocking the pore from the inside and stopping any more Na^+ flux . As the voltage-gated potassium channels open, during the falling phase of the action potential, voltage gated sodium channels are inactivated. Even when the membrane falls below the threshold voltage, closing the voltage-sensitive gate, the sodium channels' inactivation gates are still closed. This means that the sodium channels cannot re-open, and the neuron cannot fire another action potential until the inactivation gates reopen.

This period of time when firing of another action potential is impossible is called the **absolute refractory period** (Figure 3.25). Sodium channels' inactivation gates start to reopen during the falling phase of the action potential, when voltage-gated potassium channels are still open. At this stage, it becomes possible to fire another action potential, but a stronger stimulus is needed to activate the sodium channels.

This period is the **relative refractory period** (Figure 3.25). Stronger stimuli (that depolarise a neuron more) can therefore produce a faster firing rate in a target neuron than weaker stimuli by intruding into the relative refractory period.

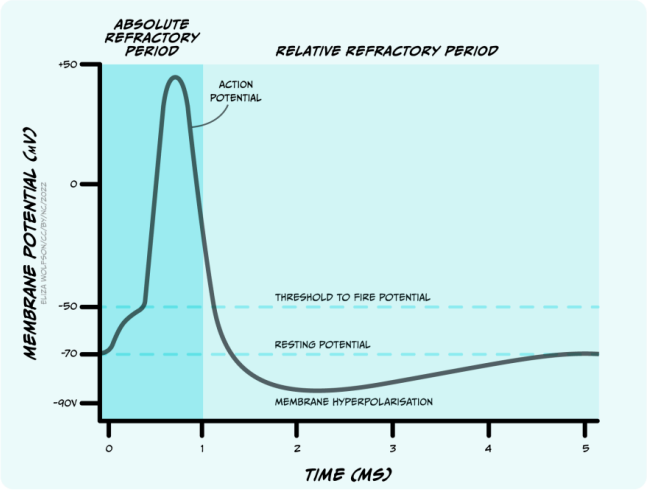

Fig 3.25. Absolute and relative refractory periods

Action potential propagation

Action potentials are initiated in the axon's initial segment near the soma, right next to the axon hillock. If the membrane potential there depolarises sufficiently to trigger voltage-gated sodium channels to open, then an action potential will fire in

that section of membrane. In an unmyelinated axon (Figure 3.26), some of the positive charge (Na+ ions) that enters the cell during the rising phase of the action potential spreads to the adjacent bit of membrane, depolarising that membrane and opening voltage-gated sodium channels there, producing an action potential, which spreads onwards to the next bit of membrane, such that a wave of depolarisation and repolarisation spreads down the axon all the way to the axon terminals. Sodium channel inactivation prevents upstream spread of the action potential back towards the soma: because the upstream membrane is in the absolute refractory period, the action potential can only spread downstream to membrane in which sodium channels are not inactivated.

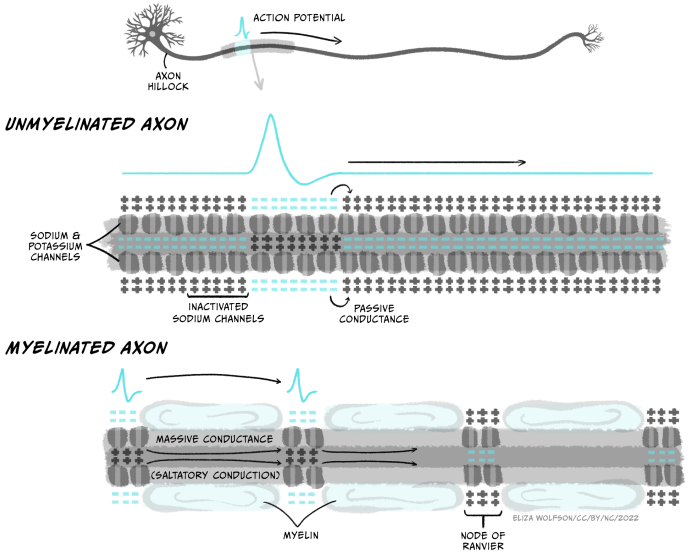

Fig 3.26. Action potential propagation in myelinated and unmyelinated axons

Increasing the axon diameter and myelinating the axon increases conduction speed

Action potentials spread quite slowly along small unmyelinated axons – around 0.5-2 m/s – because each bit of membrane has to fire an action potential and propagate it to the next bit of membrane. This speed of conduction would be too slow to get up-to-date information about what is going on

at the far end of our bodies – imagine wiggling your toe and only knowing 4 seconds later that you actually had wiggled it! Luckily, action potential conduction can be increased in two major ways.

Firstly, conduction speed is increased by increasing the diameter of the axon, which reduces the resistance to current flow within the axon, allowing depolarisation to passively spread further down the axon and therefore more rapidly activate action potential firing in downstream membrane.

Secondly, myelination of axons increases conduction speed. The layers of myelin that are tightly wrapped around axons by oligodendrocytes (in the CNS) or Schwann cells (in the PNS) insulate the axon membrane from current loss across the membrane. Axon membrane ensheathed in myelin layers does not contain ion channels – it has low permeability and high resistance to current flow. This allows current to spread further inside the axon without leaking out of the cell, allowing current to spread further down the axon without being dissipated. The myelin sheath also decreases the membrane **capacitance** – the amount of charge stored at the membrane. Charge gets stored at the membrane when positive and negative charges are attracted to each other across the thin plasma membrane, holding them near each other at opposite sides of the membrane. By wrapping tightly around the membrane, myelin increases the distance between the intracellular and extracellular fluids containing charged particles so they are less attracted to each other across the

ensheathed membrane. The lowered capacitance allows current to spread further (and faster) inside the axon as ions do not get stuck at the membrane.

The result of myelination is that depolarisation can rapidly spread passively along relatively long distances of axon, but it cannot spread down the whole length of the axon. The signal still needs to be boosted periodically by generating a new action potential. This happens at nodes of Ranvier, which are gaps in the myelin sheath that are packed with ion channels. When the nodes of Ranvier depolarise, their voltage-gated sodium channels open, triggering a new action potential which can then passively spread across the ensheathed internode region of the axon to the next node of Ranvier (Figure 3.26). Because the action potential rapidly jumps between nodes, this form of conduction is called **saltatory conduction** (from Latin 'saltare' – 'to jump'). Large diameter, myelinated axons can conduct action potentials at speeds up to 100 m/s, meaning information about toe-wiggling can reach your brain in a respectable 0.02 s. Indeed, sensory neurons carrying information about where our bodies are in space have some of the fastest propagating axons of any cell.

Demyelinating conditions, such as multiple sclerosis and Guillain-Barré Syndrome, cause a multitude of symptoms, including altered sensation, muscle weakness and cognitive impairments, due to loss of myelin sheaths, disrupted neuronal communication and eventual axonal degeneration.

Energy use by action potentials

The flow of ions through voltage-gated ion channels during the action potential occurs down their electrochemical gradients so it does not itself use energy. During an action potential very few ions actually flow so the concentration gradients do not change significantly over the short term. Over the longer term, however these ions need to be pumped back to maintain concentration gradients and the resting membrane potential so that further action potentials can fire. This is achieved by the Na+/K+ ATPase, using ATP. Myelination of axons helps speed action potential conduction, but also makes action potential firing more energy efficient, because fewer ions need to flow to depolarise the myelinated membrane. Fewer ions therefore need to be pumped back across the membrane, so less ATP is needed by the Na+/K+ ATPase.

Action potential: Key takeaways

- When the membrane reaches a threshold voltage, voltage-gated sodium channels briefly open, depolarising the cell

- Voltage-gated potassium channels open and repolarise the cell
- Depolarisation spreads along the membrane activating nearby sodium channels
- Inactivation of sodium channels means action potential propagates in one direction and sets a limit on firing frequency
- Action potentials are all-or-nothing, and only occur once the threshold for sodium channel activation is met
- Myelination speeds action potentials and makes them more energy-efficient.

Communication between neurons

Neurons signal electrically, using action potentials to communicate between the soma and the axon terminals. The action potential signals that the soma and axon's initial segment depolarised to the threshold voltage. But what generates that depolarisation in the first place? What is the signal that depolarises a neuron to make it fire an action

potential? We saw in chapter 3 that neurons integrate lots of inputs and compute whether or not to fire an action potential. In sensory neurons, these inputs to a neuron might be information from the outside (or internal) world, for example stretch of the skin, a painful heat, or a delicious smell. You will learn more about how these types of stimuli generate inputs in neurons in later chapters. But for most neurons, the inputs come from other neurons, via connections, or synapses. While neurons communicate electrically within a cell, communication between neurons is usually chemical – a chemical or neurotransmitter is released from one neuron and acts to generate a signal on the next neuron.

Synaptic transmission

During synaptic transmission, an action potential in a neuron – the **presynaptic neuron** – causes a neurotransmitter to be released into a tiny gap called the **synaptic cleft** between two neurons. The neurotransmitter diffuses across the synaptic cleft and binds to receptors on the neuron receiving the signal – the **postsynaptic neuron**, which produces a change in the postsynaptic cell. Looking into this process in more detail, we can split the processes of synaptic transmission into a number of separate steps (Figure 3.27):

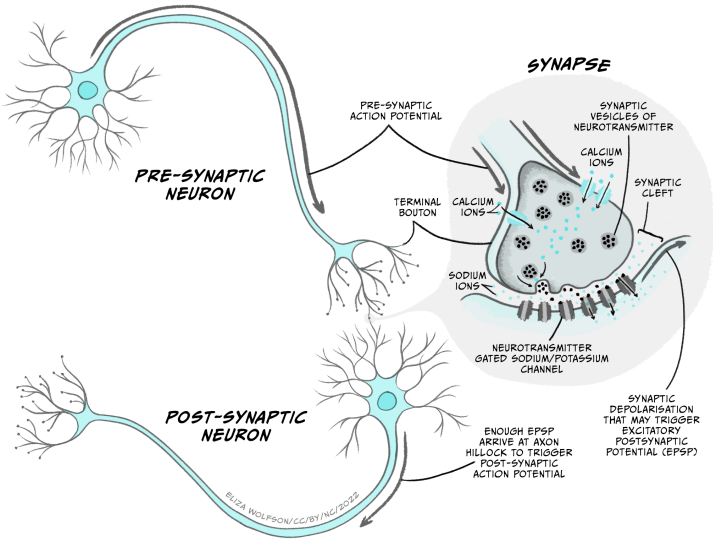

Fig 3.27. Synaptic transmission

- An action potential arrives at the axon terminal (or **presynaptic terminal**), depolarising it.
- Depolarisation of the presynaptic terminal opens a new type of voltage-gated ion channel – the **voltage-gated calcium channel**, which has a threshold for activation of around -10 mV. When these channels open, calcium (Ca^{2+}) enters the cell down its electrochemical gradient, as there is a higher concentration of Ca^{2+} in the extracellular fluid compared to the intracellular fluid (1.5-2 mM outside the cell, vs. 0.05 – 0.1 mM inside the cell), and its positive charge attracts it into the negatively charged cell. Unlike Na^{+} and K^{+}, Ca^{2+} is not present at

high enough concentrations to affect the membrane potential of the cell. Instead, an increase in intracellular Ca^{2+} concentration can trigger different signalling cascades in the cell, by binding to different proteins.

- Ca^{2+} entering through voltage-gated calcium channels binds to a protein called **synaptotagmin**.
- The presynaptic terminal contains lots of little membrane 'bags' called **synaptic vesicles**, which are packed with neurotransmitter. Some of these vesicles are close to an area on the plasma membrane of the cell called the '**active zone', ** whereas the vesicles that are still being packed with neurotransmitter are further away from the membrane and nearer the centre of the presynaptic terminal. The vesicles at the active zone are 'docked', being held close to the plasma membrane by a complex of proteins called **SNARE proteins**. When calcium binds to synaptotagmin, the membranes of the vesicle and the plasma membrane of the cell are brought even closer together and fuse, releasing the contents of the vesicle (neurotransmitter molecules) into the extracellular space of the synaptic cleft. The vesicles that are already docked at the active zone are more readily released so are the first to fuse with the membrane and release their neurotransmitter.
- The synaptic cleft is very narrow, so neurotransmitter molecules can quickly diffuse across from the presynaptic terminal to the post-synaptic cell.

- The postsynaptic cell's membrane (usually part of a dendrite) contains **receptors** for the neurotransmitter molecules that are released from the presynaptic cell. A receptor is a protein that can bind a specific molecule – termed a **ligand**. Many of these receptors are part of **ligand-gated ion channels**. These are ion channels that open when a specific molecule binds to them. Ions flow through the open ion channels, down their electrochemical gradients, producing a change in the membrane voltage in the post-synaptic cell.
- To terminate synaptic signalling, neurotransmitter must be removed from the synaptic cleft. This is achieved by **transporters** on neurons or astrocytes, proteins which take up neurotransmitter into the cell where it can be broken down, recycled or repackaged. Some neurotransmitters may also be broken down by proteins that are present in the synaptic cleft.

Excitatory synapses

Excitatory synapses make the post-synaptic neuron more likely to fire an action potential by producing a depolarisation in the post-synaptic cell, moving it towards the threshold potential for opening voltage-gated sodium channels. This happens when Na+ ions are allowed to flow into the cell.

The main excitatory neurotransmitter in the brain is **glutamate** (acetylcholine is an important excitatory

neurotransmitter in the peripheral nervous system). Glutamate's main receptors are **AMPA and NMDA receptors**. AMPA receptors are ligand-gated ion channels that let both Na^+ and K^+ pass through them. Though K^+ ions leave the cell when AMPA receptors open, the main effect is an influx of Na^+, so when glutamate binds AMPA receptors, the membrane depolarises towards the threshold for firing an action potential. This depolarising change in membrane potential is termed an **excitatory post-synaptic potential (EPSP;** Figure 3.28) and lasts several (> 10) milliseconds.

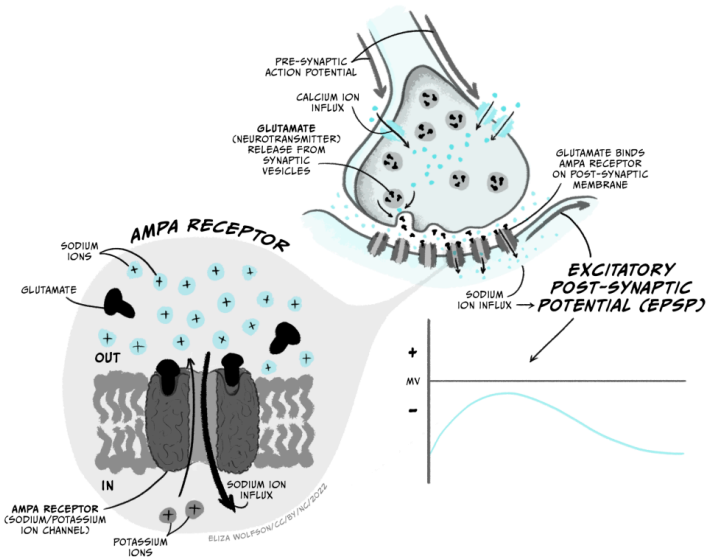

Fig 3.28. Glutamate binding to AMPA receptors causes EPSPs

NMDA receptors are also ligand-gated ion channels and are permeable to Ca^{2+} as well as Na^{+} and K^{+}. However they are also voltage-dependent, as they are blocked by Mg^{2+} ions unless the membrane potential is depolarised. They are also slower than AMPA receptors to open and close. Because of this they do not contribute much to the EPSP. However they play a really important role in altering **synaptic strength** – or how much of an effect a presynaptic action potential can have on the postsynaptic cell.

Metabotropic glutamate receptors are often also present. Metabotropic receptors are also known as **G-protein coupled receptors**. These proteins bind glutamate but do not directly open an ion channel. Instead they trigger other intracellular signalling pathways that can make other changes to the cell, for example altering the properties of other ion channels. Because their action is via intracellular signalling pathways, they have slower effects than ionotropic receptors (receptors such as AMPA and NMDA receptors that are part of, and directly activate, ion channels).

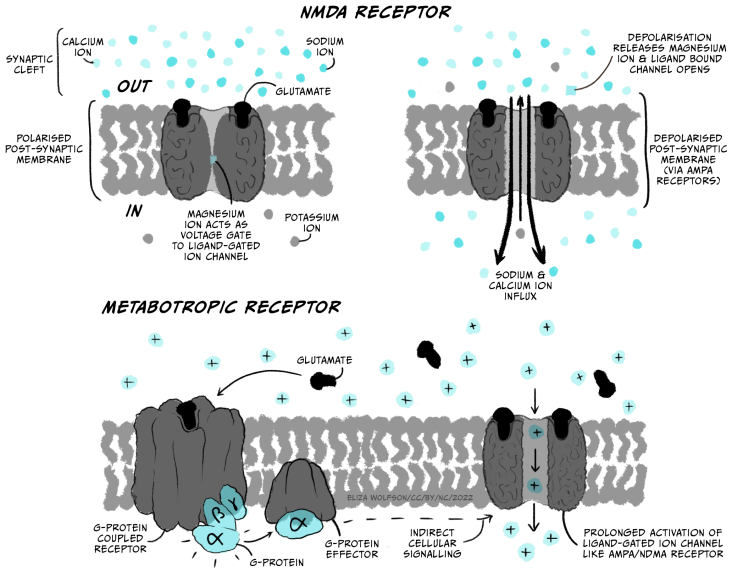

Fig 3.29. NMDA and metabotropic glutamate receptors

Usually an EPSP from a single synapse won't depolarise the post-synaptic neuron enough to reach the threshold for firing an action potential. Instead multiple synaptic inputs need to be summed together to get a big enough EPSP (Figure 3.30). If the presynaptic neuron fires lots of action potentials in a short space of time, then the inputs into a single synapse can add together to form a larger EPSP. This is **temporal summation**. Additionally, if different excitatory synapses are active at the same time, then these EPSPs can **spatially summate** to generate a larger EPSP. Both temporal and spatial summation happen to integrate the inputs onto a postsynaptic cell, to determine whether it fires an action potential.

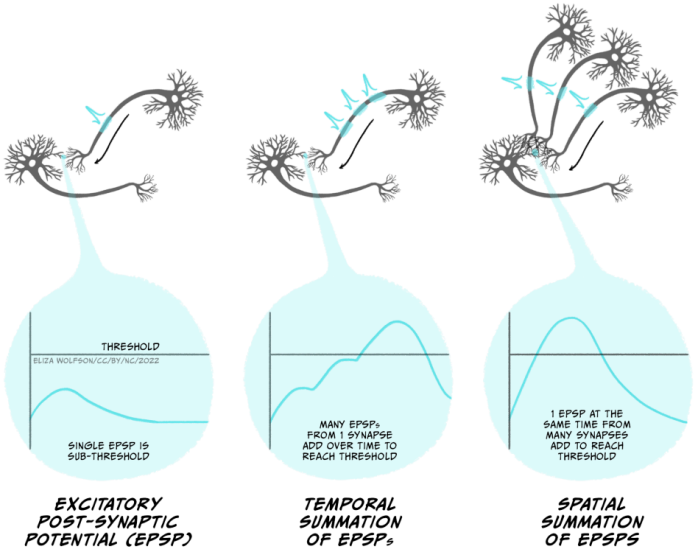

Fig. 3.30. Summation of EPSPs

Inhibitory synapses

Inhibitory synapses make the post-synaptic neuron less likely to fire an action potential, by hyperpolarising the membrane, or by preventing it from depolarising by holding the membrane below that needed to activate sodium channels.

The main inhibitory neurotransmitter in the brain is **GABA** (gamma aminobutyric acid), whose main receptors are **$GABA_A$** and **$GABA_B$ receptors**. $GABA_A$ receptors are ligand-gated ion channels that are permeable to Cl^- ions when

GABA is bound. Because Cl^- ions enter the cell on activation and the equilibrium potential for Cl^- (E_{Cl}) is -65 mV, opening $GABA_A$ channels will tend to keep the membrane potential near -65 mV. As this is below the threshold for activation of sodium channels, this will inhibit the post-synaptic neuron from firing an action potential. Depending on the membrane voltage of the cell when these channels open, the membrane potential might slightly hyperpolarise or depolarise the cell. In each case, however, this membrane potential change is inhibitory (**an inhibitory post-synaptic potential or IPSP**) because it is holding the membrane potential away from that needed to fire an action potential. For example, if the neuron's membrane potential is -75 mV when $GABA_A$ receptors open, the membrane potential will move towards E_{Cl} so will depolarise slightly to -65 mV. However the open $GABA_A$ receptors prevent the membrane from depolarising beyond -65 mV to the threshold for firing an action potential. If the membrane potential is more positive than E_{Cl}, e.g. -60 mV, then opening $GABA_A$ channels will make the membrane potential more negative or hyperpolarised, until it reaches -65 mV. In both cases, opening the $GABA_A$ channels has made the neuron less likely to reach threshold for action potential firing.

$GABA_B$ receptors are metabotropic receptors that are linked to activation of potassium channels, increasing K^+ permeability. Their activation therefore shifts the membrane potential towards E_K, or -80 mV, hyperpolarising the cell.

$GABA_B$-mediated membrane potential changes are therefore also IPSPs as they hyperpolarise the membrane away from the threshold for action potential firing, but because they require intracellular signalling these IPSPs are slower than $GABA_A$-mediated membrane potential changes.

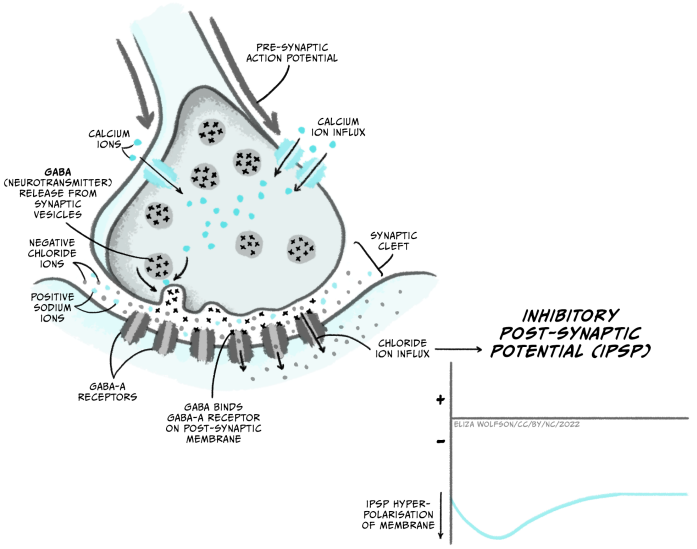

Fig 3.31. Inhibitory synapse

Synaptic integration

Postsynaptic cells use temporal and spatial summation to integrate all the different synaptic inputs to the cell. If the net effect of all the inputs is to depolarise the axon initial segment above the threshold for activating sodium channels, the cell

will fire an action potential. The way in which all these inputs are integrated to generate an output (action potential) is therefore the basis of how neurons perform the computations on which our thoughts and feelings depend.

Neurons can perform different computations based on their morphology and the spatial organisation of their excitatory and inhibitory inputs, as this alters how they are summated (Fig. 3.32). Most synaptic inputs are onto the dendrites of a neuron, but some may be onto the soma or even the axon. Synaptic inputs to the distal end of dendrites (far from the soma) will potentially have a smaller effect on the membrane potential at the axon initial segment than an input onto the soma, because the signal degrades over the distance they need to travel, while inputs onto the axon initial segment itself can have an even stronger effect than those onto the soma. Excitatory inputs onto distal dendrites can also be gated by inhibitory synapses that are more proximal to the soma on the same dendrite, so the EPSP cannot reach the soma. The ability of EPSPs and IPSPs to spread along dendrites is also determined by factors such as the number and type of ion channels in the dendritic membranes, as well as the size of the cell. If there are few ion channels, then charge cannot as easily leak out across the membrane and dissipate the potential chance. Similarly, a given input will spread further in a small cell than a larger, highly branched cell, as less charge gets lost at the membrane (the smaller cell has a lower capacitance).

However, dendrites also express voltage-gated ion channels that can boost signals from distal dendrites.

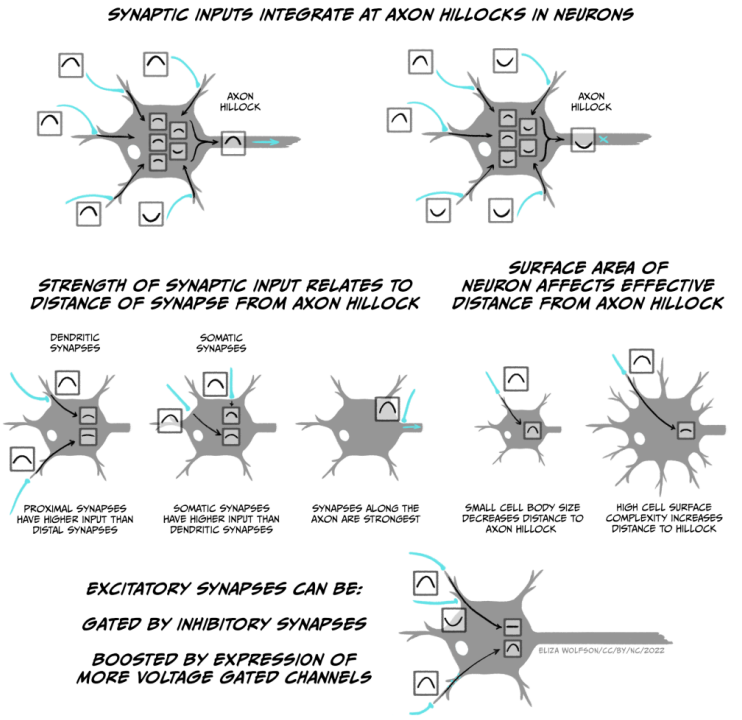

Fig 3.32 Synaptic integration

Neurons' computation can therefore be affected by many factors, from the location and strength of individual synapses, to the shape of the cell and the number and location of ion channels expressed. Many of these properties can be modified based on the cell's activity, allowing alterations to the

contribution that different synaptic connections play on the decision to fire an action potential. This **plasticity** in synaptic connectivity is critical for allowing associations to be formed and broken between neurons, forming the basis of learning and memory as well as shaping how we perceive the world.

Gap Junctions

While most connections between neurons are via chemical synapses, direct electrical connections also occur. These are called **gap junctions** and are formed by pairs of hemichannels, one on each cell, made up of a complex of proteins called connexins. Compared to other ion channels, gap junctions are relatively non-selective, allowing cations (positively-charged ions) and anions (negatively-charged ions) through as well as small molecules such as ATP. Though regulation of their opening is

possibly, they are usually open, meaning that electrical signals can spread through connected cells. Gap junctions are more common during development and are rare between excitatory cells in mature nervous systems. They are most common between certain inhibitory interneurons in the brain and the retina, as well as between glia, such as astrocytes.

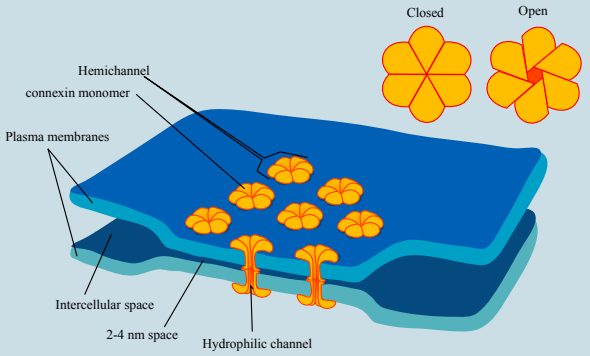

Fig. 3.33. Gap junction coupling between cells

Other neurotransmitters

While glutamate is the main excitatory neurotransmitter in the brain, and GABA is the main inhibitory neurotransmitter in the brain, there are many other neurotransmitters that can also be released at synapses. These can be broadly divided in to different categories, based on the chemical structure of the neurotransmitter molecules. All activate their own specific receptors.

Amino acid neurotransmitters include glutamate, GABA and also glycine, which is the major inhibitory neurotransmitter in the brainstem and spinal cord.

Monoamine neurotransmitters include noradrenaline, dopamine and serotonin. There are specific populations of monoaminergic neurons in the brain that originate in specific midbrain and brainstem nuclei and send projections to widespread brain regions, modulating processes such as reward, attention and alertness. Noradrenaline is also an excitatory transmitter in the peripheral nervous system.

Peptide neurotransmitters include naturally occurring opioid peptides – endorphins, enkephalins and dynorhpins – that activate the same receptors as opiate drugs such as morphine and heroin. There are numerous other peptide neurotransmitters, including oxytocin and somatostatin.

Peptide neurotransmitters are often co-released at synapses with GABA or serotonin.

Purine neurotransmitters include ATP, the cell's main energy currency, and its breakdown product adenosine.

Acetylcholine is unlike other neurotransmitters structurally. It is a common excitatory neurotransmitter in the peripheral nervous system, including at the neuromuscular junction, and is also released by many neurons in the brain, where it is involved in regulating alertness, memory and attention.

Synaptic transmission : Key takeaways

- When an action potential arrives at an axon terminal, voltage-gated calcium channels open, allowing Ca^{2+} influx into the terminal
- Ca^{2+} binds synaptotagmin, pulling synaptic vesicles very close to the plasma membrane. This triggers fusion of synaptic vesicles with the plasma membrane, releasing

neurotransmitter into the synaptic cleft.

- Neurotransmitter diffuses across the synaptic cleft and binds to ionotropic or metabotropic receptors on the postsynaptic cell.
- Receptors for excitatory neurotransmitters such as glutamate trigger Na^+ entry into the postsynaptic cell, depolarising the membrane (producing an EPSP), making it more likely the postsynaptic cell will depolarise to the threshold for firing an action potential.
- Inhibitory neurotransmitters such as GABA activate receptors that keep the membrane potential negative with respect to the threshold for firing an action potential (generating an IPSP).
- Postsynaptic neurons integrate different excitatory and inhibitory inputs to decide whether to fire an action potential.
- The location and strength of different synapses, as well as the shape of the post-synaptic cell and expression of different ion channels modify the integration of different inputs – changing these can alter the computation done by the cell.

About the author

Dr Catherine Hall
UNIVERSITY OF SUSSEX
https://twitter.com/cathnaledi

Dr Catherine Hall is a member of the Sussex Neuroscience Steering Committee, the University Senate, convenes the core first year module *Psychobiology,* and lectures on topics relating to basic neuroscience, neurovascular function and dementia.

6.

PSYCHOPHARMACOLOGY: HOW DO DRUGS WORK ON THE BRAIN?

Dr Bryan F. Singer

Learning Objectives

- To gain knowledge and understanding of how drugs enter the body and the time course of their effects
- To gain a basic understanding of how general classes of drugs interact with neurons to alter their function.

Having learnt about how neurons in the brain communicate, let's now consider how drugs can affect their function.

In a fictional example, Sam has both high cholesterol and attention deficit hyperactivity disorder (ADHD). To help alleviate their symptoms, the GP prescribes them atorvastatin to lower their cholesterol, while a psychiatrist prescribes lisdexamfetamine to help improve their attention.

Both atorvastatin and lisdexamfetamine are considered **drugs**. Researchers who design drugs and investigate how they act on the body are often called **pharmacologists** (they study **pharmacology**). While a general pharmacologist might explore the use of atorvastatin or lisdexamfetamine, someone who researches **psychopharmacology** might be more interested in understanding how lisdexamfetamine can reduce symptoms of ADHD.

These scientists don't just develop drugs or observe changes in symptoms after administration; they also ask various other questions! For example, a psychopharmacologist may consider the following:

- What parts of the brain does a drug act on?
- Does a drug have its effect because it interacts with a specific receptor type?
- How does the long-term administration of a drug impact brain biology?
- After a drug is taken, how long do its effects last?
- Can a drug's chemical structure be changed so that its

effects can be prolonged?
- Would taking medicine in a certain way (e.g., oral vs nasal) improve the drug's ability to act on the brain?
- Could brain biology explain why there is individual variation in the capacity of drugs to ameliorate certain conditions?

Building on your knowledge of neurobiology, this chapter will explore the concepts needed to understand how a psychopharmacologist might approach addressing these questions.

Exercise

Can you think of other important questions that a psychopharmacologist might investigate?

Classifying drugs

Before exploring how drugs act on the body and brain, we need to clarify how we refer to different drugs; it can be confusing because certain compounds can go by different names. For example, a psychopharmacologist may describe methylphenidate as a psychostimulant (or a phenethylamine), a norepinephrine–dopamine reuptake inhibitor. In contrast,

a chemist might refer to the drug by its chemical structure: $C_{14}H_{19}NO_2$. Furthermore, methylphenidate might be prescribed by a psychiatrist as 'Ritalin' and referred to by the UK government (2022) as a 'Class B controlled substance' (which has severe penalties for illegal possession and intent to supply). While the following is likely not an exhaustive list of methods to categorise drugs, they can broadly be referred to in the following ways:

- Source
- Chemical structure
- Relative mechanism of action in the brain
- Therapeutic use or effect
- Marketed names
- Legal or social status

We will now focus on three of these categories that you are likely to encounter in your studies of biopsychology.

Classification by source

Fig 3.34. Seedhead of opium poppy [Papaver somniferum], with milky latex sap oozing from a recent cut

Drugs come from various places – some are naturally occurring, while others are created in the laboratory. Cocaine ($C_{17}H_{21}NO_4$) is an example of a naturally occurring drug

because it is directly extracted from the leaves of the coca plant. Opium is also naturally occurring, taken from the unripe seed pods of the opium poppy. In other words, all the molecules that give cocaine and opium their psychoactive properties are already present in the plant itself.

Semisynthetic drugs are chemically derived from naturally occurring substances. An example of a semisynthetic drug is heroin (a modified molecule of morphine, the main active ingredient of opium). The drug lysergic acid diethylamide (LSD) is also semisynthetic, originally derived from the grain ergot fungus.

Finally, some drugs are entirely synthetic, made from start to finish in the laboratory. Methadone, amphetamine, and 3,4-methylenedioxymethamphetamine (MDMA or ecstasy) are all examples of synthetic psychoactive substances.

Relative mechanism of action in the brain

We will discuss the details of how drugs might act in the brain in the pharmacodynamics section (Section 4). For now, it's essential to understand that certain drugs can have very similar molecular targets in the brain.

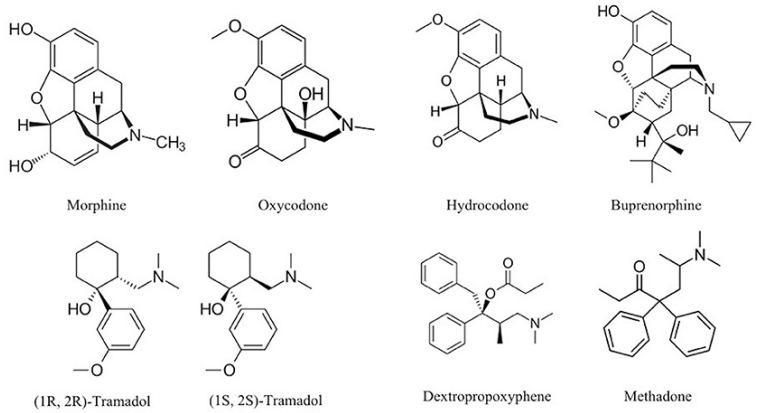

Fig 3.35. Similar chemical structures for different opioids

For example, opium (natural), heroin (semisynthetic), and methadone (synthetic) all act on opioid receptors in the brain. Therefore, these opioid-targeting drugs might be considered 'variations on a theme'. Despite their overall affinity for binding to opioid receptors, it's important to remember that the biological and psychological effects may still differ. There are different types of opioid receptors (and these might be differentially located across the brain), and certain opioid drugs might bind to some of these receptors more readily than others. These drugs may also differ in terms of how quickly they reach the brain after being administered, as well as how fast they are eliminated from the body (see Pharmacokinetics, below).

It's also crucial to remember that the brain does not express opioid receptors with the sole purpose of mediating the effects of drugs like methadone or heroin. The body already has

endogenous opioids circulating in regions of the nervous system; these molecules play essential functions, like enabling us to feel pain and pleasure and helping to regulate our respiration (Le Merrer et al., 2009; Corder et al., 2018). In contrast, drugs are **exogenous** compounds that originate outside the human body.

Therapeutic use or effect

Drugs can also be classified according to their biological, behavioural, or psychological effects. Drugs that target opioid receptors treat pain and are therefore called analgesics. Drugs that excite the central nervous system (CNS) and make us more alert are called stimulants (e.g., cocaine, amphetamine, nicotine). In contrast, substances with the opposite effect are depressants (e.g., alcohol, benzodiazepines). Some types of hallucinogens (e.g., mescaline, LSD, psilocybin) and psychotherapeutics (e.g., antidepressants like sertraline and mirtazapine) are drugs that alter psychological states.

It is also possible that some drugs can fall under different categories or are otherwise unclear what type they belong to. Ecstasy (MDMA) has a chemical structure like the stimulant amphetamine, yet it also can have hallucinogenic effects. Ecstasy is also sometimes referred to as an empathogen–entactogen because of the emotional state of relatedness, openness, or sympathy that it can create (Nichols, 2022). For all of these drugs, the effects and potential

therapeutic use depend on how much and by what method they are administered.

Pharmacokinetics

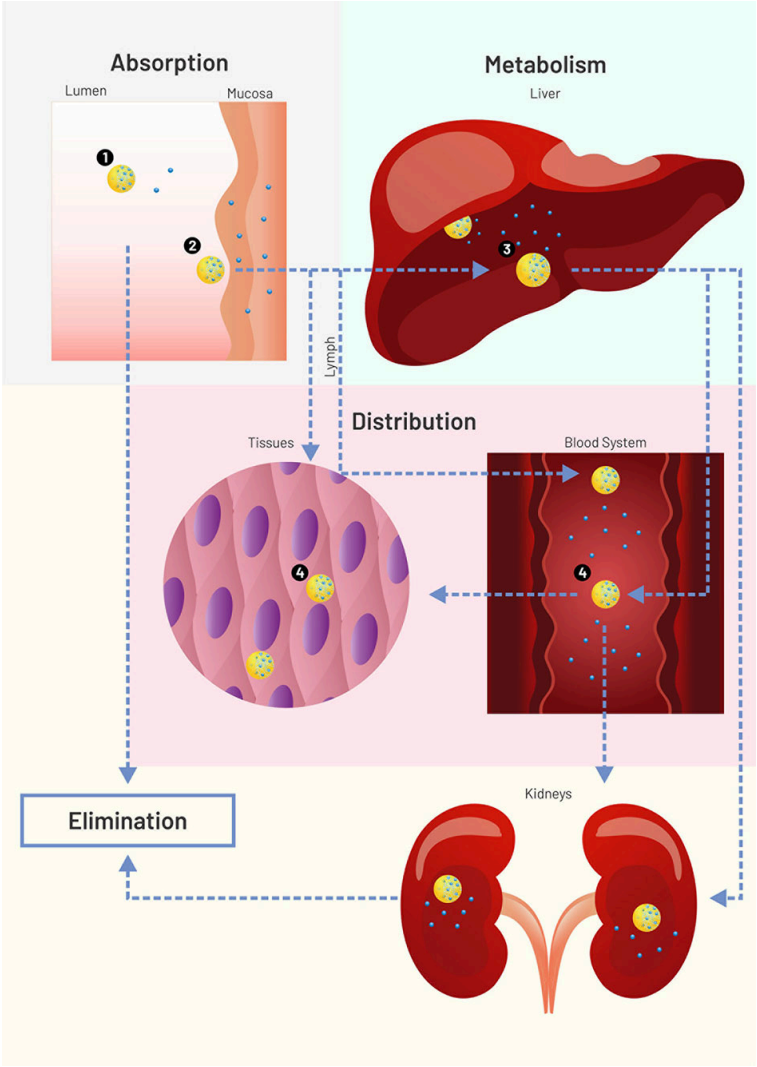

Fig 3.36. Overview of pharmacokinetics

Pharmacokinetics is a subfield of pharmacology that studies how drugs:

1. are absorbed by the body,
2. distributed,
3. metabolised, and
4. excreted from the body.

Thinking about the 'journey' a drug goes on may be helpful to understand these concepts.

A drug might first enter the body from a variety of routes. Nicotine, for example, could be smoked in a cigarette or taken via a patch applied to the skin. For this module, the effects of nicotine that we are most interested in studying are those happening in the brain. We will review how drugs like nicotine get to the brain.

Finally, you will learn how drugs, as well as their metabolites, can be removed from the body in urine.

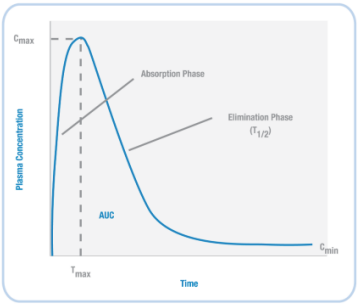

Fig 3.37. Time course of drug effects. Pharmacokinetic parameters (Cmax: maximum serum concentration; AUC: area under the curve; clearance = AUC/dose)

Based on the example timeline shown (left) for nicotine, we can plot the concentration of a drug in the body (Figure 3.37). When a drug is initially administered, the concentration in the body increases (**absorption phase**). Then, at a particular timepoint (T_{max}), the concentration reaches its highest level (C_{max}).

Next, during the **elimination phase**, the concentration of the drug in the body decreases – this happens because the drug is both metabolised and excreted. At some point, the level of the drug decreases so that it is half the value of C_{max}; the amount of time it takes to reach this point is the drug's **half-life** ($T_{1/2}$).

Absorption and distribution

Several factors influence how quickly a drug is taken up (absorbed) by the body. Perhaps the most obvious is how the drug is administered. Several non-invasive and invasive methods for drug administration are shown in Table 3.1.

Non-Invasive Methods of Drug Administration	
Oral	Into the mouth
Sublingual	Under the tongue
Nasal	Absorption through blood capillaries lining nasal cavities
Rectal	Like oral, but can be done in unconscious individuals because it doesn't require swallowing
Transdermal	On the skin via a patch
Inhalation	Into the lungs, which have a large surface area and are highly vascularised
Invasive Methods of Drug Administration	
Subcutaneous	Under the skin, but not into the muscle
Intramuscular	Into the muscle
Intravenous	Directly into the vein, so directly into the body's bloodstream
Epidural	Into the space between the dura mater and vertebrae, used in spinal anaesthesia
Injections methods primarily used in animals (e.g., in rodent models of mental health)	
Intraperitoneal	Injection into the peritoneal cavity surrounding the intestines
Intracranial	Injection into either the tissue of a specific brain region or CSF-filled ventricle; since these drugs are injected into the brain, they do not need to access the body's circulatory system

Table 3.1. Routes of administration

With so many injection methods, how is the best method for delivering a drug determined? It turns out that there are many factors influencing that decision. Intravenous (IV) infusions might be the quickest to enter the body, but a drug administered via this route might have the shortest length of action in the body (a short half-life; quickly into and out of the body). So, an IV administration of an analgesic might lead to rapid pain relief, but the effect might not last for long. Plus, some people are afraid of needles, and training is required to administer IV injections. Thus, IV injections do not allow patients to care for themselves independently.

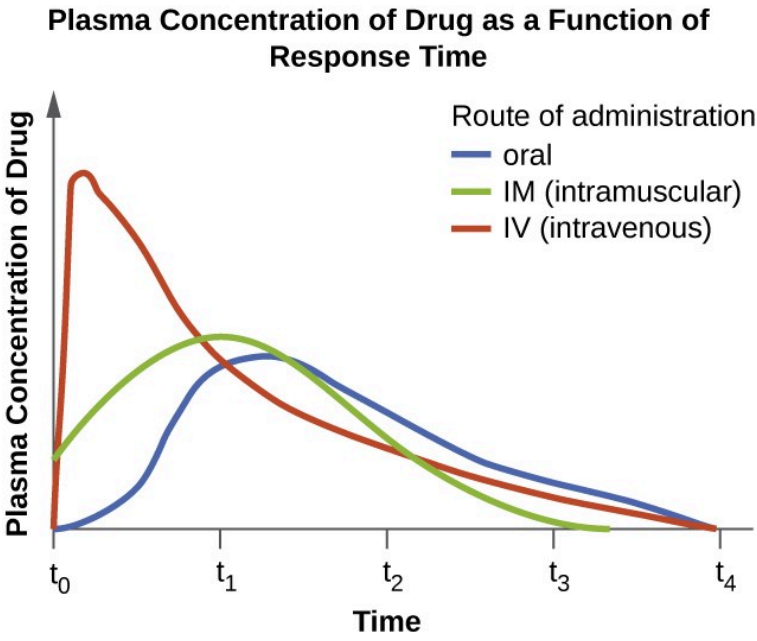

Fig 3.38. On this graph, t0 represents the time at which a drug dose is administered. The curves illustrate how plasma concentration of the drug changes over specific intervals of time (t1 through t4). As the graph shows, when a drug is administered intravenously, the concentration peaks very quickly and then gradually decreases. When drugs are administered orally or intramuscularly, it takes longer for the concentration to reach its peak.

Perhaps on the opposite end of the administration spectrum from IV injections are non-invasive oral administrations. Most individuals can swallow medications, so this method ensures a level of independent care. Unlike IV administration, however, drugs do not immediately enter the bloodstream when taken

orally. Therefore, the desired effects of medications swallowed are slower than drugs administered through IV injections.

Further complicating this is that drugs taken by the oral route are absorbed through the gastrointestinal system, and not by the mouth. This has two primary consequences. First, drugs can initially be destroyed by stomach acids, limiting the maximum effect a drug can have (e.g., for a specific medication, C_{max} might be higher after IV than after oral administration). Furthermore, the stomach environment is constantly changing (especially after meals!), impacting how much of the drug eventually reaches other parts of the body. Also, after exiting the stomach, drugs enter the liver, where they undergo **first-pass metabolism**; this can further destroy orally administered medications, reducing the concentration of the drug that reaches the rest of the body. That said, some medications (e.g., lisdexamfetamine) are designed in a certain way so that a person initially swallows an inactive **prodrug** (Mattingly, 2010). When the prodrug undergoes first-pass metabolism, it is converted into the active drug (the amphetamine) that can later impact brain function.

A few other vital implications of drug administration routes impact how quickly and for how long drugs have their effect. If a drug is administered into an area of the body with a large surface area and a high level of blood circulation (e.g., the lungs), then the drug can enter the bloodstream quicker and be faster at having its desired impact. In contrast, a drug would be much slower to act, and potentially work for a longer duration,

if it first needs to cross several cell layers before eventually arriving at a blood vessel (e.g., transdermal delivery). Furthermore, **depot binding** might occur if drugs become sequestered into inactive sites of the body where there are no receptors for them to bind to (e.g., in fat stores); these fat stores may slowly release a drug or its metabolites, further prolonging their actions on the body.

If not injected via the IV route, drugs can be slow to enter the circulatory system because they first need to pass through various membranes before entering the bloodstream (e.g., stomach wall, capillaries, etc.). While some endogenous compounds have the luxury of helper proteins designed to transport the molecule across the membrane, exogenous drugs usually do not have this mechanism. Instead, drugs most often flow from high areas of concentration to lower regions (via their **concentration gradient**) and eventually cross membranes they encounter simply through passive diffusion. Since our body's membranes are made of lipids (a **lipid bilayer**), the ability of drugs to pass through membranes is determined by their **lipid solubility** and **ionisation**.

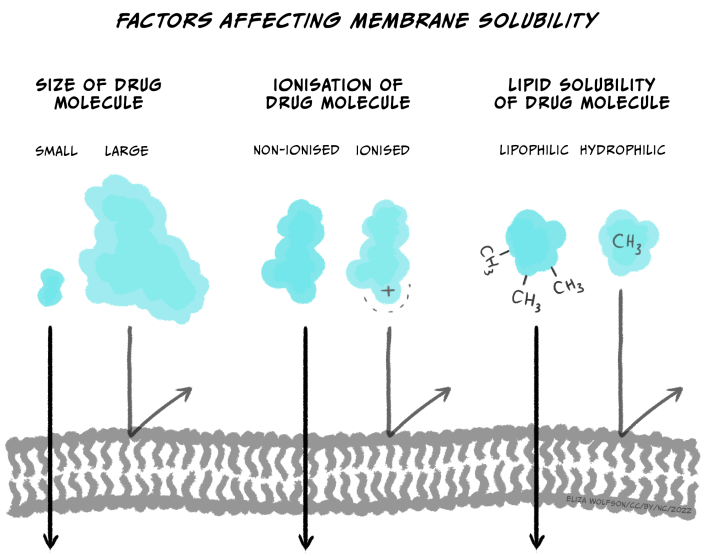

Fig 3.39. Drugs passing through membranes

Briefly, most drugs are either weak acids or bases. When a drug is dissolved in a solution, it becomes ionised (charged). The more a drug is ionised, the less lipid-soluble it becomes, decreasing its ability to cross cell membranes. In general, drugs that are acids are less ionised in more acidic solutions, while drugs that are bases are less ionised in more basic solutions. So, for example, the drug aspirin is a weak acid. If aspirin is taken orally in a tablet, it first goes to the stomach. The stomach is strongly acidic (pH 2.0), so aspirin remains primarily in its non-ionised form. Because it is non-ionized and thus lipid-

soluble, aspirin can pass through the stomach lining and enter a blood vessel. Blood, however, is slightly basic (pH 7.4); this would result in aspirin becoming ionised, making it less likely to leave the vessel because it's more challenging to cross membranes (i.e., it's now less lipid-soluble). The situation where a drug is 'stuck' in a compartment because it is highly ionised and low in lipid solubility is called **ion trapping** (Ellis & Blake, 1993). Concentration gradients can rectify ion trapping; the high concentration in one bodily compartment compared to a neighbouring compartment can encourage the drug to move across membranes to the lower concentration region. Various formulas are used to calculate drug diffusion but are beyond the scope of this module.

Finally, for drugs to enter the brain, as you've read about elsewhere, they must first cross the blood-brain barrier (BBB). Drugs that are lipid-soluble can most easily pass through the BBB. So, for example, because heroin is more lipid-soluble than morphine, it can more quickly pass through the BBB and arrive in the brain (Pimentel et al., 2020). Therefore, heroin tends to be faster acting than morphine; this may contribute to its addictive qualities.

Patients use prescribed medications to improve mental health because the drugs impact brain function. However, as described above, delivering drugs directly to the brain is challenging. Because drugs spread across our body via the bloodstream, they act in the periphery before reaching the brain; this can lead to unwanted side effects. Thus, part of the

job of a psychopharmacologist is to develop medications that can improve mental well-being via their actions on the brain while minimising undesirable and unwanted effects.

Metabolism and excretion

We have already discussed one way of inactivating drugs: first-pass metabolism in the liver, where **microsomal enzymes** can break down medications into simpler compounds. Through biotransformation, the liver can metabolise drugs so that they are more ionised; this causes them to lose their lipid solubility, further preventing them from crossing the BBB to enter the brain. Finally, metabolised drugs are primarily excreted by the body via the kidney (urine), but other excretion products include bile, faeces, breath, sweat and saliva.

Special enzymes in both the blood and the brain can also break down drugs. For example, when in the brain, heroin can be metabolised into morphine. This raises an important example – drugs can be metabolised into molecules that are also biologically active. While morphine and heroin might have similar effects, the metabolites of other drugs can have opposing effects. For example, alcohol is metabolised into acetaldehyde via the alcohol dehydrogenase enzyme (Figure 3.39). If acetaldehyde accumulates in the body, it can make a person feel sick. Acetaldehyde itself is metabolised by aldehyde dehydrogenase into acetic acid. There are drugs for alcohol use disorder that block aldehyde dehydrogenase (Disulfiram),

thereby resulting in increased levels of acetaldehyde in the body (Veverka et al., 1997). Because the effects of Disulfiram are unpleasant, it is believed that the administration of this drug might prevent people from drinking alcohol in the first place. While this might sound useful, compliance with this treatment is often an issue (Mutschler et al., 2016).

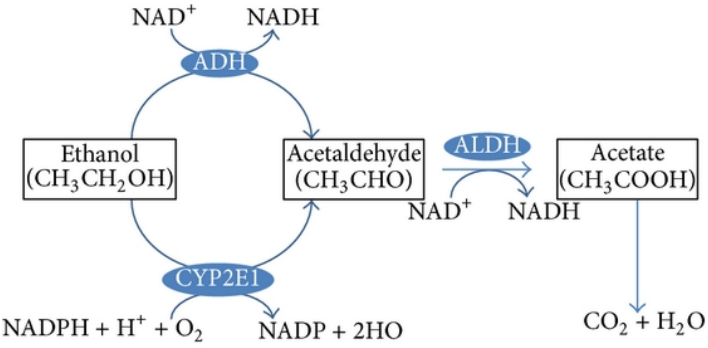

Fig 3.40. Metabolism of alcohol

Finally, there is also significant individual variation in metabolism. For example, there might be sex differences in levels of certain enzymes. Women may have lower levels of gastric alcohol dehydrogenase than men – so for a given dose of alcohol, more alcohol enters the bloodstream (Frezza et al., 1990). There is also individual adaptation – chronic drinkers have higher levels of alcohol dehydrogenase. In this example, someone with an alcohol use disorder might need more alcohol

than someone else to achieve the desired effects of alcohol – this is an example of **tolerance** resulting from a state of **enzyme induction** (increased rate of metabolism due to enhanced expression of genes for drug-metabolising enzymes; tolerance is discussed again later in this chapter). Age also impacts metabolism; older individuals have reduced liver function, and this might lead to exaggerated alcohol effects (Meier & Seitz, 2008). Finally, genetics can impact metabolism as well; some individuals have a polymorphism in the gene encoding aldehyde dehydrogenase – lack of this enzyme means there's a greater accumulation of acetaldehyde and thus more of its unpleasant effects (Goedde & Agarwal, 1987).

Pharmacodynamics

While pharmacokinetics focuses on how a drug spreads across the body and is eliminated, **pharmacodynamics** studies the effect a drug has once it reaches its target in the body. So, while pharmacokinetics explains how a drug eventually passes through the BBB to get to the brain, pharmacodynamics describes what type of receptor a drug binds to in the brain and what consequence this has on neuronal signalling. Notably, while this example discusses a drug targeting a receptor, medicines can interact with many different types of molecules in the brain, impacting brain function in numerous ways. The figure below gives a few examples of how a drug can affect synaptic transmission. Once again, it's crucial to recognise that

drugs are acting on cellular mechanisms that already exist to help us survive. For example, nicotine binds to an ionotropic receptor that usually binds the neurotransmitter acetylcholine. This particular receptor also binds nicotine, so we call it the nicotinic acetylcholine receptor.

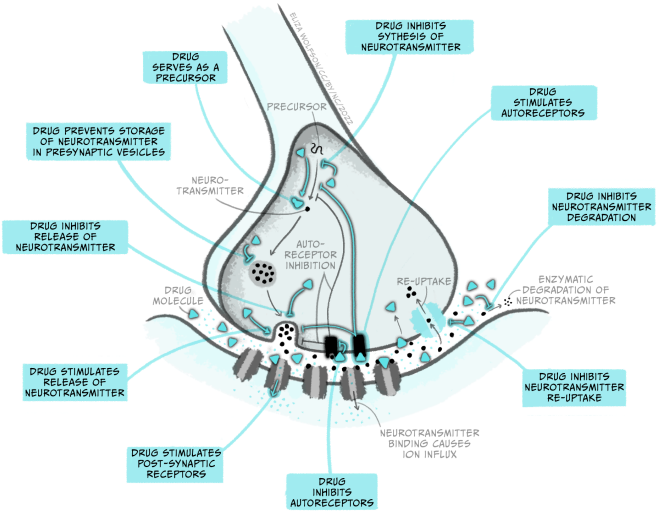

Fig 3.41. Sites of drug action

Agonist drugs and dose-response curves

Throughout this and future modules, you will learn about many different types of drugs and their impact on brain biology. For now, we will focus on two general kinds of drugs:

agonists and **antagonists** that bind to **receptors**. Receptors are molecules that a drug (or an endogenous **ligand**) bind to and initiate a biological effect. 'Receptor' is a very general term – we often think of receptors as proteins that are inserted into the membrane of cells, but this does not have to be the case (receptors can also be in the cytoplasm, for example). Some examples of where receptors can be located are shown in Figure 3.41. For example, receptors can be found on post-synaptic neurons (e.g., nicotinic acetylcholine receptors) and on pre-synaptic terminals (known as **autoreceptors**, which help self-regulate neurotransmitter release; e.g., the dopamine D3-type receptor). Drugs are often not limited to binding one particular receptor – they are often considered 'dirty', binding to multiple types of receptors to varying degrees across the body. This is one reason why drugs often have unwanted side effects. Second generation antipsychotic medications are notorious for impacting multiple types of receptors, and this may be why there is significant individual variation in their tolerability (Kishimoto et al., 2019).

Drugs that are considered agonists can bind to a receptor and initiate some type of biological effect, such as turning on an intracellular cascade of signalling events. Because of this, we often think of agonists as working via a 'lock-and-key' mechanism – inserting a drug into a receptor enables events to occur (see Figure 3.45, top). It is critical to recognise that drugs tend to bind to receptors weakly and can rapidly dissociate from the receptor. Therefore, the acute impact drugs have is

reversible. This is important because when a drug is no longer bound to a receptor, the endogenous ligand for that receptor can once again bind.

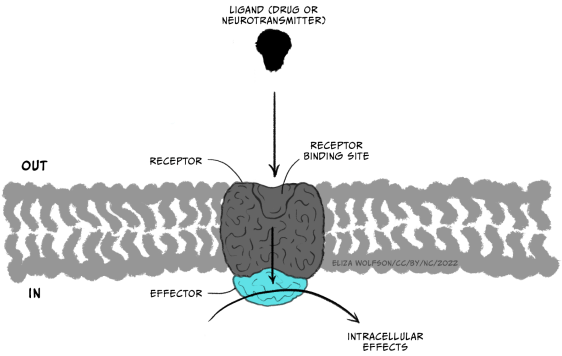

Fig 3.42. Interactions between drugs or ligands and their receptors

How much biological impact a drug has on the brain is, in part, dependent upon the number of receptors that are available to bind the drug. Therefore, increasing the number of drug molecules in the brain will also increase the probability of binding to a receptor. While larger doses of a drug can have a more significant impact on biology, there is always a limit. The maximum effect of a drug is achieved when the drug is continuously bound to all receptors; that is, a drug reaches its

maximal effect when all receptors are occupied – this is known as the **law of mass action** and can be described by a **dose-response curve**.

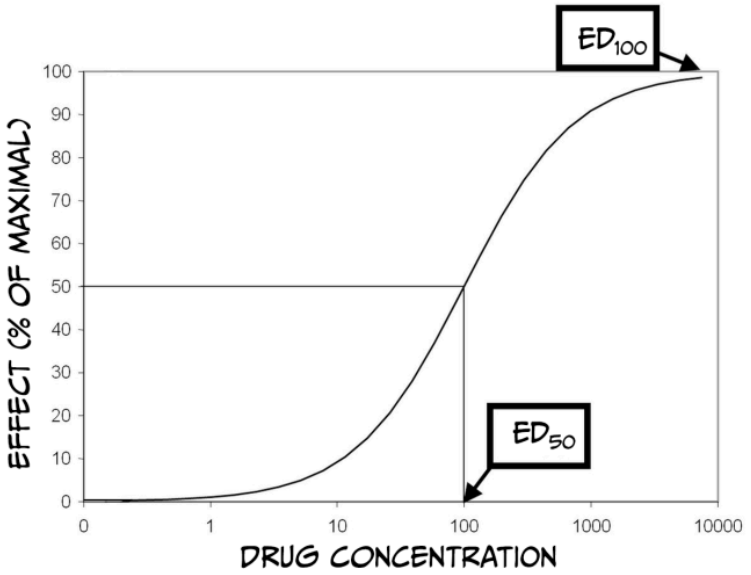

Figure 3.43 Dose-response curve

As you can see from Figure 3.43 above, dose-response curves have a typical S-shape. They are usually plotted with a logarithmic function of 'dose' of drug administered on the x-axis and a measured response on the y-axis. Looking at the figure, you can see that at some point, increasing the dose of the drug no longer produces a bigger response; at this point (known as effective dose 100, ED_{100}), the drug is occupying all of the available receptors and therefore is having its maximum effect. Another important point on the graph is the ED_{50} for

the drug. ED_{50} is a drug-potency measure representing the dose that produces half of the maximal effect. Alternatively, ED_{50} can characterise the amount of a drug that produces an effect in half of the population to which it was administered. Finally, it is also crucial to remember that most medications have various effects on the body and can even interact with multiple receptors. Binding to Receptor A might impact pain perception, while binding to Receptor B might impact blood pressure. If Receptor B is more prevalent than Receptor A, then the dose required to affect blood pressure maximally would be higher than that to alter pain perception. Accordingly, there would also be a different ED_{50} value for each response.

Drugs often have side effects that are either undesirable or dangerous; dose-response curves can also be used to characterise these effects. For example, one unwanted effect of a drug is sedation. The dose of the medicine that produces this effect in 50% of subjects is referred to as the toxic dose 50 (TD_{50}). Using this information, doctors can calculate a margin of safety, known as the **therapeutic index (or therapeutic window)** ($TI = TD_{50} / ED_{50}$), which indicates how much the dose of a drug may be raised safely. Just as drugs might have multiple desirable effects, there may also be numerous toxic effects, each with a different TD_{50}. Finally, in the therapeutic index formula, TD_{50} can be substituted with the lethal dose 50 (LD_{50}), which is the dose of a drug that can kill 50% of subjects.

Comparing dose-response curves of agonist drugs

Up until now, we have primarily discussed the **efficacy** of drugs: the maximum effect they can produce. Drugs also differ in **potency**: how much medication is needed to produce an effect. Figure 3.43 shows dose-response curves for three opioid drugs with similar efficacies; hydromorphine, morphine, and codeine can all effectively reduce pain. However, different doses of these drugs are required to relieve pain – a higher concentration of codeine is needed for pain relief than morphine. Because of this, the ED_{50} of each of these drugs are also different. Drugs with a lower ED_{50} are considered more potent than drugs with a higher ED_{50}.

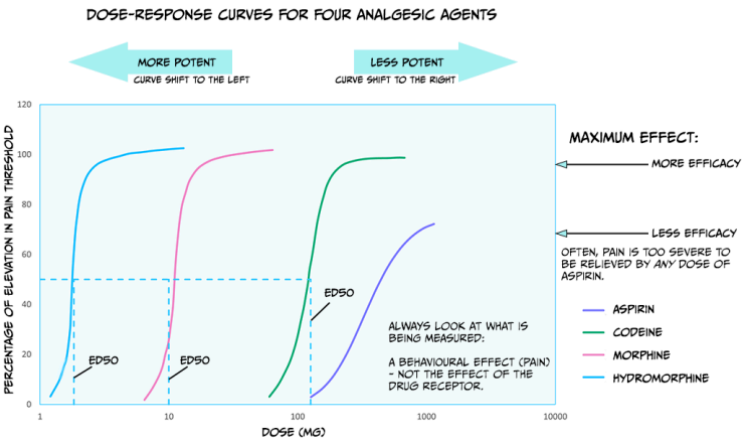

Figure 3.44. Comparing efficacy and potency of analgesic drugs

The figure also displays the dose-response curve for aspirin, a non-opioid drug that can be used to reduce pain. Not only is a higher dose of aspirin required to reach similar levels of pain relief compared to opioids like morphine, but pain can also not be entirely relieved by any dose of aspirin. So, aspirin is both less potent and less efficacious for relieving pain than morphine.

There are likely several reasons for these differences in potency between drugs. For example, the pharmacokinetics are likely different between medications; if one drug has an enhanced ability to cross the BBB, then more molecules of that drug will bind receptors, and that drug will have supreme efficacy. In addition, some drugs might have a greater **affinity** for receptors than other drugs; a drug with higher affinity will likely stay bound to the receptor for a more extended period and thus keep on having an effect. Differences in the efficacy of drugs likely signify that those medications work through different mechanisms. While both morphine and aspirin relieve pain, morphine works by binding opioid receptors, and aspirin instead inactivates the cyclooxygenase enzyme.

Antagonist drugs

Agonist drugs binding to receptors can cause a biological response – as such, they are said to have **intrinsic activity**. In contrast, **antagonists** bind to receptors and counteract either an agonist or endogenous ligand's effect on a receptor.

Therefore, one can measure the effectiveness of an antagonist by observing how its administration impacts the dose-response curves of agonist drugs. Unlike agonist drugs that follow a 'lock and key' mechanism to initiate biological effects (Figure 3.45, top row), it may be helpful to imagine an antagonist as a key that fits into a lock but does not turn (Figure 3.45, bottom row).

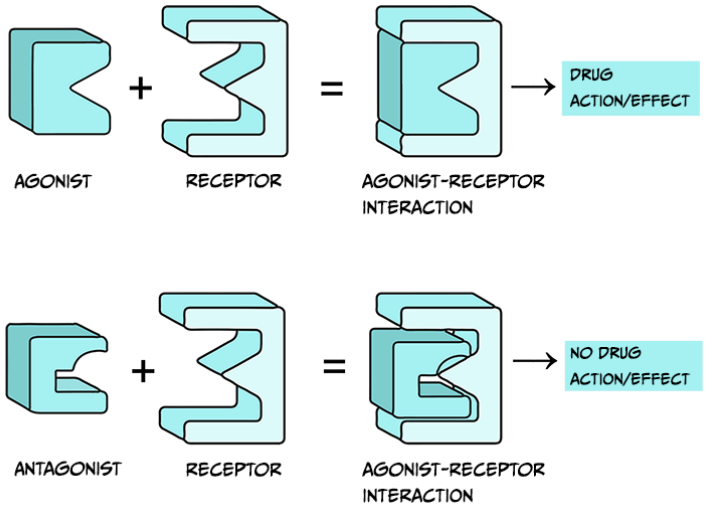

Figure 3.45. Action of agonist drugs (top), and competitive antagonists (bottom)

There are a couple of categories of antagonists that you should be familiar with (Figure 3.46). **Competitive antagonists** bind to the same site on a receptor as an agonist or endogenous

ligand. Because of this, this type of antagonist *competes with the endogenous ligand* for available binding sites. Therefore, a higher dose of the agonist drug would need to be administered to *outcompete* the presence of an antagonist; this would shift the ED_{50} of the agonist dose-response curve to the right. Theoretically, if there is so much agonist that absolutely no antagonist molecules can bind to a receptor, and if the agonist occupies all available receptors, then the agonist can reach the same ED_{100} as in the absence of an antagonist.

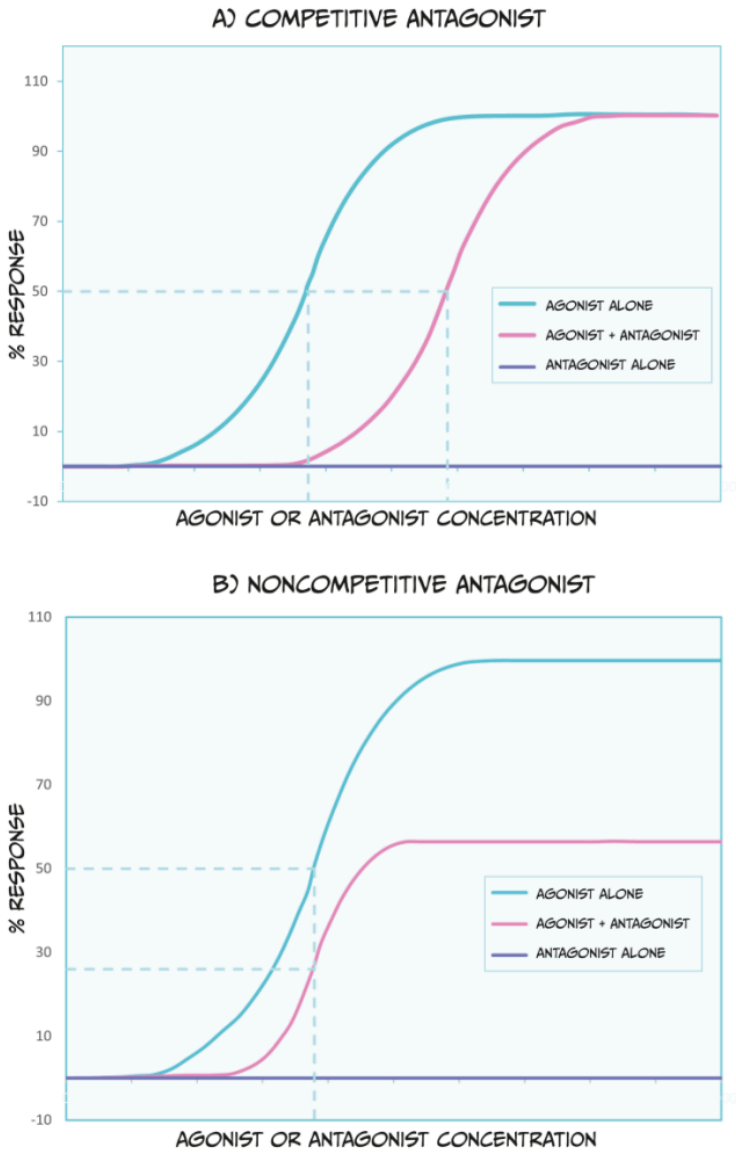

Figure 3.46 Competitive (A) vs noncompetitive (B) antagonists

Unlike competitive antagonists, **non-competitive agonists** bind to a different part of a receptor than an agonist or endogenous ligand; therefore, they do not compete for binding. In effect, non-competitive antagonists make receptors unavailable for agonist drug action. While non-competitive antagonists still shift the dose-response curves of agonists to the right (Figure 3.46 B), they can also decrease the maximum possible effect an agonist or endogenous ligand has (i.e., they reduce the ED_{100}). Because non-competitive antagonists bind to different receptor sites as agonist drugs, simply increasing the dose of an agonist cannot overcome this blockade.

When discussing agonists, we mentioned that most drugs form weak bonds with receptors. This means that the effects of the drug are reversible because the drug can easily dissociate from a receptor. Most antagonist drugs work similarly, and their interaction with receptors is temporary. However, the effects of some antagonist drugs are *irreversible* – they form a long-lasting bond with receptors. One example of such an antagonist is alpha-bungarotoxin (from banded krait venom), which blocks acetylcholine receptors at neuromuscular junctions. This can result in paralysis, respiratory failure, and death. However, there is a hypothetical scenario to overcome the effects of irreversible antagonists – synthesising new receptors. Suppose enough new receptors are formed and become functional, and the irreversible antagonist is no longer in the system (e.g., it has been eliminated via urine). In that

case, these new receptors can begin to restore biological function (when they bind agonists or endogenous ligands).

Other types of agonists

There are three other types of 'agonist' drugs you may encounter in your studies. First, **indirect agonists** (or **allosteric modulators**) can bind to a different receptor part than a (regular) full agonist or endogenous ligand. These indirect agonists help full agonists, or endogenous ligands, have their full effects. Benzodiazepines are an example of an indirect agonist because they bind to $GABA_A$ receptors, and they enhance the channel's conductance when GABA (the endogenous ligand) is also attached.

Second, **partial agonists** bind to the same receptor site as agonist drugs, but they have low efficacy (Figure 3.47). Therefore, defining a drug as a partial agonist is relative – the response to a partial agonist must be less than the maximum response produced by a full agonist. Importantly, when both full and partial agonists are administered simultaneously, they compete for the same receptor binding site. In this scenario, because the partial agonist is less effective at producing a biological response, it antagonises the effect of the full agonist. In other words, it is impossible for the body to produce a full response to the agonist because partial agonists (which are less efficacious) occupy the agonist-binding sites. Such an effect can be overcome by increasing the dose of a full agonist,

allowing it to outcompete the partial agonist for binding to receptor sites. Because of these effects, partial agonists are sometimes called **mixed agonist-antagonist** drugs.

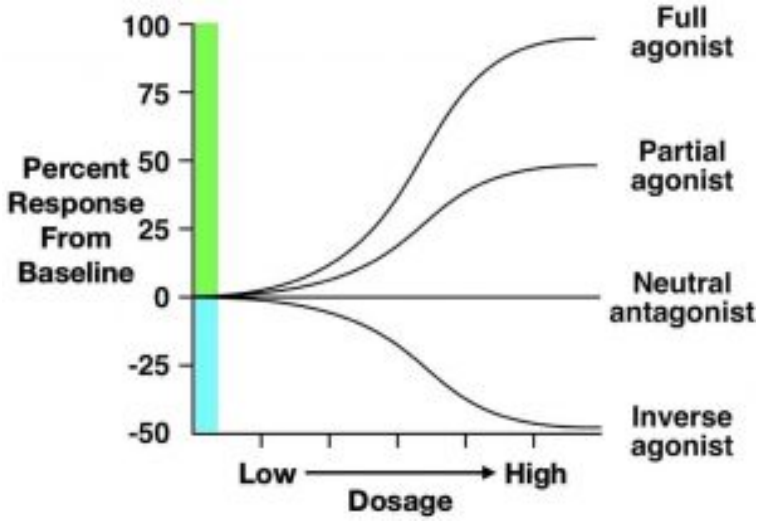

Figure 3.47. Agonist comparison

The final type of drug we will discuss is the **inverse agonist**. Some receptors in the body have substantial endogenous activity, even when ligands are not bound to them. This observation breaks the general rule that receptors have no activity when they are not bound to a ligand. Inverse agonists reduce this spontaneous activity, resulting in a descending dose-response curve (Figure 3.47). Although their mechanism is complex, some beta-carboline alkaloids are considered inverse agonists. Beta-carboline alkaloids bind to $GABA_A$

receptors at the same site as benzodiazepines. While benzodiazepines facilitate chloride conductance through the receptor channel and decrease anxiety, beta-carboline alkaloids have the opposite effects when administered (Evans & Lowry, 2007). The anxiety-inducing effects of these inverse agonists lead some people to call them 'anti-benzodiazepines'. By contrast, drugs that are competitive antagonists at the $GABA_A$ receptor do not influence the receptor's function on their own, but instead block the ability of full, partial, or inverse agonists to alter the receptor's activity.

Effects of repeated drug use

If an individual is repeatedly administered a specific dose of an agonist drug, then the ability of the drug to exert effects on the body might change. If the drug effects get smaller, this is known as **tolerance**. So, if an individual has developed tolerance to a particular drug, then the dose of the drug might need to be increased so that the drug is still efficacious. Sometimes, drug effects get bigger and bigger with repeated administrations – this finding is known as **sensitisation** or **reverse-tolerance**. Because drugs can have multiple effects on the nervous system and behaviour, some drug responses may undergo tolerance, while others are sensitised. For example, repeated administration of amphetamine can result in tolerance to the euphoria-inducing effects of the drug, but

sensitisation to specific psychomotor or psychosis-associated impacts.

Exercise

To help you think about these concepts, try drawing dose-response curves for the development of tolerance and sensitisation. Remember that, with tolerance, more drug is required to get the same effect. In contrast, with sensitisation less drug is needed to get the same effect.

Because certain drugs target similar receptors in the nervous system, sometimes **cross-tolerance** happens, where one drug also reduces the effects of another drug. For example, alcohol drinkers might be less affected by benzodiazepines since the impact of both types of drugs are dependent upon GABA transmission and the expression of GABA receptors (Lê et al., 1986). Mechanisms underlying drug sensitisation might be a bit less studied than tolerance. One example, however, of sensitisation is the ability of certain drugs (like amphetamine) to increase levels of the neurotransmitter dopamine across

administrations (Singer et al., 2009). You will learn more about drug tolerance and sensitisation when studying addiction.

Summary

Key Takeaways

- Multiple classification systems for drugs exist
- Pharmacokinetics involves the absorption, distribution, and elimination of drugs from the body
- Pharmacodynamics involves how drugs interact with receptors and alter the functional state of the receptor.

In this chapter, you have learned about different categories of drugs and how they impact the body through pharmacokinetic and pharmacodynamic processes (Figure 3.48). Entire modules are often devoted to pharmacology, and many of the concepts we described can be further quantified

via mathematical formulas, allowing for precise drug comparisons. As you study different psychiatric conditions and their biomedical treatments, be sure to refer to this chapter to help you understand how medications can be used for many individuals to improve mental wellbeing.

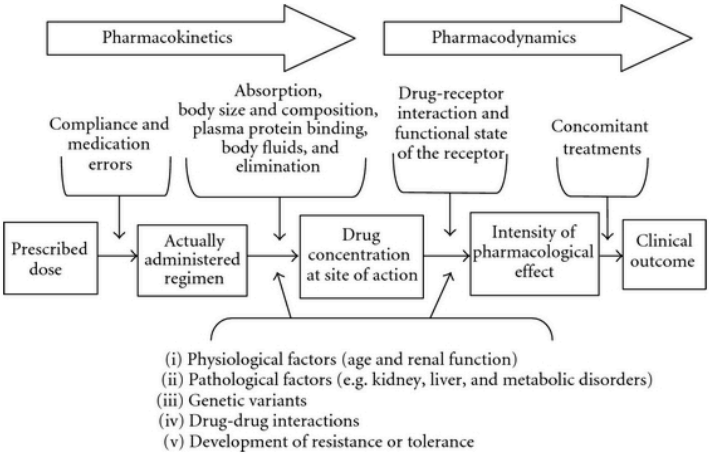

Fig 3.48 Summary of factors influencing drug action

References

Corder, G., Castro, D. C., Bruchas, M. R., & Scherrer, G. (2018). Endogenous and exogenous opioids in pain. *Annual Reviews of Neuroscience, 41*, 453–473. https://dx.doi.org/10.1146/annurev-neuro-080317-061522.

Ellis, G. A., & Blake, D. R. (1993). Why are non-steroidal anti-inflammatory drugs so variable in their efficacy? A description of ion trapping. *Annals of the Rheumatic Diseases, 52*, 241–243. https://dx.doi.org/10.1136/ard.52.3.241.

Evans, A. K., & Lowry, C. A. (2007). Pharmacology of the beta-carboline FG-7,142, a partial inverse agonist at the benzodiazepine allosteric site of the GABA A receptor: neurochemical, neurophysiological, and behavioral effects. *CNS Drug Reviews*, *13*(4), 475–501. https://doi.org/10.1111/j.1527-3458.2007.00025.x

Frezza, M., di Padova, C., Pozzato, G., Terpin, M., Baraona, E., & Lieber, C. S. (1990). High blood alcohol levels in women. The role of decreased gastric alcohol dehydrogenase activity and first-pass metabolism. *The New England Journal of Medicine*, *322*(2), 95–99. https://doi.org/10.1056/NEJM199001113220205

Goedde, H. W., & Agarwal, D. P. (1987). Polymorphism of aldehyde dehydrogenase and alcohol sensitivity. *Enzyme*, *37*(1–2), 29–44. https://doi.org/10.1159/000469239

Kishimoto, T., Hagi, K., Nitta, M., Kane, J. M., & Correll, C. U. (2019). Long-term effectiveness of oral second-generation antipsychotics in patients with schizophrenia and related disorders: a systematic review and meta-analysis of direct head-to-head comparisons. *World Psychiatry, 18*(2), 208-224. https://doi.org/10.1002/wps.20632

Lê, A. D., Khanna, J. M., Kalant, H., & Grossi, F. (1986). Tolerance to and cross-tolerance among ethanol, pentobarbital and chlordiazepoxide. *Pharmacology, Biochemistry, and Behavior*, *24*(1), 93–98. https://doi.org/10.1016/0091-3057(86)90050-x

Le Merrer, J., Becker, J. A. J., Befort, K., & Kieffer, B. L. (2009). Reward processing by the opioid system in the brain. *Physiological Reviews, 89*, 1379–1412. https://dx.doi.org/10.1152/physrev.00005.2009.

Mattingly, G. (2010). Lisdexamfetamine dimesylate: a prodrug stimulant for the treatment of ADHD in children and adults. *CNS Spectrums, 15*, 315–325. https://dx.doi.org/10.1017/S1092852900027541.

Meier, P., & Seitz, H. K., (2008). Age, alcohol metabolism and liver disease. *Current Opinion in Clinical Nutrition and Metabolic Care, 11*(1), 21-26. https:/doi.org/10.1097/MCO.0b013e3282f30564.

Mutschler, J., Grosshans, M., Soyka, M., & Rösner, S. (2016). Current findings and mechanisms of action of disulfiram in the treatment of alcohol dependence. *Pharmacopsychiatry*, *49*(4), 137–141. https://doi.org/10.1055/s-0042-103592

Nichols, D. E. (2022). Entactogens: How the name for a novel class of psychoactive agents originated. *Frontiers in Psychiatry, 13*, 863088. https://dx.doi.org/10.3389/fpsyt.2022.863088.

Pimentel, E., Sivalingam, K., Doke, M., & Samikkannu, T. (2020). Effects of drugs of abuse on the blood-brain barrier:

A brief overview. *Frontiers in Neuroscience 14,* 513. https://dx.doi.org/10.3389/fnins.2020.00513.

Singer, B. F., Tanabe, L. M., Gorny, G., Jake-Matthews, C., Li, Y., Kolb, B., & Vezina, P. (2009). Amphetamine-induced changes in dendritic morphology in rat forebrain correspond to associative drug conditioning rather than nonassociative drug sensitization. *Biological Psychiatry*, *65*(10), 835–840. https://doi.org/10.1016/j.biopsych.2008.12.020

UK Government (2022). List of most commonly encountered drugs currently controlled under the misuse of drugs legislation. (Accessed 13 Dec 2022). https://www.gov.uk/government/publications/controlled-drugs-list–2/list-of-most-commonly-encountered-drugs-currently-controlled-under-the-misuse-of-drugs-legislation

Veverka, K. A., Johnson, K. L., Mays, D. C., Lipsky, J. J., & Naylor, S. (1997). Inhibition of aldehyde dehydrogenase by disulfiram and its metabolite methyl diethylthiocarbamoyl-sulfoxide. *Biochemical Pharmacology, 53,* 511–518. https://doi.org/10.1016/S0006-2952(96)00767-8

About the author

Dr Bryan Singer
UNIVERSITY OF SUSSEX
https://twitter.com/basalganglia
https://www.linkedin.com/in/bfsinger/

Dr Bryan Singer is a lecturer in the School of Psychology at the University of Sussex (Brighton, UK). Bryan's lab is part of the highly collaborative Behavioural and Clinical Neuroscience group. He is the Director of the Sussex Addiction Research and Intervention Centre (SARIC) and a member of Sussex Neuroscience. Bryan also has an Associate role at The Open University (UK).

PART IV
SENSING THE ENVIRONMENT AND PERCEIVING THE WORLD

When you think about your senses, you will likely note that we have five senses: **touch, hearing, sight, smell** and **taste.**

As you will discover shortly, whilst this is true, it is also an oversimplification of the exquisite sensory systems our bodies possess. Even the simple names given to the senses do not do justice to the experiences they provide us with and the complexity that underpins our sensory processing.

In this section, you will learn about how the body senses the world around us. We will take each sensory system in turn and consider the sensory stimulus, how it is detected by the body, the pathways through the nervous system that the sensory information takes, and how it is processed within the brain to create a perception of the world.

Learning Objectives

By the end of this section you will be able to:

- Identify and describe the sensory stimuli for the different sensory systems
- Explain and compare how each sensory system detects sensory stimuli, converting the information into electrical signals for use within the nervous system
- Describe the pathways sensory signals take from the sense organ to the brain, noting any key processing that occurs at different points in the pathway and relating this to our perception of the stimulus
- Discuss the wider importance of our sensory systems as indicated by their functions beyond sensory perception and the impact of sensory impairment on an individual and their families.

7.

FEELING THE WORLD: OUR SENSE OF TOUCH

Dr Eleanor J. Dommett

Touch comes before sight, before speech. It is the first language and the last, and it always tells the truth.

Margaret Atwood, Poet and Novelist. *The Blind Assassin (2000)*

As the opening quote suggests, touch is fundamental to our experiences of the world, including our interactions with others. It is therefore the place we begin our journey through the senses.

Spend a moment with your eyes closed and focus on the sensations you can feel on your skin and in your body. What kinds of things can you detect?

You could have come up with a range of answers here. For example, you might have noted the feel of your clothes on your skin, how rough or smooth the fabric is or how tightly they fit. You might have felt a cool breeze across part of your body from an open door or window. You could even have realised that your body position feels a little uncomfortable, even painful, where you have sat in the same position for too long.

What you are experiencing in the exercise above is **somatosensation** which means *bodily senses*. This includes the sense of touch, but also includes the sensing of temperature, pain and **proprioception**, of which the latter can be defined as the sense of our own body position. In this first section we will focus on touch, before examining pain in the next section.

Sensing touch: getting to grips with the skin

To understand how we detect touch information, we need to understand a little about the structure of our skin. The skin is the largest organ in the human body and it incorporates the sensory receptor cells that allow us to detect touch as well

as blood vessels, sweat glands and various other specialised structures. Critically for our sense of touch, the skin can be described as a viscous liquid, much like golden syrup or honey. It is said to have viscoelasticity, which means that when forces are applied to it, stresses and strains are created within the skin that can be detected by sensory receptor cells.

Note that these sensory receptor cells are distinct from the receptors you would have read about when studying neurotransmitters. Sensory receptor cells are whole cells designed to detect sensory signals or stimuli rather than a subcellular structure which binds to neurotransmitters or other small molecules.

There are four main types of sensory receptor cells which are critical for our sense of touch, each named after the biologist who discovered them:

- Meissner's corpuscles
- Merkel's discs
- Pacinian corpuscles
- Ruffini's endings

These receptors are **mechanoreceptors** because they detect a mechanical stimulus. They can all be classed as a type of modified neuron. This means that they have a cell body and axon and are capable of producing an action potential.

The four types of receptors are shown in Figure 4.1. You can see from this figure that they are positioned at different depths

within the skin. Merkel's discs and Meissner's corpuscles are positioned superficially, whilst the Pacinian corpuscles and Ruffini's endings are positioned deep within the skin. The positioning gives us a clue about what kind of stimuli these different touch receptors detect.

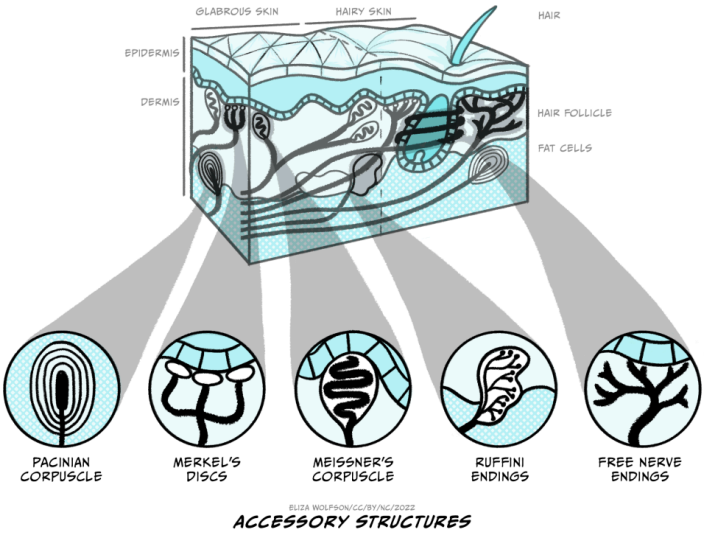

Fig 4.1. The skin is a complex structure containing the sensory receptor cells for touch. Each receptor, named after the biologist who discovered it, has a distinct structure.

If a receptor is positioned very deeply within the

skin, would you expect it to be able to detect very light or gentle touch?

No, you would not. The stresses and strains set up in the skin will be proportional to the stimulus so a very light touch to the skin will only create forces within the the superficial areas of the skin.

Scientists have now characterised these different receptors and have a good understanding of the types of stimuli they each respond to. A common feature of sensory systems is adaptation, which is a change in response of the receptor – normally a decrease – to a constant stimulus. Adaptation is very important in sensory systems because key information for our survival often comes from changing stimuli rather than constant ones. By signalling change, our sensory systems avoid wasting energy on signals that merely communicate that everything is staying the same, and therefore provide no new information. On this basis, touch receptors can be defined as fast or slow adapting receptors. The fast adapting receptors will stop responding very quickly to a constant stimulus whilst the slow adapting ones will likely continue to respond, albeit at

a lesser level, to constant stimuli. This is summarised in Table 1.

Sensory receptor cells	Location	Activating stimulus
Meissner's corpuscles	Superficial	Light touch and vibration
Merkel's discs	Superficial	Light touch and pressure
Pacinian corpuscles	Deep	Heavy pressure and vibration
Ruffini's endings	Deep	Skin stretch

Table 1. Characteristics of sensory receptor cells for touch

Before we look at how the body converts touch signals into neural signals, it is important to take note of the overall distribution of these receptors across the body, because this explains why some parts of the body are much more sensitive than others. If you have time, and someone willing to help you, have a go at the brief experiment outlined in Box 1 before you continue reading. If you do not have time, you can return to this at any point.

Box 1: Two-point discrimination experiment

This experiment demonstrates that different areas of the body differ in their ability to distinguish between a single stimulus and two stimuli placed on the skin. When two points (e.g., the ends of a cocktail stick or compass) are gently touched on the skin at the same time, they are usually felt as two different points. However, if they are very close together, they may only be detected as a single point. Different areas of the body will have different thresholds for which separation is felt and this is the 'two-point discrimination threshold'.

You will need a helper for this experiment. Read all the steps carefully before you start and gather the following pieces of equipment:

- A large paper clip or similar object with two small points that can be bent into a 'U' shape such that both points or ends are level
- Material to serve as a blindfold

- A ruler or tape measure

Now follow the steps below (this assumes you are the experimenter and your helper is the participant):

Agree which of the three body parts you will investigate from the following: index finger, palm of hand, upper arm, forehead, thigh.

Explain to your participant that, whilst they are blindfolded, you will touch their skin with the paperclip and ask them to tell you whether they felt one or two points after each touch, noting in your explanation that you will randomly select one or two points to touch them with. Let them know that you will not tell them if they are right or wrong after each guess.

Once your participant is blindfolded you can begin testing. For the first body area apply either two points or a single point in a random order. For the two-point touches you should begin with the points very close together and gradually widen them. You are looking for the point at which they correctly state that they detect two points. At this point measure the distance between the two points. Remember not to tell your participant if they are correct when they guess. You should then repeat

the process in the same body area, this time beginning with the points further apart, and bringing them together. As before, note the final distance at which they detect two points. Work out the average of the two distances and note this down as your discrimination threshold.

Now repeat all this on the other two body areas.

For reference here are some typical values in millimetres (mm):

- Index finger – 2 mm
- Upper arm – 47 mm
- Palm of the hand – 13 mm
- Forehead – 18 mm
- Thigh – 46 mm

If you map the thresholds for all areas of the body, you will find that some areas are more sensitive than others. For example, the upper lip and fingertips are much more sensitive than the back. This sensitivity arises because of the different receptor types and the receptive fields, that is the skin area where a touch will be detected by a single receptor cell. Where there is a high density of receptors with small, distinct receptive fields, areas are very sensitive. This is because two points, close together are still likely to fall in different receptive fields and therefore

be perceived separately. In contrast areas where receptive fields are large or overlapping, leave areas less sensitive because two points will likely activate the same receptor and so be perceived as a single stimulus.

So far we have focused on the skin and the receptors within it that can detect touch information but we have not yet looked at how that detection occurs. For a sensory signal to be used by the nervous system, it must be converted into a neural signal, or a change in membrane potential. The process whereby a sensory stimulus is converted into a electrical signal in the form of a membrane potential is referred to as **transduction**.

From touch to nerve impulse

Transduction is a common process across all of our sensory systems but exactly how it works varies with the sensory stimulus and receptor involved. Much of what we know about transduction in touch comes from investigations in Pacinian corpuscles because these have been the easiest to access for laboratory tests. Many of the studies done have actually focused on cells within cats, which closely resemble those in humans.

Figure 4.2 shows the structure of a Pacinian corpuscle. In this figure you can see the corpuscle is made up of multiple layers, like an onion skin. In the middle of these layers there is a sensory nerve ending with an unmyelinated tip. Remember

that myelin is a fatty substance that typically covers axons to provide electrical insulation.

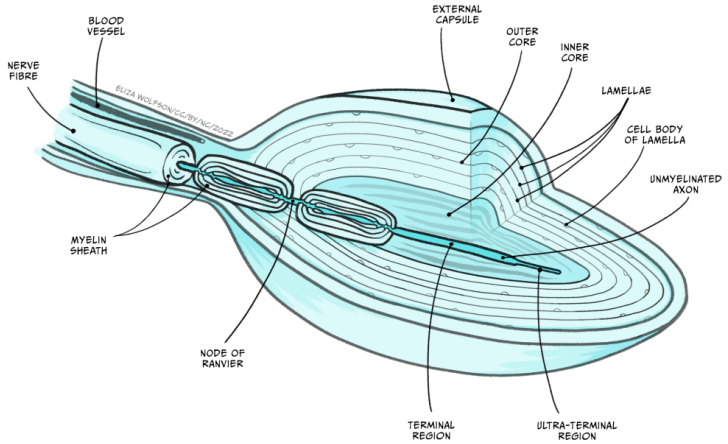

Fig 4.2. The structure of the Pacinian corpuscle showing a layered capsule containing the sensory nerve ending in the centre

When a force is applied to the skin, the layered corpuscle acts as a mechanical filter and the strain created by the force is transmitted to the unmyelinated tip through the corpuscle. The membrane of this tip contains mechano-sensitive ion channels. The term mechano-sensitive indicates that the channels will open and close depending on the mechanical force applied to them.

During transduction, the force applied causes the ion

channels to open. This causes an influx of sodium ions (Na^+) into the unmyelinated tip of the Pacinian corpuscle. You should remember from your studies of the resting and action potentials that ions will move down their electrical and chemical gradients. In this case, sodium ions, which are positively charged, are more abundant outside the corpuscle in the positively-charged extracellular space. When the ion channels open, they move into the negatively-charged cell which has a lower concentration of the ion.

Using your understanding of action potentials, what impact do you think this influx of sodium ions will have on the Pacinian corpuscle?

Hopefully you have noted that it would depolarise the cell as the inside becomes less negative than it is at rest. 'Rest' in this case means 'in the absence of any touch stimulus'.

This depolarisation is referred to as a receptor potential because it is a change in membrane potential within a sensory receptor cell caused by the presence of a sensory stimulus. The receptor potential is similar to a post-synaptic potential in that

it degrades quite rapidly, but if there is sufficient depolarisation at the point where the unmyelinated tip meets the first myelinated region (Figure 4.2), an action potential will be triggered.

You should recall from your studies of action potentials that they involve a coordinated movement of sodium and potassium ions across the membrane and this type of signal can be transmitted over long distances. This is particularly important in the senses because some of our sensory receptor cells are a very long way from our spinal cord and brain. In the case of touch, sensory receptor cells in the toe must be able to send signals over a metre to the spinal cord.

Before we look closely at the pathway touch information takes to the brain, it is useful to note the relationship between the size of the stimulus and the size of the receptor potential. Figure 4.3 shows that as the stimulus increases in intensity, the receptor potential gets larger.

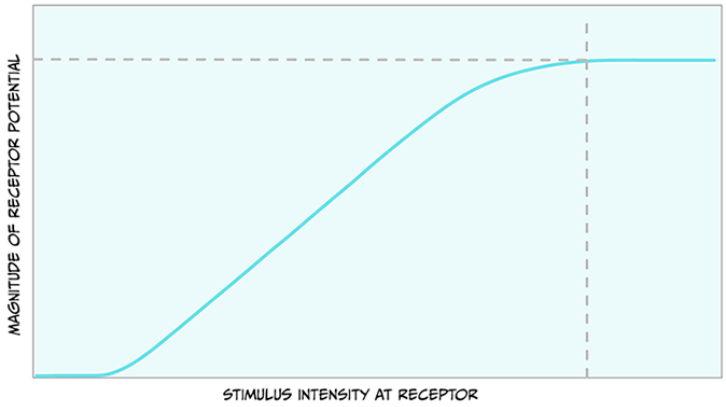

Figure 4.3. The relationship between the stimulus intensity and the receptor potential shown here indicates that as the intensity increases the receptor potential gets larger. This occurs as the ion channels remain open for longer, allowing for a greater influx of sodium ions.

Action potentials are all-or-nothing signals, meaning that they cannot change in size. What do you think a larger receptor potential means for the action potentials created?

Because the action potentials cannot get bigger, they must encode the larger stimulus in another way. The way they do this is through a greater frequency of action potentials.

Now that an action potential has been created in the neurons that detect touch, this information must travel to the brain.

Touch pathways to the brain

Whilst the sensory endings of these sensory receptor cells are found all over the body, their cell bodies are found in the dorsal root ganglion, shown in Figure 4.4.

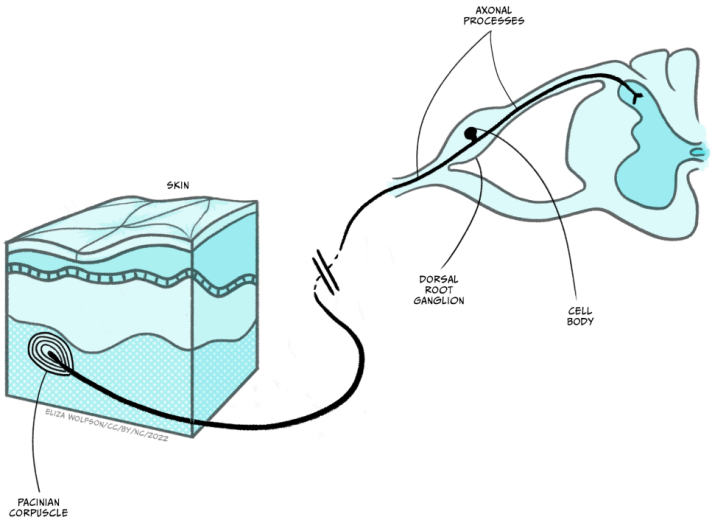

Fig 4.4. The sensory receptor cells from the skin project towards their cell body in the dorsal root ganglion which sits just outside the spinal cord before continuing into the spinal cord.

Recall that the structure of the spinal cord is relatively simple and repeats from the base (sacral regions) to the top (cervical regions). This repeating structure means that there is a dorsal root ganglion at every segment, or height, of the spinal cord. Exactly which one the information enters into depends on where in the body the information has come from. Figure 5 shows the cervical, thoracic, lumber and sacral nerves entering the different segments of the spinal cord and the regions of the body they receive information from.

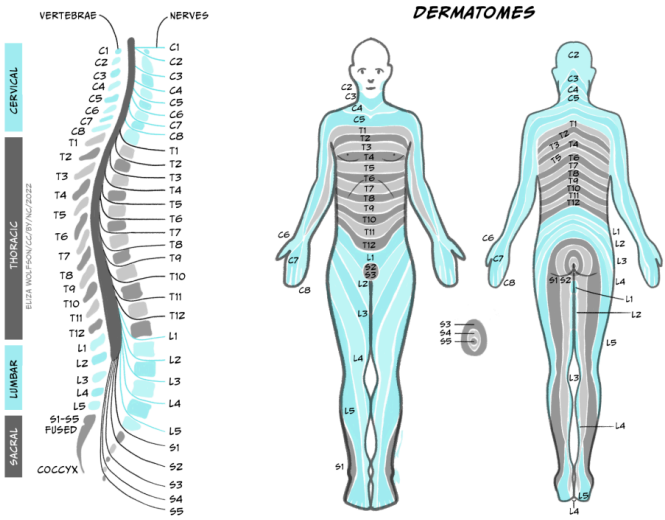

Fig 4.5 The spinal vertebrae and nerves (a) and the corresponding areas of the body sending touch information into the spinal cord (b).

Using Figure 4.5, can you identify which spinal nerve comes from the thumb?

C6 carries information from the thumb into the spinal cord.

Note that the figure makes no reference to information coming from our face. There is a separate system for carrying somatosensory information from the face, called the trigeminal system, which operates in a very similar way to that described here except that the sensory neurons enter the central nervous system at the brainstem instead of the spinal cord.

Once information from the sensory nerve ending reaches the cell body in the appropriate dorsal root ganglion, it carries on into the spinal cord via the dorsal root, which is formed of the axons of these sensory cells. These neurons have a slightly different structure to typical neurons found in the brain because they have a bifurcating axon, meaning their axon splits in two (Figure 4.6) and this allows the same neuron to transmit information from the sensory nerve ending where the mechano-sensitive channels are, beyond the cell body and into the spinal cord.

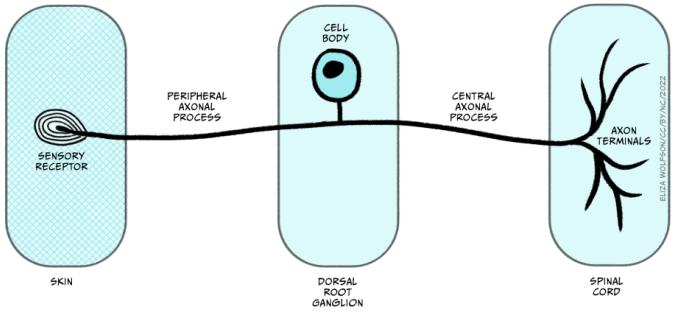

Fig 4.6. Sensory receptor cells for touch have their sensory nerve endings in the skin and a bipolar or bifurcating axon. This means that their axon is split into two parts. The first part travels from the sensory nerve ending to the dorsal root ganglion, where the cell body is found. The second part continues from here to the spinal cord. Although the axon can be considered as two parts with the cell body around half way along it, information travels uninterrupted from the sensory nerve ending to the spinal cord.

There are multiple pathways by which touch information can reach the brain, but here we will focus on the most critical pathway, called the **dorsal column/medial lemniscal pathway**. This pathway is shown in Figure 4.7.

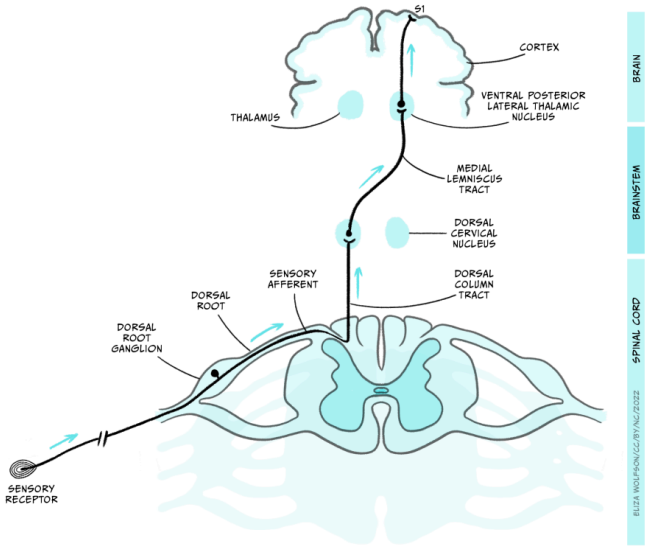

Fig 4.7. The main pathway for touch information to the brain is shown here as a line travelling into the spinal cord and up to the DCN, VPL and then S1.

The axons enter the spinal cord and pass directly up it, on the same side of the midline, until they enter the dorsal column nuclei (DCN) in the medulla where they synapse with the next neuron in the pathway. This next neuron is referred to as the 'second order neuron' because the sensory receptor cell is a modified neuron, and was therefore the first order neuron. The axons of the second order neurons travel in a pathway called the medial lemniscus to the thalamus. Specifically, they reach an area called the ventral posterior lateral thalamic nucleus (VPL), where they synapse again with the third order

neuron. The third order neuron carries the signal to the primary somatosensory cortex (S1) within the parietal lobe.

Representation of touch in the somatosensory cortex has long been understood to be topographically organised, meaning that areas of the body are represented in a way that is proportional to the input they receive, creating a mini map of the body in S1. This is known as the somatosensory homunculus or 'little man' and this was first proposed in 1937 (Figure 4.8) (Penfield & Boldrey, 1937). Much of the research that led to this proposal was conducted by Penfield, a neurosurgeon, who applied electrical stimulation to the cortical surface doing surgery in patients with epilepsy (Box 2).

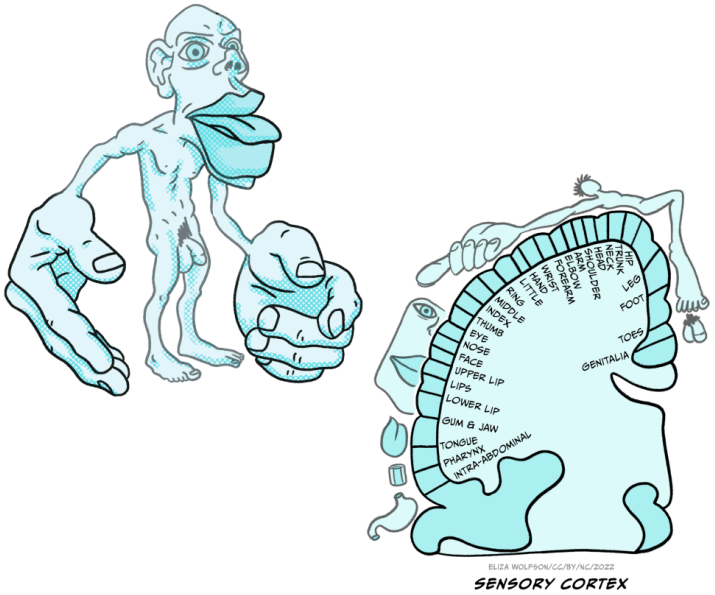

Fig 4.8. The sensory homunculus or ‘little man’ and representation of the body parts in the primary somatosensory cortex.

Box 2: Mapping the brain

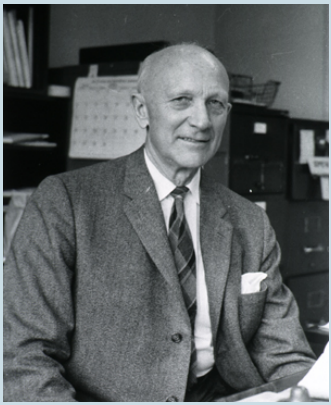

Fig 4.9. Neurosurgeon Wilder Penfield

Wilder Penfield (1891-1976, Figure 9) conducted surgery in patients with epilepsy or brain tumours. During this surgery he would apply a small electrical stimulation to the outer surface of the exposed cortex. Patients were conscious during this surgery and able to communicate with Penfield, meaning they could tell Penfield what they felt when he applied stimulation to different regions. The patient being conscious is not uncommon in brain surgery and allows the surgeons to carefully target specific areas.

Over the years, Penfield conducted cortical stimulation on over 100 patients and he kept

meticulous notes and drawings indicating responses to specific areas of stimulation. It is from this research that the **homunculus**, which is Latin for 'very small human', was born. Penfield's work is not without its limitations – it is noteworthy that exact stimulation patterns and intensity were not recorded, meaning that the final representation may not be entirely accurate (Matias, 2020). However, despite the potential inaccuracies, the idea has persisted and is still used to inspire or explain research almost 100 years later (Pan, Peck, Young, & Holodny, 2012).

The sensory homunculus is matched by a motor homunculus mapped onto the motor cortex (Figure 4.10). The two representations are connected and the connection between the two is likely to be critical for some aspects of movement including fine motor control. For example, researchers have found that impaired connectivity between these areas can underpin poor fine motor control in autism spectrum disorder (Thompson et al., 2017).

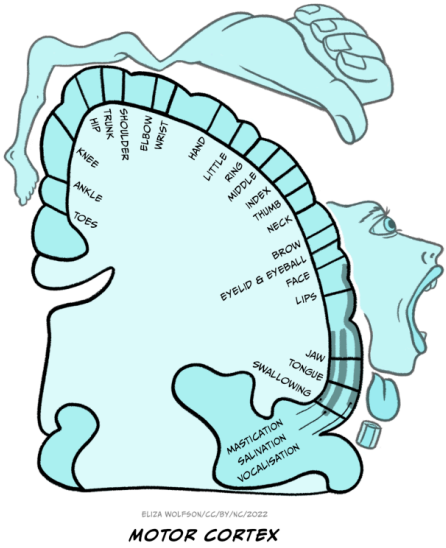

Fig 4.10. The motor homunculus showing how the body is mapped onto the motor cortex.

It is important to recognise that touch processing does not stop at the level of the primary somatosensory cortex. Research suggests an extensive cortical network is involved in processing touch information with signals from S1 continuing the secondary somatosensory cortex (S2) also located in the parietal cortex, and the insular cortex, a cortical area nestled deep within the cortical folds between the parietal and temporal lobes (Rullmann, Preusser, & Pleger, 2019).

The perception of touch

You have now considered how touch stimuli are detected by sensory receptor cells, the process of transduction and how the information travels to, and is represented in, the brain. In this final section we will look at how touch is perceived, that is, what meaning can be gained from our sense of touch.

Touch could be perceived as simply the physical encountering of objects in our environment but in fact it is much more than this. We do not simply encounter objects and identify that they are present. Rather we can glean detail of the object's size, weight, texture, stiffness and various other characteristics through our sense of touch. All of this allows us to identify specific objects and make appropriate behavioural responses to them.

Exactly what we perceive from touch may depend on the type of touch that we engage in. Touch can be categorised as either active or passive. Active touch requires the movement of the fingers over the object, that is an intentional interaction with the stimulus, whilst passive touch simply involves the object being pressed against the fingers. An example of active touch is when you reach out to feel the fabric of clothes you are considering buying and run it between your fingers to establish qualities such as smoothness, thickness and weight. By contrast, passive touch would be having the fabric pushed against your fingers.

Early research suggested that active touch may be more

informative in determining object shape (Gibson, 1962), but later work controlling for details such as the pressure with which the object was applied to the skin, showed little difference between the two types of touch, or even an advantage of passive touch (Chapman, 1994). Activation in the primary somatosensory cortex has been shown to differ between the two modes, with greater activation under active touch conditions. It has been suggested that the greater activation from active touch could arise because of activation from the motor cortex into the primary somatosensory cortex as the fingers move (Simões-Franklin, Whitaker, & Newell, 2011).

Researchers have also argued that the differences in these two types of touch can be underpinned by the role of other systems, specifically proprioception. Active touch will activate the same sensory receptor cells in the skin as passive touch, but it will also activate proprioceptive receptors, that is the ones that detect the position of the body, in this example, the fingers. It has been proposed that both the touch and proprioceptive inputs converge in the brain, causing greater excitation (Cybulska-Klosowicz et al., 2020). However, other researchers have countered this, suggesting that the two inputs compete rather than combine within the brain (Dione & Facchini, 2021). There is still much work to be done to fully understand the effects of active and passive touch on the brain.

Touch and social bonding

It should be clear from what you have read so far that touch is extremely important for perceiving the world around us including identifying the objects that our bodies come into contact with. This kind of touch can be considered as discriminatory touch. However, we do not just come into contact with inanimate objects! Much of the physical contact we have in the world is with other people and in this context, touch perception is often about affective experience, rather than discriminatory. This affective touch plays a critical role in social bonding (Portnova, Proskurnina, Sokolova, Skorokhodov, & Varlamov, 2020).

Affective touch begins from a very early age, with parent-infant touch a key part of the nurturing process. A large body of research now shows links between early nurturing tactile interactions to later life social and emotional functioning. This relationship is thought to be mediated in part by the hypothalamic-pituitary-adrenal (HPA) axis which underpins our body's stress responses and can be suppressed by various hormones, including oxytocin (Walker & McGlone, 2013). Research in rodents has shown that greater nurturing behaviour in the form of licking, grooming, huddling, and playing results in a greater density of connections in the somatosensory cortex of the offspring (Seelke, Perkeybile, Grunewald, Bales, & Krubitzer, 2016).

Such comparisons are difficult to conduct in people because

it would be unethical to divide human infants and parents into high and low nurturing conditions. However, one approach is to consider a group who would likely have experienced reduced nurturing without any experimental intervention. This approach was taken by a group of researchers who compared care-leavers with non care-leavers in the UK (Devine et al., 2020). They noted that the main reason for entering care in the UK is neglect and abuse and inferred from this that those in care may have received reduced tactile nurturing in infancy. The researchers used a range of measures to look at sensitivity and found care-leavers to be less sensitive to the affective components of touch.

It is not just early nurturing that can alter touch sensitivity and affective touch. Research has found that levels of empathy (Schaefer, Kühnel, Rumpel, & Gärtner, 2021) and loneliness (Saporta et al., 2022) can also have an effect on how people perceive affective touch. Additionally, the presence of certain diagnoses may also impact on touch perception. For example, individuals with Autism Spectrum Disorder have impaired responses to interpersonal touch (Baranek, David, Poe, Stone, & Watson, 2006).

You have now completed the first section on the senses with this section on touch. This section has introduced you to some key concepts including transduction, sensory pathways and the wider social implications of our senses.

Key takeaways

In this section you have learnt:

- The sense organ for touch is the skin, the largest organ in the body, which contains the sensory receptor cells critical for touch – Meissner's corpuscles, Merkel's discs, Pacinian corpuscles and Ruffini's endings
- Each type of sensory receptor cell for touch can be found in a specific location within the skin and can give rise to a specific sensation. Receptors also differ in how quickly they adapt
- The type of receptors found and how densely they are packed determines the sensitivity of different parts of the body
- Sensory information enters the spinal cord and rises to the level of dorsal column nuclei in the medulla before crossing to the contralateral side in the medial lemniscus. After this it enters the ventral posterior lateral

thalamic nuclei before continuing to the primary somatosensory area and other cortical regions for further processing
- Touch can serve both discriminatory and affective functions and can be considered in terms of active and passive touch
- Affective touch is critical for social and emotional functioning through its role in social bonding. Early adversity in the form of lack of nurturing tactile stimulation can have a long-lasting impact. Altered perception of affective touch can also be related to empathy, loneliness and the presence of diagnoses such as autism spectrum disorder.

References

Baranek, G. T., David, F. J., Poe, M. D., Stone, W. L., & Watson, L. R. (2006). Sensory experiences questionnaire: discriminating sensory features in young children with autism, developmental delays, and typical development. *J*

Child Psychol Psychiatry, 47(6), 591-601. https://doi.org/10.1111/j.1469-7610.2005.01546.x

Chapman, C. E. (1994). Active versus passive touch: factors influencing the transmission of somatosensory signals to primary somatosensory cortex. *Canadian Journal of Physiology and Pharmacology, 72*(5), 558-570. https://doi.org/10.1139/y94-080

Cybulska-Klosowicz, A., Tremblay, F., Jiang, W., Bourgeon, S., Meftah, E.-M., & Chapman, C. E. (2020). Differential effects of the mode of touch, active and passive, on experience-driven plasticity in the S1 cutaneous digit representation of adult macaque monkeys. *Journal of Neurophysiology, 123*(3), 1072-1089. https://doi.org/10.1152/jn.00014.2019

Devine, S. L., Walker, S. C., Makdani, A., Stockton, E. R., McFarquhar, M. J., McGlone, F. P., & Trotter, P. D. (2020). Childhood Adversity and Affective Touch Perception: A Comparison of United Kingdom Care Leavers and Non-care Leavers. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.557171

Dione, M., & Facchini, J. (2021). Experience-driven remodeling of S1 digit representation in awake monkeys: the challenge of comparing active and passive touch. *J Neurophysiol, 125*(3), 805-808. https://doi.org/10.1152/jn.00380.2020

Gibson, J. J. (1962). Observations on active touch.

Psychological Review, 69(6), 477-491. https://doi.org/10.1037/h0046962

Matias, C. M. (2020). Edwin Boldrey and Wilder Penfield's homunculus: From past to present. *World Neurosurgery, 135*, 14-15. https://doi.org/10.1016/j.wneu.2019.11.144

Pan, C., Peck, K. K., Young, R. J., & Holodny, A. I. (2012). Somatotopic organization of motor pathways in the internal capsule: a probabilistic diffusion tractography study. *AJNR American Journal of Neuroradiology, 33*(7), 1274-1280. https://doi.org/10.3174/ajnr.A2952

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain, 60*(4), 389-443. https://doi.org/10.1093/brain/60.4.389

Portnova, G. V., Proskurnina, E. V., Sokolova, S. V., Skorokhodov, I. V., & Varlamov, A. A. (2020). Perceived pleasantness of gentle touch in healthy individuals is related to salivary oxytocin response and EEG markers of arousal. *Experimental Brain Research 238*(10), 2257-2268. https://doi.org/10.1007/s00221-020-05891-y

Rullmann, M., Preusser, S., & Pleger, B. (2019). Prefrontal and posterior parietal contributions to the perceptual awareness of touch. *Scientific Reports, 9*(1), 16981. https://doi.org/10.1038/s41598-019-53637-w

Saporta, N., Peled-Avron, L., Scheele, D., Lieberz, J., Hurlemann, R., & Shamay-Tsoory, S. G. (2022). Touched by loneliness-how loneliness impacts the response to

observed human touch: a tDCS study. *Social Cognitive and Affective Neuroscience, 17*(1), 142-150. https://doi.org/10.1093/scan/nsab122

Schaefer, M., Kühnel, A., Rumpel, F., & Gärtner, M. (2021). Dispositional empathy predicts primary somatosensory cortex activity while receiving touch by a hand. *Scientific Reports, 11*(1), 11294. https://doi.org/10.1038/s41598-021-90344-x

Seelke, A. M. H., Perkeybile, A. M., Grunewald, R., Bales, K. L., & Krubitzer, L. A. (2016). Individual differences in cortical connections of somatosensory cortex are associated with parental rearing style in prairie voles (*Microtus ochrogaster*). *Journal of Comparative Neurology, 524*(3), 564-577. https://doi.org/10.1002/cne.23837

Simões-Franklin, C., Whitaker, T. A., & Newell, F. N. (2011). Active and passive touch differentially activate somatosensory cortex in texture perception. *Human Brain Mapping, 32*(7), 1067-1080. https://doi.org/10.1002/hbm.21091

Thompson, A., Murphy, D., Dell'Acqua, F., Ecker, C., McAlonan, G., Howells, H., Baron-Cohen, S., Meng-Chuan, L., Lombardo, M. V., the MRC AIMS Consortium, & Catani, M. (2017). Impaired communication between the motor and somatosensory homunculus is associated with poor manual dexterity in autism spectrum disorder. *Biological Psychiatry, 81*(3), 211-219. https://doi.org/10.1016/j.biopsych.2016.06.020

Walker, S., & McGlone, F. (2013). The social brain: neurobiological basis of affiliative behaviours and psychological well-being. *Neuropeptides, 47*(6), 379-393. https://doi.org/10.1016/j.npep.2013.10.008

About the author

Dr Eleanor Dommett
KING'S COLLEGE LONDON
https://twitter.com/EllieJane1980?ref_src=twsrc%5Egoogle%7Ctwcar
https://www.linkedin.com/in/eleanor-dommett-33193011a/?originalSubdomain=uk

Dr Ellie Dommett studied psychology at Sheffield University. She went on to complete an MSc Neuroscience at the Institute of Psychiatry before returning to Sheffield for her doctorate, investigating the superior colliculus, a midbrain multisensory structure. After a post-doctoral research post at Oxford University she became a lecturer at the Open University before joining King's College London, where she is now a Reader in Neuroscience. She conducts research into Attention Deficit Hyperactivity Disorder, focusing on identifying novel management approaches.

8.

FROM PHYSICAL INJURY TO HEARTACHE: SENSING PAIN

Dr Eleanor J. Dommett

> *There are wounds that never show on the body that are deeper and more hurtful than anything that bleeds.*
>
> Laurell K. Hamilton, Novelist

The opening quote illustrates the complexity of pain. When we think of pain in simple terms, we might think of it as the experience that arises from a cut or bruise to the body. It is, after all, one of the bodily senses that comes under the banner of somatosensation. This simple idea is correct, and it is our starting point in this section, but it does not fully encompass the experience of pain, as you will soon learn. Therefore in this section we will discuss three different types of pain, beginning with **nociceptive** pain, which refers to the kind of pain that arises from a bodily injury.

Nociception: detecting bodily injury

In the previous section you learnt about four different types of modified neurons which act as sensory receptor cells responsible for our sense of touch. There is also a fifth type of modified neuron in the skin which is responsible for detecting tissue damage, and this is called the nociceptor. Unlike the touch receptors, nociceptors do not have any associated structures such as capsules: instead, they are referred to as **free nerve endings**. On the surface of the free nerve endings there are different types of channels, which means they can detect different types of stimuli.

Keeping in mind the idea of damage to the body, as opposed to heartbreak or other types of pain, what type of stimuli might be detected by nociceptors?

You could have come up with a range of ideas here. In the opening paragraph we mentioned cuts and bruises, so you may have identified stimuli that cause damage to tissue or apply great pressure. You

might also have noted that extreme temperatures can cause pain, or some chemical substances. All of these would have been correct.

For each type of stimulus that causes bodily damage, the nociceptor must detect the stimulus and produce a receptor potential through the process of **transduction**.

This process varies according to the stimulus type. The first type of noxious stimulus that can be detected is intense pressure, for example, pinching or crushing. This type of stimulus is detected by mechano-nociceptors. The process of transduction here is very similar to that which was outlined for touch receptors.

Which kind of ion channels open in the nerve endings of touch receptors to produce a receptor potential?

Mechano-sensitive ions channels open, allowing

sodium ions into the nerve ending causing a depolarising receptor potential.

Another type of transduction occurs when tissue is damaged. The damage results in cell membranes being ruptured, so that the chemical constituents of a cell which are typically found within the intracellular space spill into the extracellular space.

Can you identify an ion which is normally found in high concentration inside neurons but at a lower concentration outside?

Potassium

Substances that can be released from the inside of the cell include potassium ions, which are critical to the function of neurons, but there are other ions. For example, hydrogen ions also increase in the extracellular space when tissue damage occurs, causing a decrease in pH and an increase in acidity.

Substances such as bradykinin and prostaglandins can also be released from damaged cells.

These chemicals directly increase as a consequence of tissue damage, but there are also other indirect changes which impact on the chemical constituents of the extracellular space. When tissue is damaged the immune system makes a response to protect the body. This response includes release of several chemicals in the area: histamine, serotonin, and adenosine triphosphate (ATP). All these changes give the nociceptors plenty to detect! Some of the substances directly activate the nociceptor (e.g., potassium and bradykinin) whilst others sensitise them (e.g., prostaglandins).

The final type of stimuli that can activate nociceptors are those that are very hot (>45°C) or very cold (<5°C). Transduction of these stimuli depends on heat-sensing channels. When these channels detect hot or cold stimuli, they open and allow both sodium and calcium ions into the nociceptor. The consequence of this is depolarisation in the form of the receptor potential.

What would you expect to happen after the receptor potential is induced?

If the receptor potential is large enough, an action potential will be triggered.

As you might expect, the receptor potential can trigger an action potential in the nociceptor and this information can then be transmitted to the central nervous system. Before we look at the pathway to the spinal cord and the brain, it is helpful to note a few features of nociceptive pain.

Firstly, although our starting point here was nociceptors being located in the skin, they are in fact found throughout the body. The only two areas where they are not found are inside bones and in the brain. The latter explains why brain surgery can be conducted with awake patients, as described in the previous section on touch. Other organs, such as the heart, lungs or bladder do have nociceptors and activation of these is referred to as **visceral pain**. Visceral pain is a special type of nociceptive pain. It is much rarer than the typical nociceptive pain that arises from our muscles, skin or joints, which is technically referred to as **somatic pain**. The rarety of visceral pain contributes to an interesting phenomenon called **referred pain** (Box 3).

Secondly, nociceptors can alter their sensitivity following tissue damage, resulting in **hyperalgesia** or **allodynia**.

Hyperalgesia refers to an increased sensitivity following injury (Gold & Gebhart, 2010).

From an evolutionary perspective, why might hyperalgesia be beneficial?

It is likely to support greater period of rest to allow recovery.

Hyperalgesia is considered to be part of 'sickness behaviour', that is, the behaviour that we have evolved to allow any infection or illness to run its course (Hart, 1998). Allodynia refers to nociceptors becoming sensitive to non-noxious stimuli, for example, a gentle touch, after injury.

The distinction between the two is illustrated in Figure 4.11.

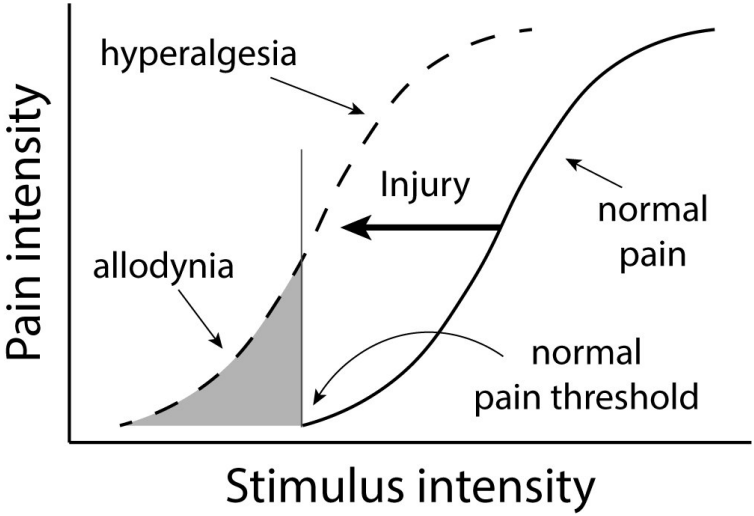

Figure 4.11. The sensitivity of nociceptors can change after injury to result in allodynia or hyperalgesia.

Box 3: Referred pain – what hurts anyway?

Referred pain is an interesting phenomenon where injury to one area of the body creates a perception of pain arising from a different area. There are some

well-known examples of this shown in the table below:

Site of injury	Site of pain perception
Heart	Left arm, shoulder and jaw
Throat	Head
Gall bladder	Right shoulder blade
Lower back	Legs

Table 2. Examples of referred pain

Referred pain can be problematic because it can make it harder for health professionals to diagnose and treat the problem if they are unable to locate the source of the pain. Some examples are common – for example, a heart attack presenting as pain down the left arm and shoulder – meaning that these are easily identified, but for others, it can cause delays to treatment.

There is no consensus on exactly how referred pain arises but it is thought that convergence of information from the visceral site and the somatic site, for example, the heart and the shoulder respectively, onto a single neuron, resulting in an ambiguous signal in the brain. Perception is driven both by the sensory stimulus (bottom-up processing) and the information from memory and past experiences (top-down processing) and when faced with an ambiguous signal, the brain interprets the situation according to what it most expects. We are more likely to have hurt our shoulder than our heart and so the injury to our heart is perceived as a pain in our shoulder.

Pain pathways: getting pain information to the brain

Once an action potential has been produced in the nociceptor the information can be transmitted to the brain. As with the somatosensory neurons responsible for detecting touch, nociceptors also have their cell bodies in the dorsal root ganglion. From here they enter the spinal cord. There are then

two key routes information can take. The simplest route is shown in Figure 12 and shows that the nociceptor connects to an interneuron, which in turn synapses with a motor neuron, forming a reflex arc. This pathway is responsible for the pain withdrawal reflex, for example, moving your hand away from a shard of glass or hot pan handle. The pathway does not travel via the brain and therefore works below conscious awareness to allow a fast response, withdrawing the body from any source of pain.

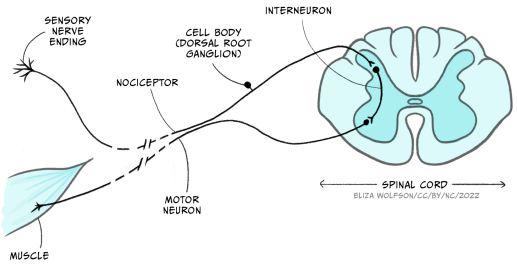

Fig 4.12. The spinal reflex pathway sees the nociceptor axon synapses with an interneuron within the spinal cord, which then excites a motor neuron to withdraw the area of the body from the noxious stimulus.

The remaining pathway travels to the brain and is shown in Figure 4.13. It is this pathway that underpins our conscious

perception of pain, which can only occur when the signal reaches the brain. This pathway is referred to as the **spinothalamic tract**. Nociceptors feeding into this pathway terminate in the superficial areas of the spinal cord and synapse with lamina I neurons, also referred to as transmission cells, which form the second order neuron that crosses the midline and travels up to the thalamus, hence the name spinothalamic.

Using Figure 4.13, can you identify which thalamic nuclei are within the spinothalamic tract?

The ventroposterior lateral nuclei (VPL) and the central laminar nuclei.

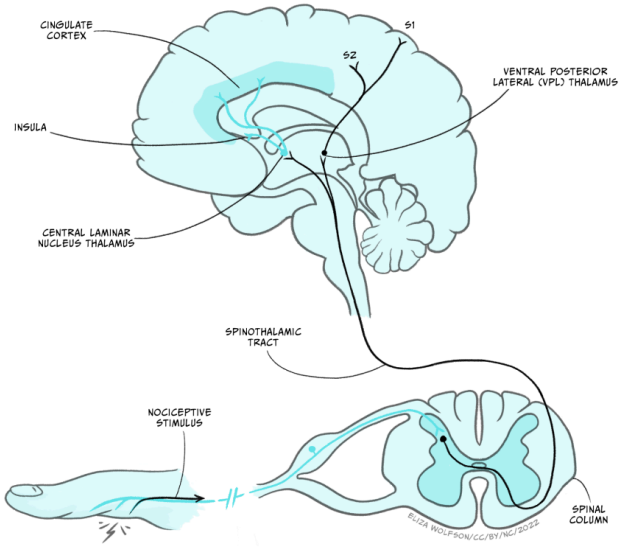

Fig 4.13. The nociceptor synapses with a lamina I neuron in the superficial layers of the spinal cord. This neuron, also referred to as a transmission cell, transmits the signal up to two different areas of the thalamus. From these, the signal is sent to several cortical areas.

From the VPL, third order neurons continue to the secondary somatosensory cortex. Additionally, neurons synapsing in the central laminar nucleus in the thalamus connect with neurons which carry the signal onwards to the insula and cingulate cortex. As with touch information, pain arising from the face region travels separately in the trigeminal pathway.

You may have spotted that the pathways responsible for touch travel up the spinal cord ipsilaterally (i.e. on the same side at the sensory input) and cross the midline in the

brainstem, whilst the pathways for nociception travel up contralaterally (i.e. on the opposite side), having crossed the midline in the spinal cord. This gives rise to an unusual condition called **Brown-Séquard Syndome** (Box 4).

Box 4: Brown-Séquard Syndome

Brown-Séquard Syndome is named after the Victorian scientist (Figure 4.14) who first described and explained a rare spinal injury, presenting his case study at the British Medical Association's annual meeting in 1862.

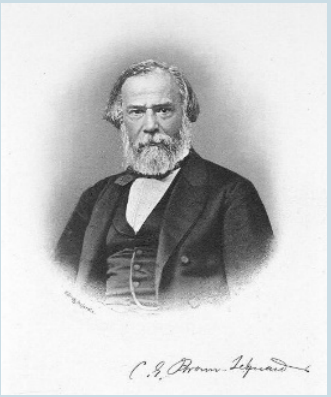

Figure 4.14. Charles-Édouard Brown-Séquard, the Victorian scientist who first discovered the condition named after him caused by damage to the spinal cord.

In the case study he presented the syndrome, characterised by loss of touch on one side of the body and loss of pain sensation on the other, was caused by a traumatic injury (Shams & Arain, 2020). However, the syndrome can, albeit less commonly, arise through non-traumatic injury, such as multiple sclerosis or decompression sickness. The syndrome occurs due to incomplete damage to the spinal cord, such that only one side is severed. The unusual pattern of lost sensation arises because touch information and pain information ascend on different sides of the midline within the spinal cord. Touch ascends ipsilaterally to the stimulus whilst pain ascends contralaterally. When one side of the spinal cord is cut, touch sensation is lost on the same side as the injury and pain on the opposite side.

Prognosis for individuals with Brown-Séquard

Syndome varies depending on the cause and the extent of the damage, but because the syndrome only sees partial damage to the spinal cord, it is possible to have significant recovery, provided complications such as infections can be avoided.

Clearly, it is very important that information about bodily injury can reach the brain, but pain is also an unpleasant sensation, which is not always beneficial.

Can you think of a situation when it is helpful not to be aware of, or focus, on a bodily injury you are experiencing?

You could have come up with lots of ideas here. Perhaps the most obvious one is when your survival depends on being able to mobilise. Battlefield injuries are associated with this situation, where an

individual reports not being fully aware of their injuries until they are away from the frontline. Similar reports can be found for people in sports matches.

The fact that there are situations when pain can be modulated suggests that it is not a simple case of the nociceptor sending an uninterruptible signal to the brain. In fact you have already learnt about the one type of input to the spinothalamic tract that can interrupt the signal about a noxious stimulus before it reaches the brain.

Think about when you knock your elbow, head or knee on something. What do you typically do immediately, without thinking?

The most common reaction here is to rub the site that you have knocked. That is, provide touch stimulation. You often see this when young children

fall or hurt themselves, with the caregiver 'rubbing it [the injured site] better'.

A key explanation of why touching the site of injury may provide pain relief comes from the **Gate Control Theory**. This theory was first put proposed by Ronald Melzack and Patrick Wall in 1965 and describes how pain perception can be modulated by touch. In their theory, Melzack and Wall suggest that there is a gating mechanism within the spinal cord which, when activated, results in a closing of the gate, and can prevent pain signals reaching the brain.

Recall that in the spinal cord the nociceptor synapses with lamina I cells, also called transmission cells. However, other sensory inputs also enter the spinal cord, in the form of touch sensory receptor neurons. According to Melzack and Wall, activation of the touch receptors can result in the signal from the nociceptor being blocked.

Figure 4.15 shows the proposed circuitry for this. In the figure you can see the touch receptor and the nociceptor. Both are synapsing with the lamina I transmission cell and another neuron in lamina II of the spinal cord. This small lamina II neuron, called an interneuron, is critical because it is thought to act as the gate due to its ability to inhibit the transmission

cell. When the nociceptor alone is activated, Wall and Melzack proposed that it will inhibit the lamina II cell and simultaneously excite the transmission cell. The effect of inhibiting the lamina II cell is to remove the inhibition that cell normally exerts on the transmission cell. In effect this results in direct excitation of the transmission cell by the nociceptor, and indirect dishibition, that is, removal of all inhibition, on the same cell through the interneuron. This means the transmission cell is excited and a signal reaches the brain causing the perception of pain.

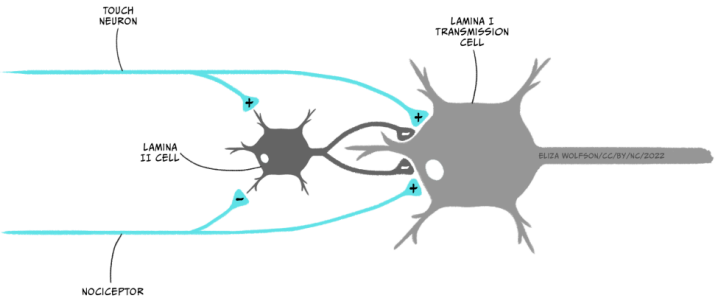

Fig 4.15. The original Gate Control Theory proposed that there was a gate in the spinal cord in the form of a lamina II interneuron which could receive excitatory (+) and inhibitory (-) inputs from different types of sensory input and could inhibit the lamina I transmission cell.

If the touch receptor neuron is also excited, this has the effect of exciting the inhibitory interneuron in lamina II, which results in the inhibition of the transmission cell directly – in effect a two-pronged attack on transmission cell excitation, reducing the likelihood of an action potential being produced and the signal about pain reaching the brain.

The Gate Control Theory is just a theory, and like all theories there is evidence both for and against the theory being accurate. For example, lamina II interneurons have been found to contain GABA, an inhibitory neurotransmitter, supporting the theory. However, nociceptors have so far only been found to form excitatory synapses, against the proposed theory. Although elements of this theory may be inaccurate, it has proved highly influential.

We indicated above that you have learnt about the first way that pain signals can be modulated when we described the role of touch. The other way pain signals can be altered is by actions of the brain itself.

Pain pathways: descending control of pain from the brain

Within the brainstem there are several structures which can modulate our experience of pain. We will begin with the periaqueductal grey (PAG), which is also referred to as the central grey. This structure is activated by activity in the spinothalamic tract and research has shown that electrical

stimulation of the PAG results in a powerful pain-relieving or analgesic effect (Fardin et al., 1984). It is here that some endogenous opioids are thought to act.

This analgesia is thought to occur because of the connections between the PAG and two structures called the locus coeruleus and raphe nuclei. The locus coeruleus contains noradrenergic neurons and the raphe nuclei contains serotoninergic neurons. It is thought that both of these neurons send signals down to the spinal cord to the lamina II interneurons, which in turn suppress lamina I transmission cells that form the spinothalamic tract.

What would be the impact of suppressing the spinothalamic tract?

Reduced activity in this pathway would reduce the amount of activity reaching the thalamus and other areas of brain, minimising our perception of pain.

This descending pathway is an example of negative feedback. A noxious stimulus causes excitation in the spinothalamic tract which in turn activates the PAG. The PAG then sends a signal

down to the spinal cord to interrupt, or silence, the incoming signal from nociceptors.

We have already discussed some situations where it is beneficial to not experience the full unpleasantness of pain, for example, in the case of a battlefield injury or where survival depends on being able to focus on getting to safety. It is likely this pathway from the PAG is critical in these situation. However, there are other, more everyday situations, where it is helpful not to focus on pain and therefore to have a way of modulating the experience.

Think back to the last time you did any of these things: a) went to the dentist for a filling b) had a vaccination by an injection or c) had a piercing or tattoo. How did you manage the pain associated with this experience?

You could have come up with all kinds of things here, but one thing they would likely all have had in common is distraction. Your dentist may have placed a poster on the ceiling of a busy image for you to look at, or had music playing. The doctor or nurse

giving the vaccination or your piercer or tattooist will likely have chatted away to distract you.

In these situations you are experiencing attentional analgesia, that is where attention is directed away from the threatening event, resulting in decreased pain perception. Researchers have shown that this kind of analgesia does involve the PAG but also involves several other structures. Oliva et al. (2022) have demonstrated that attentional analgesia involves parallel descending pathways from the anterior cingulate cortex (ACC) to the locus coeruleus and from the ACC to the PAG and onto a region of the medulla, both extending down to the spinal cord.

Before we leave pathways behind, we need to introduce the second type of pain. Recall that we said we would examine three types of pain at the start of this section. The first is the nociceptive pain, that is pain that arises through actual bodily injury. The second type of pain is neuropathic pain. This type of pain arises through damage to the nociceptors and pathways that carry nociceptive information. One example of this is so-called thalamic syndrome. This syndrome arises when the thalamus is damaged, for example, by a stroke. Individuals with this syndrome experience intense burning or crushing pain from any sort of contact with the skin at a specific

location. The pain is neuropathic because there is no actual damage to the location on the body where the sensation is felt but there is damage to the neurons forming the pathways that would typically carry pain information to the brain about this bodily region.

We now turn our attention to the final type of pain, the kind that arises without any damage to the body, including the neurons that process pain. This is **psychogenic pain.**

Pain without physical damage: psychogenic pain

The opening quote to this section indicated that pain can extend beyond physical damage. This idea is in keeping with our own day-to-day experiences and our use of language. For example, we talk of heartache and life events breaking or damaging us. In these cases, there can be no physical injury to the body or the nerves that normally carry nociceptive information, meaning there is no nociceptive or neuropathic pain, and yet, our experiences are best described as painful. This is psychogenic pain and it refers to a type of pain that can be attributed to **psychological factors**.

There is a huge range of psychological factors that could result in feelings described as pain. In the previous paragraph we mentioned heartache which could arise from a relationship breakdown or bereavement, but there are other, perhaps less obvious factors as well. For example, the Social Pain Theory

(MacDonald and Leary, 2005) suggests that being excluded from a social group or desirable interpersonal relationship can cause social pain, due to rejection, which is similar to physical pain. They also suggest that this social pain serves the same purpose as physical pain which is to respond to any threat to survival including reproduction. Furthermore, there is some evidence that social exclusion and rejection involve similar areas of the brain.

Researchers used functional magnetic resonance imaging (fMRI) to demonstrate overlap between the areas involved in physical pain and the experience of social exclusion. Eisenberger and colleagues asked people to play a virtual ball game whilst having their brain scanned. They found that the anterior cingulate cortex was more active when participants were excluded from the game and that this activity was positively correlated with the self-reported distress felt by participants (Eisenberger et al., 2003). Recall that this area of the cortex is also activated by physical pain. Future studies went on to demonstrate paracetamol, which can provide relief from physical pain, can also decrease activity in this region and the perceived social pain felt (De Wall et al., 2010).

This provides an important link to our last section on pain – its treatment.

Treating pain: medication and beyond

It would not be appropriate to talk about pain without discussing pain treatment. Whilst some pain will resolve without treatment, other pain will require treatment or management. The importance of pain treatment is illustrated by examining the consequences of not treating pain. Failure to treat chronic pain, that is a pain persisting for more than three months, can result in altered mood, mental health disorders, cognitive impairments, sleep disruption and, overall, a reduction in quality of life (Delgado-Gallén et al., 2021).

You have now read about three different types of pain: nociceiptive (encompassing somatic and visceral), neuropathic and psychogenic. It is probably not surprising to learn that with such a range of pain experiences, there is no single treatment that will be effective for all types of pain in all individuals. Additionally, psychogenic pain is rarely treated by healthcare professionals, although the underlying psychological factors may be addressed through talking therapies.

One important consideration when treating pain is whether the pain is acute, for example, from a cut or even broken bone, or chronic, for example, from nerve damage that cannot repair. Some treatments may be effective in the short term, and therefore suitable for acute pain, but not suitable for chronic

pain, for example, because of side effects of long term pain medication.

Treatment of acute pain is typically the most straightforward and is often achieved with drug treatments. These can be categorised according to where in the pain pathway they act:

- **Acting at the sensory nerve ending**: Medicines such as non-steroidal anti-inflammatory drugs (NSAIDS e.g., ibuprofen) act at the sensory nerve ending of nociceptors to block the sensitization of nociceptors by prostaglandins.
- **Acting on the nociceptor axon**: Medicines such as local anaesthetics (e.g., lidocaine) act to block sodium channels in the cell membrane, prevent depolarisation and, therefore, action potentials.
- **Acting in the spinal cord**: Medicines such as opioids, gabapentin and ketamine act in the spinal cord, likely through a range of mechanisms.
- **Acting in the brain**: Opioids may also act on the brain in the thalamus and sensory cortex, along with antidepressant drugs. These drugs can also alter mood meaning the pain may continue but its impact is reduced.

Some of these drugs may also be used to treat chronic pain but consideration needs to be given to side effects. For example,

long term use of NSAIDS is associated with stomach problems, and long term use of opioids comes with the risk of addiction. Decisions about long term drug use must therefore be made carefully and on an individual basis. For example, long term opioid use may be deemed appropriate where the pain is due to a terminal condition.

Other treatments that can be used for acute or chronic pain include stimulation techniques such Transcutaneous Electrical Nerve Stimulation or TENS. TENS machines provide a low-voltage electrical stimulation to the site of pain. It is thought that this low level of stimulation, activates the touch receptors, which, as you should recall from the discussion of the Gate Control Theory, could in turn reduce perceived pain. Although TENS is not typically used for acute injuries, it is sometimes used to treat the acute pain during labour and period pains. A systematic review of the literature on labour pains, which included data from 1671 women found little difference in the pain perceived by women receiving TENS compared to those in control groups, not receiving TENS (Dowswell et al., 2009). However, results for period pains are more positive with results from 260 individuals indicating that when compared to a sham TENS condition (i.e., the machine is attached but not switched on), TENS provided significant pain relief (Arik et al., 2022).

Studies into the effectiveness of TENS in chronic pain have looked at a range of conditions. For example, a review of the literature investigating osteroarthritis in the knee, a condition

which affects 16% of individuals over 15 years of age worldwide (Cui et al., 2020), found TENS to be effective at reducing pain and improve walking ability (Wu et al., 2022).

Surgical approaches may also be taken to treating chronic pain. Clearly any surgery carries risks and therefore this type of treatment is only used in extreme cases. One situation in which surgical approaches may be used is in the treatment of intractable pain found in up to 90% of individuals with terminal cancer. In this situation, surgery may be deemed an appropriate treatment. The most common types of surgery conducted are cordotomy and myelotomy (Bentley et al., 2014). In the cordotomy surgeons cut the spinothalamic tract on one side of the spinal cord.

If only one side of the spinal cord has the spinothalamic tract cut, would pain from both sides of the body be reduced?

No, only pain from the contralateral side of the body as the spinothalamic tract cross the midline immediately on entering the spinal cord.

A cordotomy is a suitable treatment for unilateral pain, that is, pain on one side of the body. In a myelotomy, the surgeons

cut at the middle of the spinal cord, again targeting the spinothalamic neurons, this time at the point they cross.

As you will likely have gathered, chronic pain typically requires a multifaceted approach which may include psychological interventions including cognitive behavioural therapy. This kind of multifaceted treatment is typically delivered at pain management clinics where individuals are supported by a team of professionals including pain consultants, physiotherapists, psychologists and occupational therapists. Such clinics keep the individual at the centre of treatment and they are active in their pain management, with the view to educating them about their pain and finding suitable, but often minimal, analgesic requirements.

The exact cause of the chronic pain will determine, in part, how successfully it can be treated. One type of chronic pain that is still considered very hard to treat, even with a multifaceted approach, is **phantom limb pain**. This type of pain consists of ongoing painful sensations that appear to be coming from part of the limb that is no longer there. This can occur in up to 80% of amputees (Richardson & Kulkarn, 2017).

However, the name 'phantom limb' is actually quite misleading because these kind of painful sensations are not limited to missing limbs. Up to 80% of patients who have had a mastectomy (a breast removed), typically for the treatment of breast cancer, may experience both non-painful and painful sensations arising from the missing breast (Ramesh &

Bhatnagar, 2009). Exactly why phantom pain happens is still not fully understood, but is likely to be due to changes in how the nervous system is wired or connected following the amputation or mastectomy. A review of studies investigating treatments for phantom limb pain by Richardson and Kulkarn (2017) found that over 38 different therapies had been investigated including a range of drug treatments and transcutaneous magnetic stimulation (TMS), a technique similar to TENS, applying a magnetic pulse instead of an electrical one, and mirror therapy (Box 5). They concluded that despite the range of therapies test, results were insufficient to support use of any of these treatments.

Box 5: Novel approaches to pain management: Mirror Therapy

This novel treatment was first described by neuroscientist Vilayanur Ramachandran. In this treatment the patient positions a mirror box between their intact limb and the missing limb. They then look into the mirror to see a reflection of their intact limb, creating a visual representation of the

missing limb (Figure 16). They can then make movements with their intact limb whilst looking at the reflection. This movement can create the perception of individual re-gaining control over the missing limb. Where the pain arises from a clenched or cramped feeling in the phantom limb, movement of the intact limb to a different position, could relieve pain.

Figure 4.16. Mirror therapy relies on the visual representation of an intact limb where the amputated limb should be.

All the treatments available for pain could, in part, have their effects attributed to the placebo effect. This is where an individual gains some benefit without receiving any real

treatment. The placebo effect is not specific to pain treatment, it can occur in treatment for any condition. In the context of pain, the placebo effect could be responsible if an individual gains pain relief from swallowing a tablet, even if that tablet had no impact on pain processing, or from being connected to a TENS machine, even if it is not switched on. The placebo effect is a complicated phenomenon (see also the chapter Placebos: a psychological and biological perspective); there are several reasons it might occur, including (Perfitt & Plunkett, 2020):

- **Conditioned behaviour**: people learn to associate pain relief with taking a tablet or receiving and injection so if when it is an inert or inactive substance, they experienced a conditioned response of pain relief.
- **Expectation**: people expect to get better after seeing a doctor or receiving treatment and so experience pain relief because of this expectation.

Given we know that pain perception can be modulated by descending pathways from the brain, either of these top-down mechanisms are plausible.

It is also important to recognise that other effects could occur which are mistaken for the placebo effect. For example, people may just get better over time because that is the natural trajectory of their condition, meaning there is no placebo effect, they just recovered. There is also an effect called the

Hawthorne effect, which refers to the fact that simply observing people in an experiment or trial will change their behaviour. For example, in a study on drug treatment for osteoarthritis, those who are part of the trial may be more likely to complete recommended exercises than those who are not and so may experience pain relief from a placebo drug treatment, not because of the placebo effect but because they are mobilising the joint more and regularly providing an account of their behaviour, in comparison to individuals not part of a trial.

What do you think the placebo effect, or even the Hawthorne effect, means for clinical trials trying to test the effectiveness of new treatments?

These trials need to be very carefully designed to ensure that the group of people receiving the new treatment are compared to an appropriate control group. For example, it might be appropriate to have a TENS group, a sham TENS group and a third group on a waiting list who are assessed but do not receive a real or placebo treatment.

We have now reached the end of our exploration of the somatosensory system, covering touch and pain.

Key takeaways

In this section you have learnt:

- Nociceptive pain arises when there is actual physical injury to the body and it is detected by nociceptors capable of responding to mechanical, chemical and thermal signals
- Nociceptive pain can be divided into pain arising from the muscles, skin or joints, called somatic pain, and pain arising from the internal organs, which is called visceral pain. We can sometimes struggle to identify the location of visceral pain and misattribute it to somatic pain, a phenomenon known as referred pain
- Nociceptors can alter their sensitivity giving rise to hyperalgesia and allodynia. Both of these may serve to the protect the body to

allow any injury or damage to pass

- When nociceptive information enters the spinal cord it can form a reflex arc with a motor neuron or be transmitted up to the brain via the spinothalamic tract. From the thalamus, information about noxious stimuli is sent onto various cortical areas
- The Gate Control Theory proposes that signals in the spinothalamic tract can be blocked by activation of lamina II interneurons in the spinal cord, which are activated by touch
- Descending control of pain, by areas such as the PAG and the anterior cingulate cortex can also exert a powerful methods of pain control
- Neuropathic pain arises when the pathways which process pain information are damaged, creating a perception of pain in the absence of damage to that body part
- The final type of pain is psychogenic pain, that is pain arising from psychological factors such as relationship breakdown or social exclusion. Imaging from brain scanning suggests the experience of psychogenic pain activates similar areas of the brain to physical pain

- Pain treatment focuses on nociceptive and neuropathic pain and can include a range of approaches including drug treatment, stimulation approaches, surgery and psychological therapy. Several treatments may be combined and delivered by specialised clinics where the pain is chronic.

References

Arik, M. I., Kiloatar, H., Aslan, B., & Icelli, M. (2022). The effect of TENS for pain relief in women with primary dysmenorrhea: A systematic review and meta-analysis. *Explore*, *18*(1), 108–113. https://doi.org/10.1016/j.explore.2020.08.005

Bentley, J. N., Viswanathan, A., Rosenberg, W. S., & Patil, P. G. (2014). Treatment of medically refractory cancer pain with a combination of intrathecal neuromodulation and neurosurgical ablation: case series and literature review. *Pain Medicine*, *15*(9), 1488–1495. https://doi.org/10.1111/pme.12481

Cui, A., Li, H., Wang, D., Zhong, J., Chen, Y., & Lu, H.

(2020). Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. e*ClinicalMedicine*, *29*, 100587. https://doi.org/10.1016/j.eclinm.2020.100587

Delgado-Gallén, S., Soler, M. D., Albu, S., Pachón-García, C., Alviárez-Schulze, V., Solana-Sánchez, J., Bartrés-Faz, D., Tormos, J. M., Pascual-Leone, A., & Cattaneo, G. (2021). Cognitive reserve as a protective factor of mental health in middle-aged adults affected by chronic pain. *Frontiers in Psychology*, *12*, 752623. https://doi.org/10.3389/fpsyg.2021.752623

DeWall, C. N., MacDonald, G., Webster, G. D., Masten, C. L., Baumeister, R. F., Powell, C., Combs, D., Schultz, D. R., Stillman, T. F., Tice, D. M., & Eisenberger, N. I. (2010). Acetaminophen reduces social pain: Behavioral and neural evidence. *Psychological Science*, *21*(7), 931–937. https://doi.org/10.1177/ 0956797610374741

Dowswell, T., Bedwell, C., Lavender, T., & Neilson, J. P. (2009). Transcutaneous electrical nerve stimulation (TENS) for pain relief in labour. *The Cochrane database of systematic reviews*, (2), CD007214. https://doi.org/10.1002/14651858.CD007214.pub2

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Advancement of Science*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Fardin, V., Oliveras, J. L., & Besson, J. M. (1984). A

reinvestigation of the analgesic effects induced by stimulation of the periaqueductal gray matter in the rat. II. Differential characteristics of the analgesia induced by ventral and dorsal PAG stimulation. *Brain Research, 306*(1-2), 125–139. https://doi.org/10.1016/0006-8993(84)90361-5

Gold, M. S., & Gebhart, G. F. (2010). Nociceptor sensitization in pain pathogenesis. *Nature Medicine, 16*(11), 1248–1257. https://doi.org/10.1038/nm.2235

Hart, B. L. (1988). Biological basis of the behavior of sick animals. *Neuroscience and Biobehavioral Review, 12*(2), 123–137. https://doi.org/10.1016/S0149-7634(88)80004-6

Macdonald, G., & Leary, M. R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, *131*(2), 202–223. https://doi. org/10.1037/0033-2909.131.2.202

Melzack, R., & Wall, P. D. (1965). Pain mechanisms: a new theory. *Science*, *150*(3699), 971–979. https://doi.org/10.1126/science.150.3699.971

Oliva, V., Hartley-Davies, R., Moran, R., Pickering, A. E., & Brooks, J. C. (2022). Simultaneous brain, brainstem, and spinal cord pharmacological-fMRI reveals involvement of an endogenous opioid network in attentional analgesia. *eLife*, *11*, e71877. https://doi.org/10.7554/eLife.71877

Perfitt, J. S., Plunkett, N., & Jones, S. (2020). Placebo effect

in the management of chronic pain. *BJA Education*, *20*(11), 382–387. https://doi.org/10.1016/j.bjae.2020.07.002

Ramesh, Shukla, N. K., & Bhatnagar, S. (2009). Phantom breast syndrome. *Indian Journal of Palliative Care*, *15*(2), 103–107. https://doi.org/10.4103/0973-1075.58453

Richardson, C., & Kulkarni, J. (2017). A review of the management of phantom limb pain: challenges and solutions. *Journal of Pain Research*, *10*, 1861–1870. https://doi.org/10.2147/JPR.S124664

Shams, S., & Arain, A. (2020). Brown Sequard Syndrome. *StatPearls [Internet]*. https://www.ncbi.nlm.nih.gov/books/NBK538135/

Wu, Y., Zhu, F., Chen, W., & Zhang, M. (2022). Effects of transcutaneous electrical nerve stimulation (TENS) in people with knee osteoarthritis: A systematic review and meta-analysis. *Clinical Rehabilitation*, *36*(4), 472–485. https://doi.org/10.1177/02692155211065636

About the author

Dr Eleanor Dommett
KING'S COLLEGE LONDON
https://twitter.com/EllieJane1980?ref_src=twsrc%5Egoogle%7Ctwcan
https://www.linkedin.com/in/eleanor-dommett-33193011a/?originalSubdomain=uk

Dr Ellie Dommett studied psychology at Sheffield University. She went on to complete an MSc Neuroscience at the Institute of Psychiatry before returning to Sheffield for her doctorate, investigating the superior colliculus, a midbrain multisensory structure. After a post-doctoral research post at Oxford University she became a lecturer at the Open University before joining King's College London, where she is now a Reader in Neuroscience. She conducts research into Attention Deficit Hyperactivity Disorder, focusing on identifying novel management approaches.

9.

LIGHTING THE WORLD: OUR SENSE OF VISION

Dr Eleanor J. Dommett

Just by seeing is believing, I don't need to question why.
Sung by Elvis Presley. Lyrics by Red West & Glen Spreen

The song lyric above from 'Seeing is Believing', made famous by Elvis Presley, encapsulates the power we give our sense of vision. This lyric is one of many examples in our language which indicates how important we consider vision to be. For example, phrases such as 'I see', intended to mean that we understand, or 'A picture paints a thousand words,' rely on the metaphor of vision. Similarly, in business and industry, organisations typically have vision statements, which outline what they want to achieve. All these phrases point to the importance of vision in our everyday lives. In keeping with our approach to the other senses, we will now begin our journey to understanding vision, with the signal that reaches our senses – the visual stimulus.

Light: the wave and the particle

The signal detected by the visual system is light that is either reflected from a surface or emitted from a source, such as a light bulb or natural sources of light like the sun. We can detect light ranging in intensity or luminance, measured in candela per metre squared (cd m^{-2}), from 10^{-6} to 10^{8} cd m^{-2}. To give some context to this, this incorporates everything from a dimly lit night sky to the sun. A typical computer screen, like the one you may be reading from now, has a luminance of 50-300 cd m^{-2}. The light we can detect is just a small part of electromagnetic spectrum (Figure 4.31).

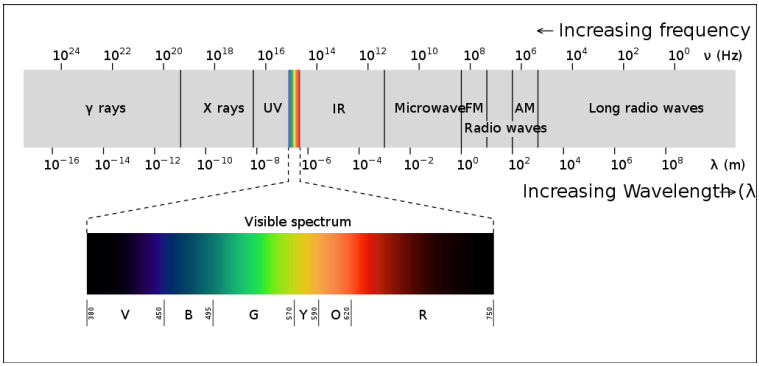

Fig 4.31. The electromagnetic spectrum includes visible light which can be detected by our visual system

This electromagnetic spectrum includes other signals you may be familiar with, such as radio waves, X-rays and microwaves, but the visible light spectrum spans the wavelengths of

380-780 nm, which corresponds to a frequency range of 7.9 x 10^{14} – 3.8 x 10^{14} Hz (790000000000000 – 380000000000000 Hz), what we see as the colours from violet to red.

Looking at Figure 4.31, which wavelength and frequency is associated with violet?

Violet is at the end of the visible light spectrum with a wavelength of 380nm and a frequency of 7.9 x 10^{14} Hz.

Waves in this spectrum are **transverse waves** and consist of simultaneous variations in electrical and magnetic fields at right angles to each other (Figure 32). Unlike the sound waves you learnt about for hearing, electromagnetic waves do not require a medium to be transmitted, so light can travel through a vacuum.

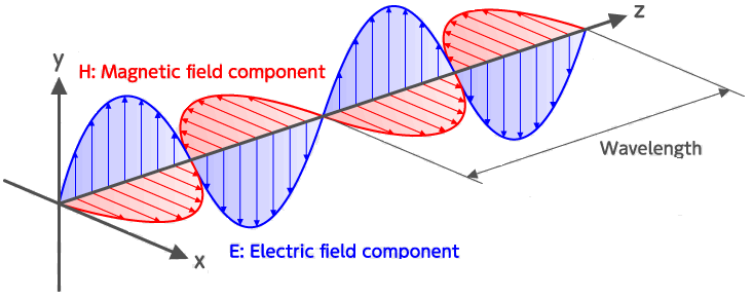

Fig 4.32. The electromagnetic wave consists of fluctuating and perpendicular electrical and magnetic fields

The perceptive amongst you will have spotted that the title to this section indicates that light is not just a wave, but also a particle. This subheading hints at a fierce scientific debate between some of the most famous scientists in history – a *Who's Who* of the Royal Society. Isaac Newton believed that light was made up of particles, referred to as photons, whilst his rival Robert Hooke believed light was a wave. Over time, experiments and calculations by James Clerk Maxwell appeared to prove Hooke right, and the once-fierce debate was calmed.

However, the discovery of the photoelectric effect (see Box: The photoelectric effect, below) reignited this debate and drew the attention of Albert Einstein. He proposed a new theory: that light was not made of waves or particles, but *both* – light was made of wave packets or photons. This idea was elaborated on to show that some experimental findings are best explained when we conceive light as a wave, whilst others work best when

it is described as a particle, and others still can work with either explanation. It was this work that led to Einstein's Nobel Prize in physics.

The Photoelectric Effect

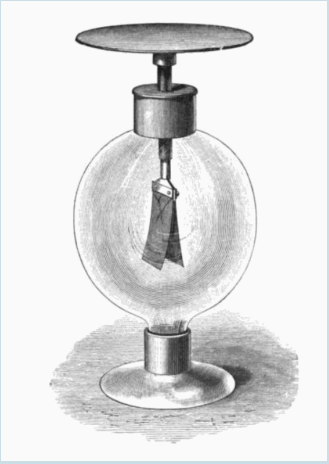

Fig 4.33. A gold-leaf electroscope which can be used to demonstrate the photoelectric effect

The photoelectric effect refers to the emission of electrons from a material when an electromagnetic radiation hits that material. It is best demonstrated when the material is a metal. When an electromagnetic radiation hits a metal, the energy within it can transfer to the electrons within the metal, and if that energy exceeds the binding energy (the energy keeping the electron in the metal), the

electrons can be ejected from the substance. Critically though this will not happen with just any radiation, it only happens with very high energy sources. The energy within electromagnetic spectrum is related to its frequency and wavelength, such that the waves with the highest frequency, and therefore lowest wavelength, have the most energy. This means light at the violet end of the visible spectrum has more energy than light at the red end. The effect can be demonstrated using an electroscope (Figure 4.33). In this gold-leaf electroscope, the two gold leaves hanging down are separated when negatively charged but, if high energy photons are delivered to the plate above, causing the loss of electrons, the leaves fall back together. The fact that this effect only worked for some wavelengths of light was critical in understanding light as both a wave and a particle.

Now that we have examined the nature of the signal in vision, it is helpful to look at how that signal can be detected. From this it might be expected that the next section will focus on transduction, but in the visual system there is quite a lot to do

before we reach the sensory receptor cells for transduction, so we start by looking at the structure of the eye.

From light source to retina: bringing the world into focus

The sense organ of the visual system is the eye and, just like the ear, it is made up of several different parts, all of which play a critical role in ensuring we have accurate vision (Figure 4.34).

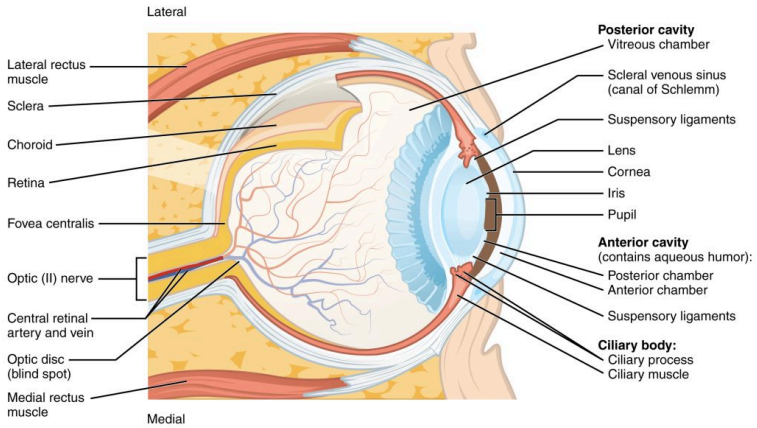

Figure 4.34. The structure of the eye

Figure 4.34 shows that there are several structures which the light must pass through before it gets to the retina, where the sensory receptor cells, referred to as photoreceptors, are located. These structures have a dioptric effect and are referred to as the dioptric apparatus, which simply means that they are

involved in refracting or bending the light to a focal point. Despite there being several structures involved, the refractive power in the eye comes almost entirely from the cornea and the lens. The cornea has a fixed refractive power, but the lens can alter its power by becoming fatter or flatter. When the ciliary muscles contract the lens becomes rounder, increasing its refractive power and the ability to bend the light waves. This allows sources at different distances from the eye to be brought into focus, which means that the light waves are brought to a focal point on the surface of the retina (Figure 4.35).

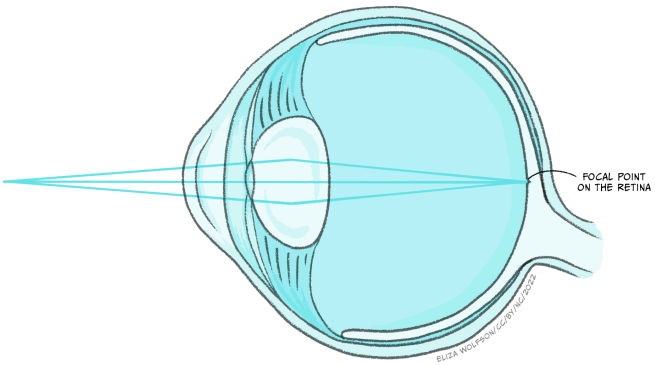

Fig 4.35. The first stage of successful vision is for the light waves to reach a focal point on the surface of the retina.

Despite this process appearing quite simple in comparison to much of what you have learnt in this chapter about the senses, refractive errors are extremely common. There are different types of refractive errors, but you are mostly likely to have heard of:

- **Myopia** or short-sightedness which makes distant objects look blurry
- **Hyperopia** or long-sightedness which makes nearby objects look blurry
- **Presbyopia** which makes it hard for middle-aged and older adults to see things up close

Collectively these conditions are thought to impact 2.2 billion people worldwide (World Health Organisation, 2018), with 800 million people with an impairment that could be addressed with glasses or contact lenses (World Health Organisation, 2021). To correct for these refractive errors requires lenses to be produced that can increase (hyperopia) or decrease (myopia) the overall refraction of light (see Box: Refractive errors, below).

Refractive errors and corrective lenses

Myopia and hyperopia are two of the most common refractive errors. Myopia arises when the refractive power of the eye is too great and the focal point occurs before the retina (Figure 4.36a), whilst hyperopia occurs when the refractive power of the eye is too low, and the image has therefore not been focused by the time it reaches the retina – it would effectively have a focal point behind the retina (Figure 4.36c). To address this, lenses need to be placed in front of the eye in the form of glasses or contact lenses. For myopia, the lens counters the normal refractive power of the eye (Figure 4.36b) whilst for hyperopia it bends the light in the same direction as the eye (Figure 4.36d).

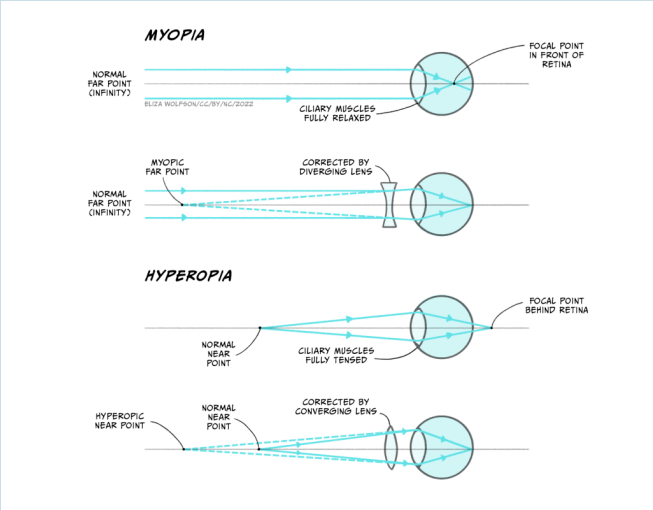

Fig 4.36. The focal point falls in front of the retina in myopia (a) and behind it in hyperopia (c). To correct for this diverging (b) and converging (d) lenses are needed.

Presbyopia typically arises with age and is caused by the gradual hardening of the lens in the eye. As it hardens, flexibility is lost which means that it is difficult to focus. The solution is bifocal or varifocal lenses which have different refractive powers at different positions of the lens.

Assuming that the light waves can be brought to a focal point on the retina, the visual system can produce an unblurred image. As stated above, the retina contains the photoreceptors

that form the sensory receptor cell of the visual system. However, it also contains many other types of cells in quite a complex layered structure (Figure 4.37).

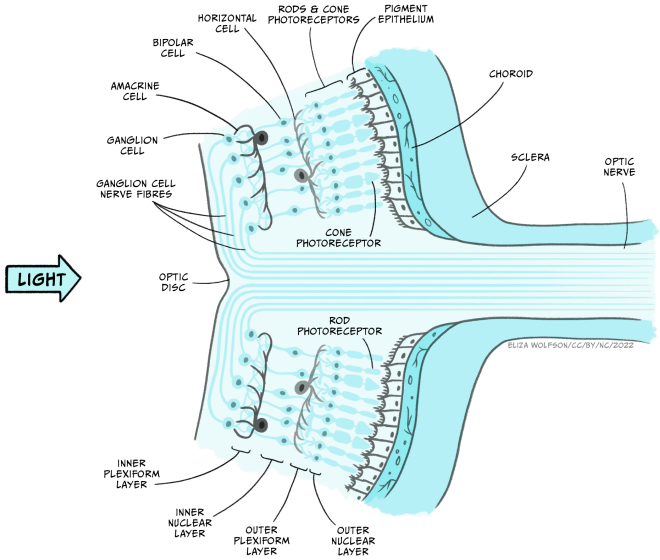

Fig 4.37. The layered cellular structure of the retina

If you examine Figure 4.37 you will see that the photoreceptors form the deepest layer of the retina, that is, the one furthest from the light source. You should also spot that there are two different types of photoreceptors: **rods** and **cones**. These two different types of photoreceptors allow the visual system to operate over a wide range of luminance and wavelength conditions.

Rods outnumber cones by around 20:1 and they are found predominantly in the peripheral area of the retina rather than

the **fovea** or central point of the retina. They are much more sensitive to light than cones, meaning they are suitable for scotopic vision – that is night vision or vision in dimly lit environments. They also provide lower acuity visual information because they are connected in groups rather than singularly to the next type of cell in the retina. This means that the brain cannot be sure exactly which of a small number of rods a signal originated from. There is only one type of rod in the human eye, and it is most sensitive to light with a wavelength of 498 nm.

Look back at Figure 4.31. What colour does this wavelength correspond to?

This corresponds to a green-blue colour.

In contrast to rods, cones are found in a much greater number within the fovea and provide us with high acuity vision due to one-to-one connections with other cells creating very small receptive fields. They are less sensitive than rods and so best suited to **photopic** or day vision. There are, however, three different types of cones, each with different spectral sensitivities (Figure 4.38).

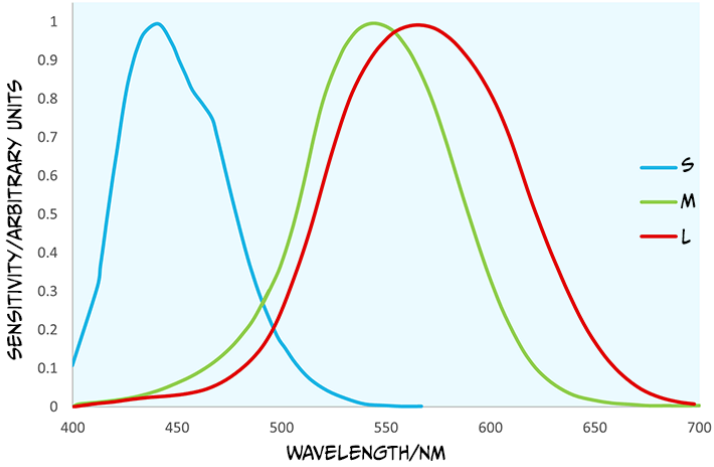

Fig 4.38. The spectral sensitivities of short (S), medium (M) and long (L) wave human cone cells

Although the cones are often referred to as short (S), medium (M) and long wave cones (L), indicative of their wavelength sensitivity, they are sometimes called blue, green and red cones, corresponding to the colours we perceive of the wavelengths that optimally activate them.

Looking at Figure 4.38, what cone would you expect to react if a light with a wavelength of 540 nm was to be detected by the retina?

You would expect to see that the red and the green cone would react because this is within their spectral sensitivity.

In the question above, we asked you about photoreceptors reacting, which leads on to the next stage of our journey through the visual system to look at this process in detail as we turn our attention to transduction.

Photoreceptors and visual transduction

To understand transduction it is helpful to look at the structure of the photoreceptors in a little more detail (Figure 4.39).

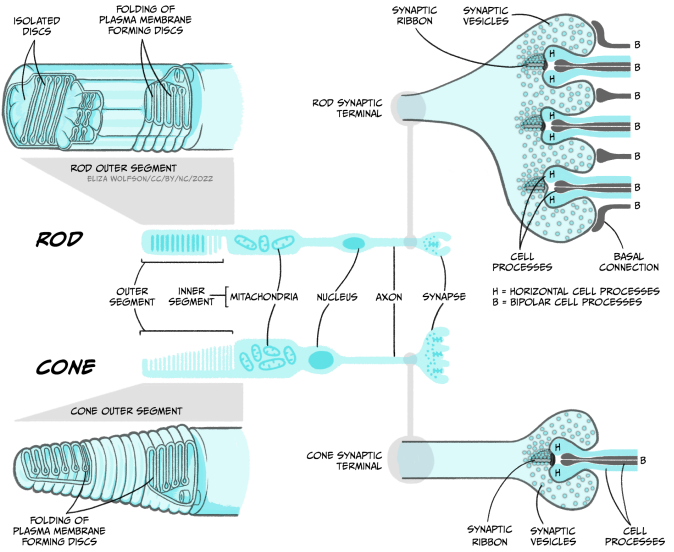

Fig 4.39. The structure of rods and cones

Both rods and cones contain an outer segment which includes a photosensitive pigment which can be broken down by light. The pigment in rods, which is referred to rhodopsin, contains a protein – opsin – attached to a molecule called 11-cis-retinal. The pigment in cones is generally referred to as iodopsin and still consists of opsin and 11-cis-retinal but these are three slightly different opsin molecules that have different spectral sensitivies. The process of phototransduction is similar for all rods and cones. It is described below in detail for rods, referring to rhodopsin rather than the different cone opsins, though the process is analogous in cones.

The process of visual transduction (or phototransduction)

is more complex than the process for touch, pain or hearing so we will need to break this down into a series of steps, but it is also helpful to have an oversight (no pun intended) from the start (Figure 4.40).

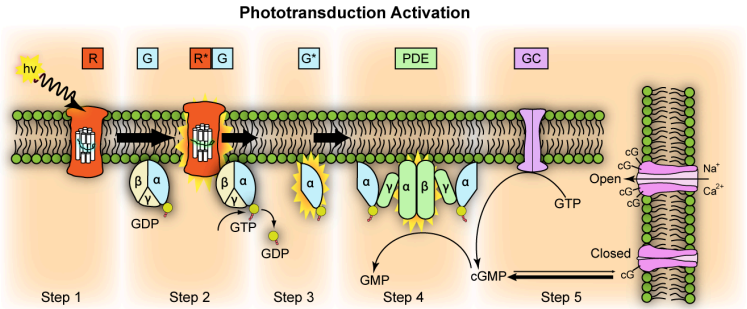

Fig 4.40. The process of transduction in the visual system. R = rhodopsin, R* = activated rhodopsin, G = transducin, G* = activated transducin, PDE = phosphodiesterase, GC – Guanylyl cyclase

The first stage of transduction happens when the energy from a photon of light reaching the retina is absorbed by rhodopsin. The absorption of energy forces the 11-cis-retinal to undergo a transformational change and become all-trans-retinal – this 'activates' rhodopsin.

In the second stage, the newly-activated rhodopsin interacts with a G-protein called transducin. We briefly met G-proteins in the "Neurotransmission" chapter. Like rhodopsin, metabotropic neurotransmitter receptors are G-protein coupled receptors (GPCRs). G-proteins are a group of

proteins which are involved in transmitting signals from outside of a cell to inside it and are so-called as they bind guanine nucleotides. In phototransduction, activation of the G protein transducin by rhodopsin transmits information about the light from outside the photoreceptor to inside it, while in neurotransmission, ligand binding to the receptor activates the G protein to transmit information about neurotransmitter presence at the synapse. In each case the activated G protein releases guanosine diphosphate (GDP) bound to it and instead binds a guanosine triphosphate (GTP) molecule.

In the third stage of transduction, GTP binding to the G protein, transducin, results in the β and γ subunits of transducin dissociating from the α subunit and bound GTP molecule. In the fourth stage, the α subunit and bound GTP interact with a second protein called phosphodiesterase (PDE) which in turn becomes activated.

What would you expect to happen at some point during transduction if a receptor potential is to be produced?

You would need to see ions channels open or close

to allow a change in ions moving across the membrane, carrying the charge that makes up the receptor potential.

The final step sees activated PDE breakdown a molecule called cyclic guanosine monophosphate (cGMP). cGMP is produced by guanylyl cyclase and opens cGMP-gated ion channels in the cell's membrane that allow sodium and calcium ions to enter the cell. Thus when cGMP is broken down by PDE, these ion channels close and no more calcium or sodium can enter the cell. This might not be quite what you expected to happen because in the previous senses we have looked at, transduction involves channels opening and positively charged ions coming into the cell, depolarising the cell. However in light detection, the reverse occurs: light detection causes cessation of a depolarising current and a hyperpolarisation of the membrane. The current that flows when no photons of light are being absorbed is called the 'dark current'. One suggested reason that the visual system operates in this way is to minimise background noise. To explain this further, when there is a dark current, there is a steady flow of sodium into the cell in the absence of light. This means that any minor fluctuations in sodium channel openings will not impact the cell very much – the noise will effectively be ignored. It is only when a large

number of channels close in the light that the cell membrane potential will be affected, giving rise to a clear signal.

In any event, the decrease in intracellular calcium that occurs because of channels closing when light hits the retina results in a reduction in the release of glutamate from the photoreceptor. This in turn impacts on the production of action potentials in the bipolar cells that synapse with the rods and cones. There are broadly two types of bipolar cells – ON cells and OFF cells. OFF bipolar cells respond to the decrease in glutamate release during light stimulation with a decrease in action potential firing, i.e. a decrease in glutamate causes a decrease in excitation and reduced firing. However, ON bipolar cells respond to the decrease of glutamate during light stimulation with an increase in action potential firing. Glutamate is usually excitatory, so how can a decrease in glutamate during light cause an increase in bipolar cell firing? This happens because instead of expressing ionotropic AMPA glutamate receptors, ON cells express a specific metabotropic glutamate receptor, mGluR6. When mGluR6 is activated, its G-protein subunits close a non-specific cation (positive ion) channel, hyperpolarising the cell. When glutamate release is reduced, mGluR6 is inactive, allowing the cation channel to open, and sodium ions to enter the bipolar cell, depolarising it and causing action potentials to fire. Bipolar cells in turn connect to the retinal ganglion cells, whose axons form the optic nerve and transmit action potentials from the eye to the brain.

Visual pathways: to the visual cortex and beyond

The axons of the retinal ganglion cells form the optic nerve and leave the eye through the blind spot. From the optic nerve, two routes that can be taken, a cortical and a subcortical route. The cortical route is the pathway that is responsible for much of our higher processing of visual information and has been the focus of a large amount of research and, as such, it is a logical starting point. Figure 41 shows the route that visual information typically takes from the eye to the primary visual cortex, located in the occipital lobe. In contrast to the pathway from the ear to the primary auditory cortex this pathway looks quite simple, but information is very carefully sorted throughout the pathway.

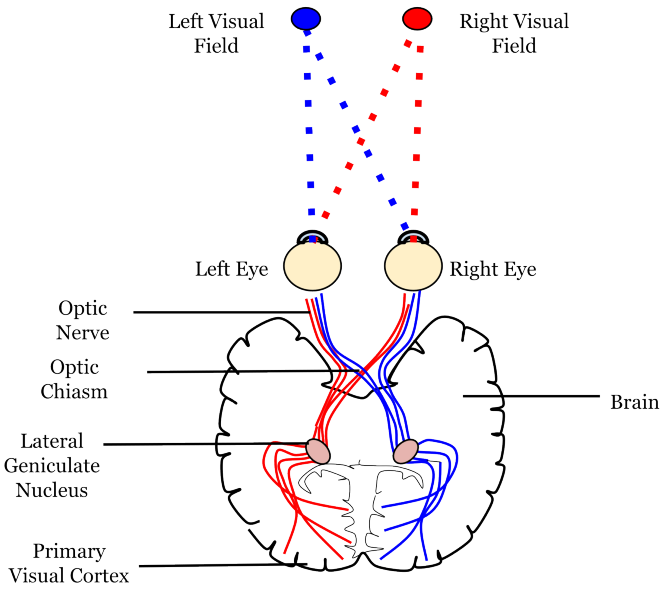

Fig 4.41. The route visual information takes from the eye to the visual cortex

Starting from the eye, information leaves via the optic nerve. The optic nerves from both eyes meet at the optic chiasm which can be seen on the underside of the brain (Figure 4.41). At this point information is arranged such that signals from the left visual field of both eyes continues its pathway via the right side of the brain, whilst information from the right visual field of both eyes travels onwards in the left side of the brain. The first stop in the brain is the lateral geniculate nucleus (LGN) which is part of the thalamus. Each LGN is divided into six layers. Three of these layers receive information from one eye and three receive it from the other. These layers are

said to be retinotopically mapped, which means that adjacent neurons will receive information about adjacent regions in the visual field.

From the LGN, information travels, via the optic radiation, to the primary visual cortex (V1), sometimes also referred to as the striate cortex because of its striped appearance. As we learnt in an earlier chapter (Exploring the brain), the cortex consists of a series of layers from the outside of the brain to the inside. The most dorsal or outer layer is labelled Layer I and the deepest or innermost layer is layer VI. Information from the LGN enters the primary visual cortex in layer IV where different layers of the LGN enter different subsections of layer IV (Figure 4.42).

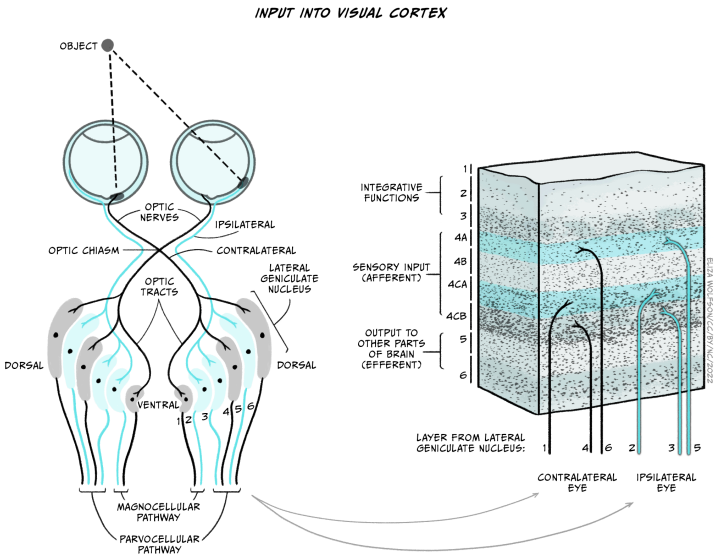

Fig 4.42. The input from the LGN into the primary visual cortex

There will be many thousands of cortical neurons receiving information from each small region of the retina and these cells are organised into columns which respond to specific stimulus features such as orientation. This means that cells in one orientation column preferentially respond to a specific orientation (e.g. lines at 45° clockwise from the vertical) whilst those in the next column will respond to a slightly different orientation. Across all columns, all orientations can be represented. Primary visual cortex can also be divided into columns that respond preferentially to one eye or the other – these are termed 'ocular dominance columns'. Theoretically,

the cortex can be split into 'hypercolumns' each of which contains representations from both ocular dominance columns and all orientations for each part of the visual field, though these do not map as neatly onto the cortical surface as was once theorised (Bartfeld and Grinvald, 1992).

However, despite the exquisite organisation of the primary visual cortex, information does not stop at this point. In fact visual information travels to many different cortical regions – with 30 identified so far.

Can you recollect how auditory information was divided after the primary auditory cortex?

It was divided into a dorsal and ventral pathway.

Visual information can also be divided into a dorsal and ventral pathway. The ventral pathway which includes V1, V2, V4, and further regions in inferior temporal areas is thought to be responsible for object identity i.e., a 'what' pathway. The dorsal stream which includes V1, V2, V3, V5 supports detection of location and visually-controlled movements (e.g., reaching for an object), i.e., a 'where' pathway (Figure 4.43).

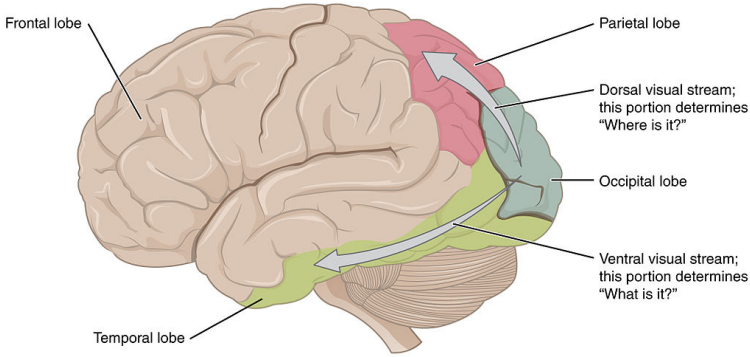

Fig 4.43. The dorsal and ventral streams of the visual system

We mentioned that there is also a subcortical pathway that visual information can take through the brain. In fact there are several different subcortical structures that receive visual information but one of the main ones is a structure call the superior colliculus. This name may sound familiar because you have already learnt about the inferior colliculus in your exploration of hearing. The superior colliculus sits just above the inferior colliculus, on the surface of the midbrain. Although often overlooked when describing visual processing, the superior colliculus is thought to be involved in localisation and motion coding. It has also been implicated in an interesting phenomenon termed Blindsight (see Box below, Blindsight: I am blind and yet I see).

Blindsight: I am blind and yet I see

Blindsight was first described in the 1970s by researchers who had identified residual visual functioning in individuals who were deemed to be clinically blind due to damage to the visual cortex (Pöppel, Held, & Frost, 1973; L Weiskrantz & Warrington, 1974). These individuals reported being unable to see, but could detect, localise or discriminate stimuli that they were unaware of at higher than chance levels (i.e., greater than the levels that would be expected if they were just guessing). Later work allowed a further distinction to be made into blindsight Type 1, where the individual could guess certain features of the stimulus at higher levels than chance e.g., type of motion without any conscious awareness of it, and Type 2 where individuals could detect a change in the visual field but do not develop any perception of that change (L. Weiskrantz, 1997).

Several explanations have been proposed for this interesting phenomenon:

- Areas other than primary visual cortex underlie the responses, including the superior colliculus, which has been shown to provide quick crude responses to visual stimuli.
- Whilst much of the primary visual cortex is destroyed in people with blindsight, small pockets of functionality remain, and this explains the residual abilities.
- The LGN is capable for detecting key visual information and passing this directly to other cortical areas which could explain the phenomenon.

Research continues into blindsight and the role of several brain structures in visual processing, but the existence of this phenomenon has demonstrated that subcortical pathways and structures outside the primary visual cortex can still play a significant role in visual processing.

We have now discussed transduction and pathways for vision but not said much about specific features of the visual scene are detected. As you will probably have guessed this is an extremely complicated process and so we will focus just on three components of the visual scene in the next section: colour, motion and depth.

Perceiving the world: colour, motion and depth

Colour processing is critical to our perception of the world, and you learnt earlier in this section that we have three types of cones with distinct but overlapping spectral sensitivities. These three types of cones in the retina are the start of our colour perception journey. The presence of three types of cones is referred to as trichromacy. The development of trichromacy is thought to offer an evolutionary advantage because it can help identify suitable foods and better discriminate their ripeness. For example, the ability to differentiate red-green is thought to be important as reddish colours in fruits are indicative of higher energy or greater protein content. Work with humans suggest colour remains important in food preferences (Foroni et al., 2016).

The output from the three types of cones is thought to be translated into an opponent colour system in the retinal ganglion cells which can then give rise to specific channels of information in the visual system, referred to as the opponent processing theory of colour processing (Figure 4.43):

- A red-green channel which receives opposing inputs from red and green channels.
- A luminance channel which receives matching inputs from red and green channels.
- A blue- yellow channel which receives excitatory input

from blue channels and inhibitory input from the luminance channel (which in turn is created from excitation from red and green cones).

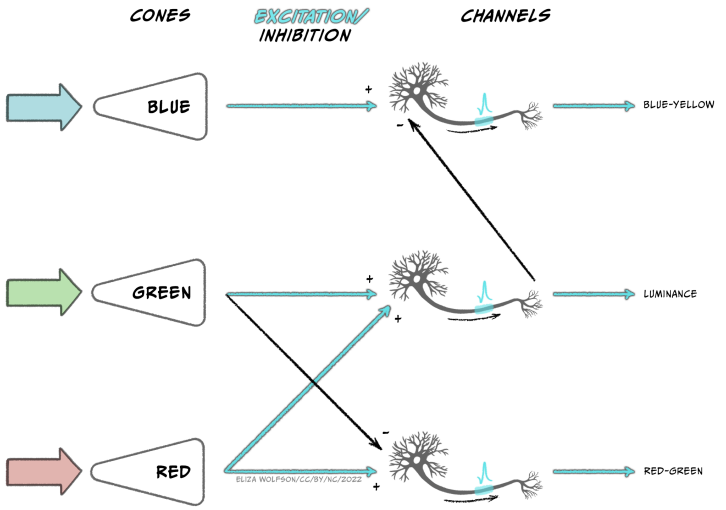

Fig 4.44. Opponent processing channels in the visual system

Cells that respond in line with this theory have been found in the LGN and the primary visual cortex. Further along in the ventral pathway, V4 has been found to contain neurons which respond to a range of colours i.e., not just red, green, blue and yellow (Zeki, 1980). This area receives input from V2 and sends information onwards to V8. The latter of which appears to combine colour information with memory information (Zeki & Marini, 1998).

Imagine looking out of your window on a bright morning to the leaves on the trees outside in the sunshine. Now consider looking out a few hours later when the weather has become dull and overcast. Do you perceive that the leaves have changed colour?

Hopefully you answered 'No' this question because you know that the leaves have not changed colour. But how do you know this?

The fact that we can perceive colour as unchanging despite overall changes in luminance is because of a phenomenon called **colour constancy**. The brain compensates for differences in luminance by taking into account the average colour across the visual scene.

We mentioned previously that some cells in V1 respond to specific orientations of stimuli. In addition to these cells, other cells in V1 have been found to respond to specific movement of stimuli indicating that motion detection begins early in cortical processing. However, it is V5, also known as MT (for medial temporal area), and the adjacent region V5a or medial superior temporal area, that are thought to be critical in

motion detection. V5 region receives input from V1 but also from the superior colliculus which is involved in visual reflexes that are important for motion. Information from V5 is then sent onwards to V5a which has been found to have neurons that respond to specific motion patterns including spiral motion (Vaina, 1998).

Before we move on to the final subsection on visual impairment, we will look briefly at depth perception. The image created on the retina is two dimensional and yet we can perceive a three-dimensional world. This is possible because we use specific depth cues. Spend a moment looking at the visual scene in Figure 4.45.

Fig 4.45. Chatsworth House in Derbyshire

The image shown in Figure 4.45 is complex, with multiple components including the house, fountain cascade, trees and the landscape beyond the house leading to the horizon. But how do we know how all the components fit together? For example, how do we know which trees are in front of the house and which are behind it or whether the trees in the distance are far away or just small? We can interpret the scene using depth cues including:

- Interposition: Objects which obscure other objects are closer to the viewer than the ones they obscure.
- Linear perspective: Parallel lines will converge as they move further away. This is illustrated with the sides of the fountain cascade in the image.
- Size constancy: Objects which appear smaller are likely be further away so trees in the distance will be smaller than those nearby because they are in the distance rather than because they differ in size.
- Height in the field: The horizon tends to appear towards the middle of the image with objects below the horizon nearer to the observer than the horizon and those close to the bottom of the image the nearest to the viewer.

We also obtain visual cues by comparing the images we have from the left and right eye – these are binocular cues. For example, when our left and right eye receive slightly different images, referred to as binocular disparity, this can alter our

depth perception. It is through compiling cues like this, along with discrete information about colour, orientation and motion, that we can create a perception of the world around us.

Blindness: causes, impact and treatment

Globally, the main causes of visual impairment are uncorrected refractive errors, as discussed in Box 8. However, these do not typically result in blindness. The leading cause of blindness is cataracts which accounts for 51% of blindness worldwide (Pascolini & Mariotti, 2012). Cataracts occur when the lens of the eye develops cloudy patches, losing the transparency which is critical for transmitting light. Individuals may experience cataracts in one or both eyes. As the lens becomes cloudy, light cannot reach the retina. The National Institute for Health Care and Excellence (NICE, 2022) have identified several risk factors for cataracts including:

- Ageing – most cataracts occur in people over 60 years of age
- Eye disease – in this case the cataracts can occur because of other conditions
- Trauma – the cataracts arise due to injury to the eye
- Systemic disease – the cataracts arise because of other conditions, for example, diabetes

Aside from a decline in visual abilities to the point of blindness, cataracts have been associated with wider impact on health. In age-related cataracts the presence of cataracts is related to cognitive decline and increased depression (Pellegrini et al., 2020).

At the time of writing the only proven effective treatment for cataracts is surgery to replace the lens of the eye with a synthetic lens. These lenses cannot adjust like the natural lens so glasses will typically need to be worn after the surgery. The surgery is short (around 30 mins) and carried out under a local anaesthetic with a 2-6 week recovery period. These operations are considered routine in countries like the UK, but in low- and middle-income countries, eye care is often inaccessible, and cataracts can cause blindness.

After cataracts the next leading cause of blindness is glaucoma, accounting for 8% of cases followed by age-related macular degeneration (AMD), accounting for 5% of cases (Pascolini & Mariotti, 2012). Glaucoma refers to a build-up of pressure within the eye that can lead to damage to the optic nerve. This build-up happens because the fluid in the eye cannot drain properly, and it typically happens over time. As with cataracts, the condition is more common in older people. Several treatment options exist for glaucoma including use of eye drops, laser treatment and surgery, all aiming to reduce the intraocular pressure, but damage may be irreversible. Perhaps unsurprisingly, this condition is also associated with poorer quality of life (Quaranta et al., 2016).

For both cataracts and glaucoma, the site of damage is not specifically the retina and the photoreceptors. However, AMD does result from damage to the retina. In this case, the macular region of the retina deteriorates causing blurred central vision, although peripheral vision is intact, meaning it only causes complete blindness in a small percentage of people. As indicated by the name, this is an age-related condition such that older people are more likely to develop it, but other risk factors include smoking and exposure to sunlight. Whilst the remaining vision might suggest less impact on individuals than other types of visual impairment, it is still associated with reduced quality of life, anxiety and depression (Fernández-Vigo et al., 2021).

There are two types of age-related macular degeneration: dry and wet. Dry AMD occurs because of a failure to remove cellular waste products from the retina. These products build up causing deterioration of blood vessels and cell death of the rods and cones. This type of AMD accounts for around 90% of the AMD cases and there is no treatment for this type of AMD. Wet AMD arises in around 10% of people with AMD as a progression from dry AMD. Here new blood vessels form in the eye, but they are weak and prone to leaking. This type of AMD can be treated with regular injections into the eye to reduce the growth of new blood vessels. An alternative to injections, or to be used alongside the injections, is Photodynamic Therapy (PDT) where a laser is directed to the back of the eye to destroy the abnormal blood vessels there.

Key Takeaways: Summarising Vision

- Our sense of vision uses light as a sensory stimulus. Visible light is part of the electromagnetic spectrum and can be conceptualised as both a wave and a particle
- Light emitted from objects or reflected off them enters the eye through the dioptric apparatus where it is bent to a focal point on the retina at the back of the eye. Most of the refractive power comes from the cornea, but the lens provides an adjustable amount of power
- Refractive errors such as myopia can arise when the dioptric apparatus is too weak or powerful, causing the focal point to be in front of, or behind, the retina. Although refractive errors are a leading cause of visual impairment worldwide, they do not typically result in blindness
- Visual transduction occurs in the photoreceptors at the back of the retina of

which there are two classes: rods and cones. Rods outnumber cones overall and are more sensitive, providing vision in scotopic conditions, but provide lower acuity and are found predominantly in the peripheral retinal areas. In contrast, cones are largely found in the fovea, and are specialised for high acuity, photopic vision. There are three types of cones, each with differing spectral sensitivity, giving rise to our colour perception

- The process of visual transduction begins with activation of photosensitive pigment in the photoreceptors. After this a series steps involving G-proteins results in the closure of ion channels and therefore a reduction of calcium entering the cell. This results in reduced glutamate release. Unlike the other senses, the presence of a stimulus results in hyperpolarisation of the receptor
- Retinal ganglion cells carry information away from the retina in the optic nerve to the lateral geniculate nucleus and onto the primary visual cortex. Information is arranged according to the eye and visual field and retinotopically mapped. After leaving the

primary visual cortex over 30 cortical regions will receive visual input, including those forming the dorsal and ventral stream. Subcortical pathways also exist, most notably the pathway from the retina to the superior colliculus

- Specific features of the visual scene are identified by specific neural processes. For example, colour is believed to arise through opponent processing creating red-green, blue-yellow and luminance channels in the ventral pathway. Motion sensitive cells have been found in the dorsal pathway
- Different components of a visual scene can be combined and use of specific cues e.g., linear perspective can be used to create a 3D perception from the 2D image on the retina
- Leading causes of blindness are age related and include cataracts, glaucoma and age-related macular degeneration. In all cases the condition can have a significant impact on quality of life and result in distress. Treatments exist for most of these conditions but access to those treatments varies widely across the world.

References

Bartfeld, E., & Grinvald, A. (1992). Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proceedings of the National Academy of Sciences USA*, 89, 11905-11909. Neurobiology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50666/pdf/pnas01098-0266.pdf

Fernández-Vigo, J. I., Burgos-Blasco, B., Calvo-González, C., Escobar-Moreno, M. J., Shi, H., Jiménez-Santos, M., Valverde-Megías, A., Reche-Frutos, J., López-Guajardo, L., & Donate-López, J. (2021). Assessment of vision-related quality of life and depression and anxiety rates in patients with neovascular age-related macular degeneration. *Archivos de la Sociedad Española de Oftalmología (English Edition), 96*(9), 470-475. https://dx.doi.org/10.1016/j.oftale.2020.11.008

Foroni, F., Pergola, G., & Rumiati, R. I. (2016). Food color is in the eye of the beholder: the role of human trichromatic vision in food evaluation. *Scientific Reports, 6*(1), 37034. https://doi.org/10.1038/srep37034

NICE. (2022). *Cataracts*. Retrieved from https://cks.nice.org.uk/topics/cataracts/background-information/causes-risk-factors/

Pascolini, D., & Mariotti, S. P. (2012). Global estimates of visual impairment: 2010. *British Journal of Ophthalmology, 96*(5), 614-618. https://doi.org/10.1136/bjophthalmol-2011-300539

Pellegrini, M., Bernabei, F., Schiavi, C., & Giannaccare, G. (2020). Impact of cataract surgery on depression and cognitive function: Systematic review and meta-analysis. *Clinical & Experimental Ophthalmology, 48*(5), 593-601. https://doi.org/10.1111/ceo.13754

Pöppel, E., Held, R., & Frost, D. (1973). Residual visual function after brain wounds involving the central visual pathways in man. *Nature, 243*(5405), 295-296.

Quaranta, L., Riva, I., Gerardi, C., Oddone, F., Floriani, I., & Konstas, A. G. (2016). Quality of life in glaucoma: A review of the literature. *Advances in Therapy, 33*(6), 959-981. https://doi.org/10.1007/s12325-016-0333-6

Vaina, L. M. (1998). Complex motion perception and its deficits. *Current Opinion in Neurobiology, 8*(4), 494-502. https://doi.org/10.1016/S0959-4388(98)80037-8

Weiskrantz, L. (1999). *Consciousness lost and found: A neuropsychological exploration.* OUP https://doi.org/10.1093/acprof:oso/9780198524588.001.0001

Weiskrantz, L., & Warrington, E. (1974). K., Sanders, M. D., & Marshall, J. Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain, 97*(1), 709-728. https://doi.org/10.1093/brain/97.1.709

World Health Organisation. (2018). *Blindness and visual*

impairment. Retrieved from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

World Health Organisation. (2021). *Global eye care targets endorsed by member states at the 74th World Health Assembly*. Retrieved from https://www.who.int/news/item/27-05-2021-global-eye-care-targets-endorsed-by-member-states-at-the-74th-world-health-assembly

Zeki, S. (1980). The representation of colours in the cerebral cortex. *Nature, 284*, 412-418. https://doi.org/10.1038/284412a0

Zeki, S., & Marini, L. (1998). Three cortical stages of colour processing in the human brain. *Brain, 121*(9), 1669-1685. https://doi.org/10.1093/brain/121.9.1669

About the author

Dr Eleanor Dommett
KING'S COLLEGE LONDON
https://twitter.com/EllieJane1980?ref_src=twsrc%5Egoogle%7Ctwcam
https://www.linkedin.com/in/eleanor-dommett-33193011a/?originalSubdomain=uk

Dr Ellie Dommett studied psychology at Sheffield University. She went on to complete an MSc Neuroscience at the Institute of Psychiatry before returning to Sheffield for her doctorate, investigating the superior colliculus, a midbrain multisensory

structure. After a post-doctoral research post at Oxford University she became a lecturer at the Open University before joining King's College London, where she is now a Reader in Neuroscience. She conducts research into Attention Deficit Hyperactivity Disorder, focusing on identifying novel management approaches.

10.

PERCEIVING SOUND: OUR SENSE OF HEARING

Dr Eleanor J. Dommett

> *I always say deafness is a silent disability: you can't see it, and it's not life-threatening, so it has to touch your life in some way in order for it to be on your radar.*
>
> Rachel Shenton, Actress and Activist

Rachel Shenton, quoted above, is an actress who starred in, created and co-produced *The Silent Child* (2017), an award-winning film based on her own experiences as the child of a parent who became deaf after chemotherapy. The quote illustrates the challenge of deafness, which in turn demonstrates our reliance on hearing. As you will see in this section, hearing is critical for safely navigating the world and communicating with others. Consequently, hearing loss can have a devastating impact on individuals. To understand the importance of hearing and how the brain processes sound, we begin with the sound stimulus itself.

Making waves: the sound signal

The stimulus that is detected by our auditory system is a sound wave – a longitudinal wave produced from fluctuations in air pressure by vibration of objects. The vibration creates regions where the air particles are closer together (compressions) and regions where they are further apart (rarefractions) as the wave moves away from the source (Figure 4.17).

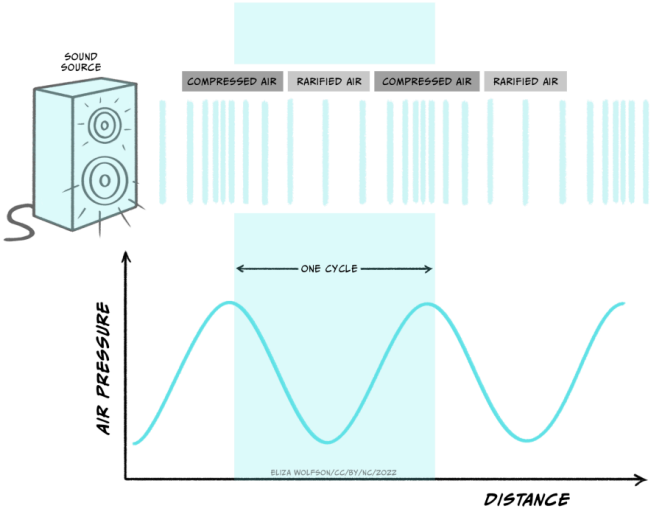

Fig 4.17. Sound waves created by vibration of an object

The nature of the sound signal is such that the source of the sound is not in direct physical contact with our bodies. This is different from the bodily senses described in the first two

sections because in the senses of touch and pain the stimulus contacts the body directly. Because of this difference touch and pain are referred to as proximal senses. By contrast, in hearing, the signal originates from a source not in direct contact with the body and is transmitted through the air. This makes hearing a distal, rather than proximal, sense.

The characteristics of the sound wave are important for our perception of sound. Three key characteristics are shown in Figure 4.18: frequency, amplitude and phase.

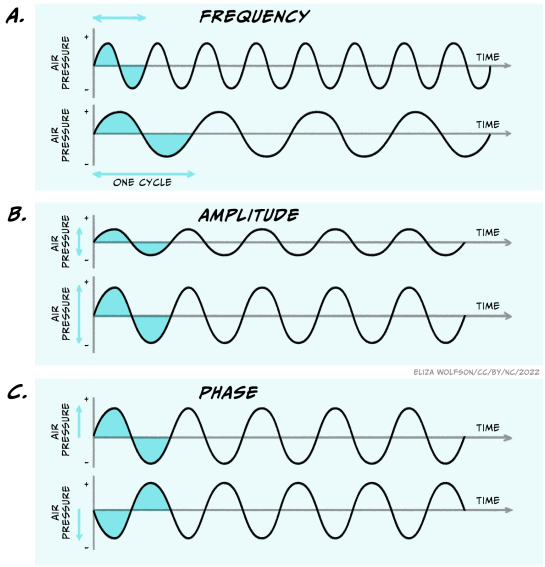

Fig 4.18. Key characteristics of sound are a) frequency, b) amplitude and c) phrase

The frequency is the time it takes for one full cycle of the

wave to repeat, and is measured in Hertz (Hz). One Hertz is simply one cycle per second. Humans can hear sounds with a frequency of 20 – 20000Hz (20 KHz). Examples of low frequency sounds, which are generally considered to be under 500 Hz, include the sounds of waves and elephants! In contrast, higher frequency sounds include the sound of whistling and nails on a chalk board. Amplitude is the amount of fluctuation in air pressure that is produced by the wave. The amplitude of a wave is measured in pascals (Pa), the unit of pressure. However, in most cases when considering the auditory system, this is converted into intensity and intensities are discussed in relative terms using the unit of the decibel (dB). Using this unit the range of intensities humans can typically hear are 0-140 dB. Above this level can be very harmful to our auditory system. Although you may see sound intensity expressed in dB, another expression is also commonly used. Where the intensity of sound is expressed with reference to a standard intensity (the lowest intensity a young person can hear a sound of 1000 Hz), it is written as dB SPL. The SPL stands for sound pressure level. Normal conversation is typically at a level of around 60 dB SPL.

Unlike frequency and amplitude, phase is a relative characteristic because it describes the relationship between different waves. They can be said to be in phase, meaning they have peaks at the same time or out of phase, meaning that they are at different stages in their cycle at anyone point in time.

The three characteristics above and the diagrams shown

indicate a certain simplicity about sound signals. However, the waves shown here are pure waves, the sort you might expect from a tuning fork that emits a sound at a single frequency. These are quite different to the sound waves produced by more natural sources, which will often contain multiple different frequencies all combined together giving a less smooth appearance (Figure 4.19).

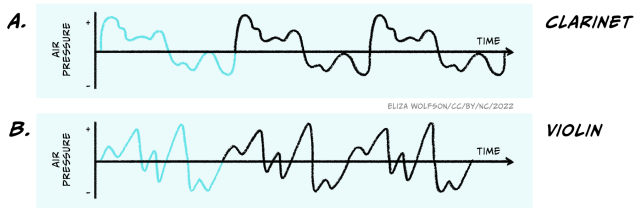

Fig 4.19. Examples of sound waves produced by the clarinet and clarinet

In addition, it is rare that only a single sound is present in our environment, and sound sources also move around! This can make sound detection and perception a very complex process and to understand how this happens we have to start with the ear.

Sound detection: the structure of the ear

The human ear is often the focus of ridicule but it is a highly specialised structure. The ear can be divided into three different parts which perform distinct functions:

- The outer ear which is responsible for gathering sound and funnelling it inwards, but also has some protective features
- The middle ear which helps prepare the signal for receipt in the inner ear and serves a protective function
- The inner ear which contains the sensory receptor cells for hearing, called hair cells. It is in the inner ear that transduction takes place.

Figure 4.20 shows the structure of the ear divided into these three sections.

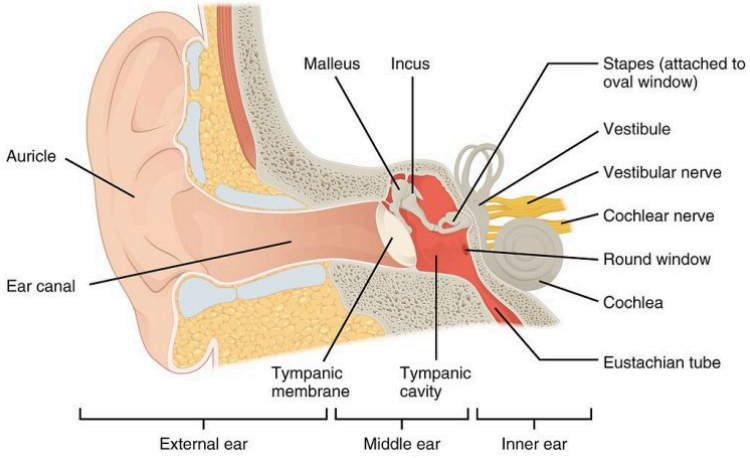

Fig 4.20. The human ear can be divided into the outer, middle and inner ear, each of which has a distinct function in our auditory system.

Although transduction happens in the inner ear, the outer and middle ear have key functions and so it is important that we briefly consider these.

The outer ear consists of the pinna (or auricle), which is the visible part that sticks out of the side of our heads. In most species the pinna can move but in humans they are static. The key function of the outer ear is in funnelling sound inwards, but the ridges of the pinna (the lumps and bumps you can feel in the ear) also play a role in helping us localise sound sources. Additional to this and often overlooked is the protective function of the outer ear. Ear wax found in the outer ear provides a water-resistant coating which is antibacterial and

antifungal, creating an acidic environment hostile to pathogens. There are also tiny hairs in the outer ear, preventing entry of small particles or insects.

The middle ear sits behind the tympanic membrane (or ear drum) which divides the outer and middle ear. The middle ear is an air-filled chamber containing three tiny bones, called the ossicles. These bones are connected in such away that they create a lever between the tympanic membrane and the cochlea of the inner ear, which is necessary because the the cochlear is fluid-filled.

Spend a moment thinking about the last time you went swimming or even put your head under the water in a bath. What happens to the sounds you could hear beforehand?

The sounds get much quieter, and will likely be muffled, if at all audible, when your ear is filled with water.

Hopefully you will have noted that when your ear contains water from a pool or the bath, sound becomes very hard to

hear. This is because the particles in the water are harder to displace than particles in air, which results in most of the sound being reflected back off the surface of the water. In fact only around 0.01% of sound is transmitted into water from the air, which explains why it is hard to hear underwater.

Because the inner ear is fluid-filled, this gives rise to a similar issue as hearing under water because the sound wave must move from the air-filled middle ear to the fluid-filler inner ear. To achieve this without loss of signal, the signal is amplified in the middle ear, by the lever actions of the ossicles, along with changes in the area of the bones contacting the tympanic membrane and cochlea, both of which result in a 20-fold increase in pressure changes as the sound wave enters the cochlea.

As with the outer ear, the middle ear also has a protective function in the form of the middle ear reflex. This reflex is triggered by sounds over 70 dB SPL and involves muscles in the middle ear locking the position of the ossicles.

What would happen if the ossicles could not move?

The signal could not be transmitted from the outer ear to the inner ear.

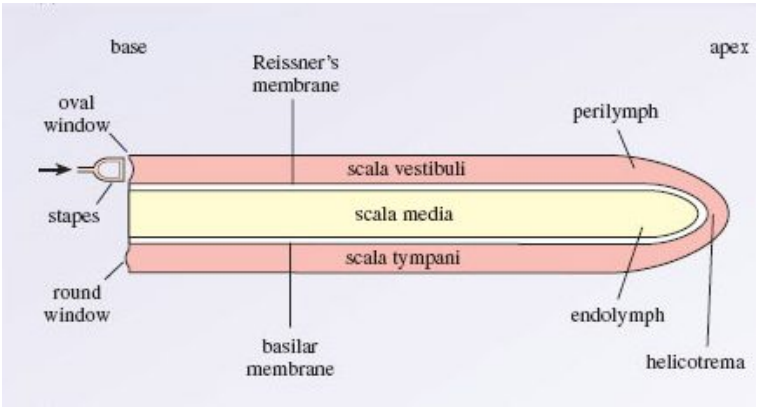

Fig 4.21a. Diagrammatic representation of the three scalae of the cochlea (uncoiled)

We now turn our attention to the inner ear and, specifically, the cochlea, which is the structure important for hearing (other parts of the inner ear form part of the vestibular system which is important for balance). The cochlea consists of a tiny tube, curled up like a snail. A small window into the cochlea, called the oval window (Figure 4.21a), is the point at which the sound wave enters the inner ear, via the actions of the ossicles.

The tube of the cochlea is separated into three different chambers by membranes. The key chamber to consider here is the **scala media** which sits between the **basilar membrane** and **Reissner's membrane**, and contains the **organ of corti** (Figures 4.21b, c).

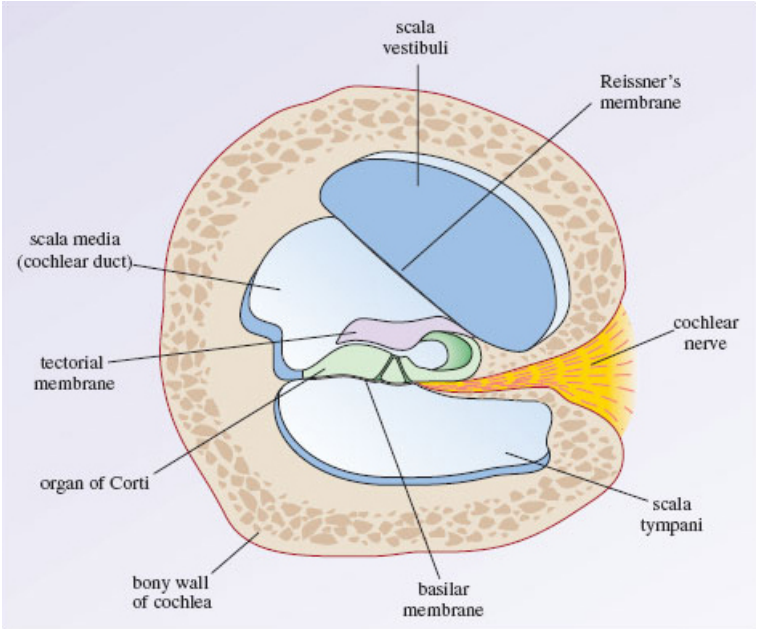

Fig 4.21b. The cross-sectional structure of the cochlea

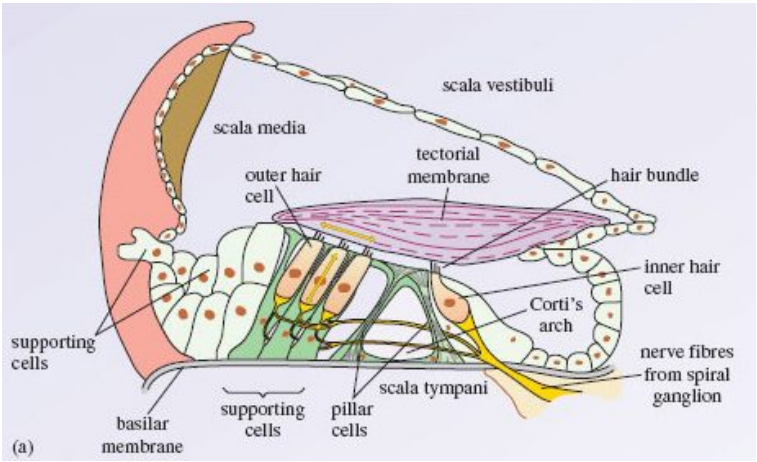

Fig 4.21c. The cross-sectional structure of the organ of Corti

The cells critical for transduction of sound are the inner hair cells which can be seen in Figure 4.21c.

These cells are referred to as **hair cells** because they contain hair-like **stereocillia** protruding from one end. The end from which the stereocillia protrude is referred to as the **apical** end. They project into a fluid called **endolymph**, whilst the other end of the cell, the **basal** end, sits in **perilymph**. The endolymph contains a very high concentration of potassium ions.

How does this differ from typical extracellular space?

Normally potassium is at a low concentration outside the cell and a higher concentration inside, so this is the opposite to what is normally found.

When a sound wave is transmitted to the cochlea, it causes the movement of fluid in the chambers which in turn moves the basilar membrane upon which the inner hair cells sit. This movement causes their stereocilia to bend. When they bend, mechano-sensitive ion channels in the tips open and potassium floods into the hair cell causing depolarisation (Figure 4.22). This is the auditory receptor potential.

Spend a moment looking at Figure 4.22. What typical neuronal features can you see? How are these cells different from neurons?

There are calcium gated channels and synaptic vesicles but there is no axon.

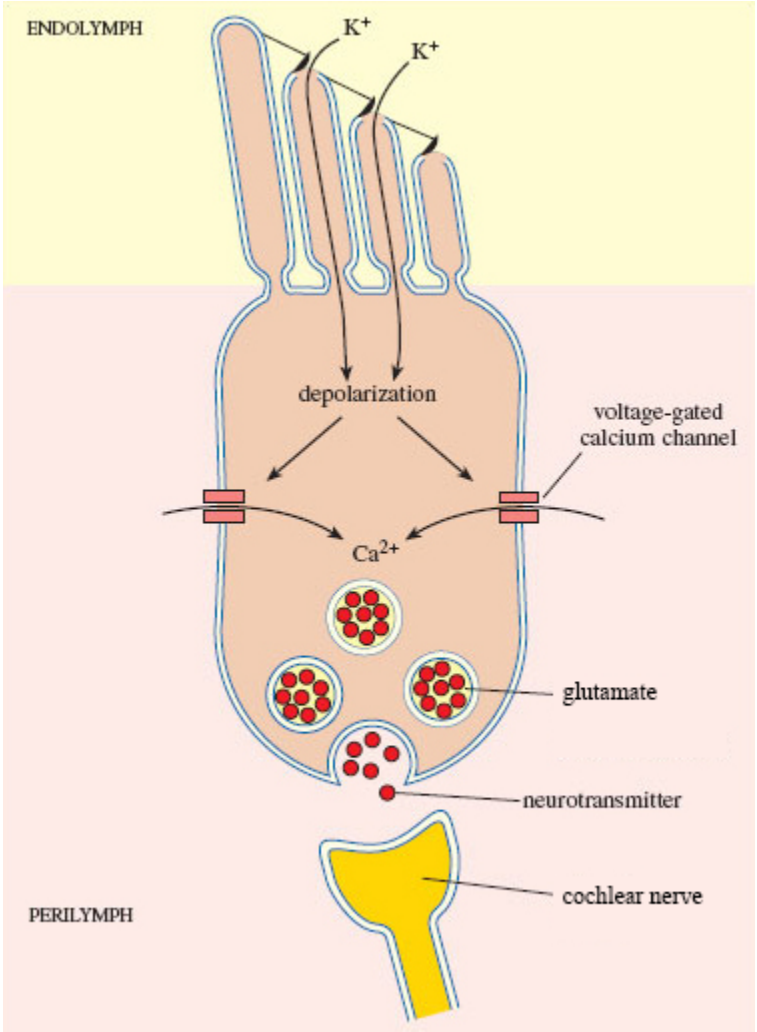

Fig 4.22. The inner hair cell responsible for transduction of sound waves.

You should have noted that the inner hair cells only have some of the typical structural components of neurons. This is because, unlike the sensory receptor cells for the somatosensory system, these are not modified neurons and they cannot produce action potentials. Instead, when sound is detected, the receptor potential results in the release of glutamate from the basal end of the hair cell where it synapses with neurons that form the cochlear nerve to the brain. If sufficient glutamate binds to the AMPA receptors on these neurons, an action potential will be produced and the sound signal will travel to the brain.

Auditory pathways: what goes up must come down

The cochlear nerve leaves the cochlea and enters the brain at the level of the brainstem, synapsing with neurons in the cochlear nuclear complex before travelling via the trapezoid body to the superior olive, also located in the brainstem. This is the first structure in the pathway to receive information from both ears. Prior to this in the cochlear nuclear complex, information is only received from the ipsilateral ear. After leaving the superior olive, the auditory pathway continues in the lateral leminiscus to the inferior colliculus in the midbrain before travelling to the medial geniculate nucleus of the thalamus. From the thalamus, as with the other senses you

have learnt about, the signal is sent onto the cortex. In this case, the primary auditory cortex in the temporal lobe. This complex ascending pathway is illustrated in Figure 4.23.

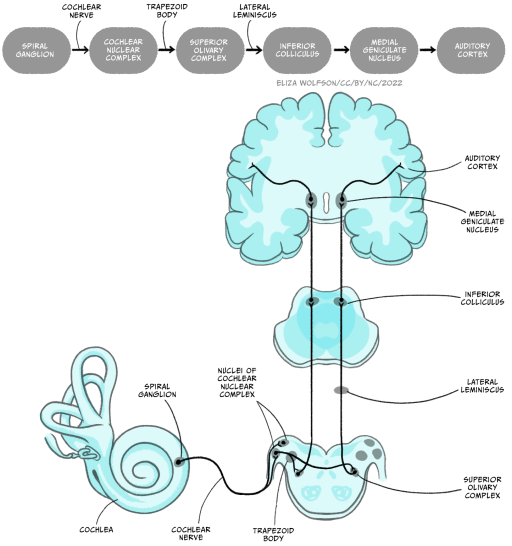

Fig 4.23. The ascending pathway from the cochlea to the cortex

You will learn about the types of processing that occurs at different stages of this pathway shortly but it is also important to recognise that the primary auditory cortex is not the end of the road for sound processing.

Where did touch and pain information go after the primary somatosensory cortex?

In both cases, information was sent onto other cortical regions, including secondary sensory areas and areas of the frontal cortex.

As with touch and pain information, auditory information from the primary sensory cortex, in this case the primary auditory cortex, is carried to other cortical areas for further processing. Information from the primary auditory cortex divides into two separate pathways or streams: the ventral 'what' pathway and the dorsal 'where' pathway.

The ventral pathway travels down and forward and includes the superior temporal region and the ventrolateral prefrontal cortex. It is considered critical for auditory object recognition, hence the 'what' name (Bizley & Cohen, 2013). There is not yet a clear consensus on the exact role in recognition that the different structures in the pathway play, but it is known that activity in this pathway may be modulated by emotion (Kryklywy, Macpherson, Greening, & Mitchell, 2013).

In contrast to the ventral pathway the dorsal pathway travels up and forward, going into the posterodorsal cortex in the parietal lobe and forwards into the dorsal lateral prefrontal cortex (Figure 24). This pathway is critical for identifying the location of sound, as suggested by the 'where' name. As with the ventral pathway, the exact role of individual structure is not clear but it too can be modulated by other functions. Researchers have found that whilst it is not impacted by emotion (Kryklywy et al., 2013) it is, perhaps unsurprisingly, modulated by spatial attention (Tata & Ward, 2005).

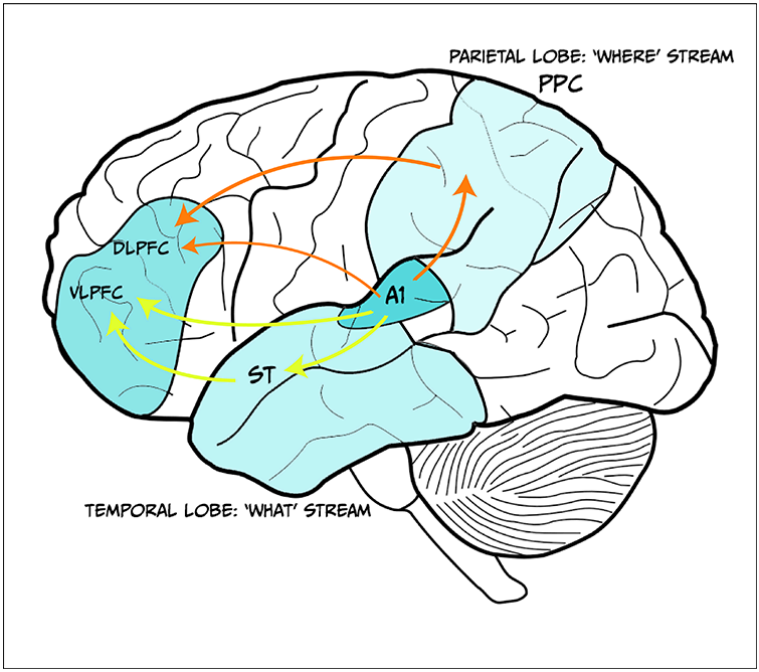

Fig 4.24. The dorsal and ventral streams of auditory information

Recall that when discussing pain pathways you learnt about a pathway which extends from higher regions of the brain to lower regions – a descending pathway. This type of pathway also exists in hearing. The auditory cortex sends projections down to the medial geniculate nucleus, inferior colliculus, superior olive and cochlear nuclear complex, meaning every structure in the ascending pathway receives descending input. Additionally, there are connections from the superior olive directly onto the inner and outer hair cells. These descending connections have been linked to several different functions including protection from loud noises, learning about relevant auditory stimuli, altering responses in accordance with the sleep/wake cycle and the effects of attention (Terreros & Delano, 2015).

Perceiving sound: from the wave to meaning

In order to create an accurate perception of sound information we need to extract key information from the sound signal. In the section on the sound signal we identified three key features of sound: frequency, intensity and phase. In this section we will consider these as you learn about how key features of sound are perceived, beginning with frequency.

The frequency of a sound is thought to be coded by the auditory system in two different ways, both of which begin

in the cochlea. The first method of coding is termed a place code because this coding method relies on stimuli of different frequencies being detected in different places within the cochlea. Therefore, if the brain can tell where in the cochlea the sound was detected, the frequency can be deduced. Figure 4.25 shows how different frequencies can be mapped within the cochlea according to this method. At the basal end of the cochlea sounds with a higher frequency are represented whilst at the apical end, low frequency sounds are detected. The difference in location arises because the different sound frequencies cause different displacement of the basilar membrane. Consequently, the peak of the displacement along the length of the membrane differs according to frequency, and only hair cells at this location will produce a receptor potential. Each hair cell is said to have a characteristic frequency to which it will respond.

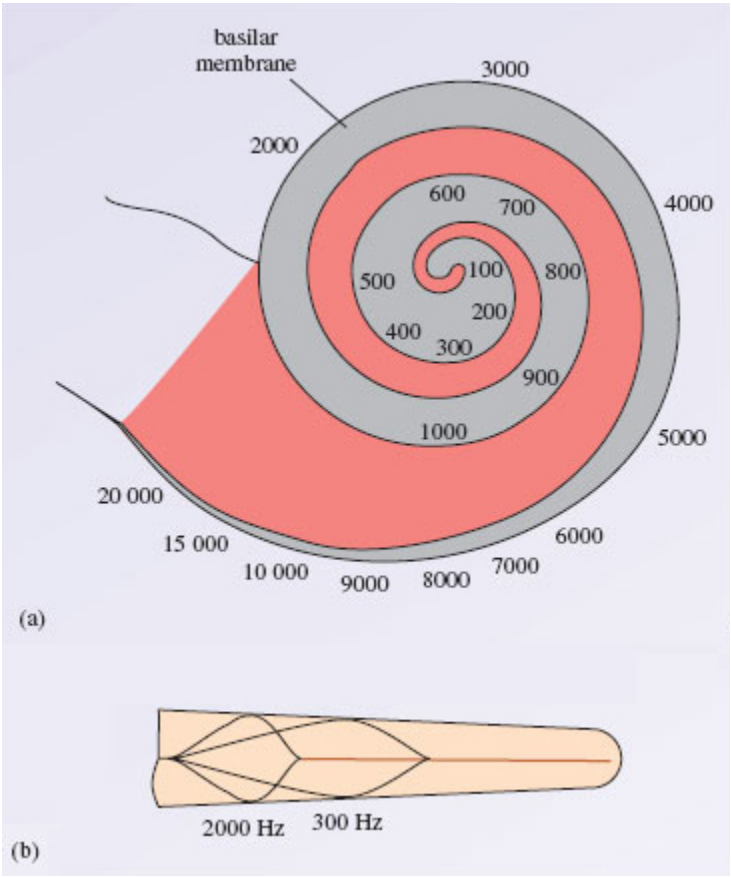

Fig 4.25. A schematic diagram of how different frequencies are located along the length of the cochlea (a). Distinct displacement patterns for signals of different frequencies (b)

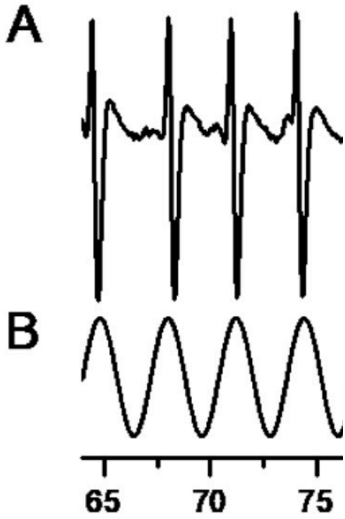

Fig 4.26. Temporal code assumes a direct relationship between stimulus intensity and firing of action potentials in the cochlear nerve. The neural response phase, locked to every cycle of the sound wave, is shown in the upper trace (A) and the frequency of the stimulus sound waveform (in ms) in the lower trace (B).

Although there is some support for a place code of frequency information, there is also evidence from studies in humans that we might be able to detect smaller changes in sound frequency that would be possible from place coding alone.

This led researchers to consider other possible explanations and to the proposal of a temporal code. This proposal is based on research which shows a relationship between the frequency of the incoming sound wave and the firing of action potentials in the cochlear nerve (Wever & Bray, 1930), which is illustrated in Figure 4.26. Thus when an action potential occurs, it provides information about the frequency of the sound.

Recall that we can hear sounds of up to 20,000 Hz or 20 KHz. How does this compare to the firing rate of neurons?

This is much higher than the firing rate of neurons. Typical neurons are thought to be able to fire at up to 1000 Hz.

Given the constraints of firing rate, it is not possible for temporal code to account for the range of frequencies that we can perceive. Wever and Bray (1930) proposed that groups of neurons could work together to account for higher frequencies, as illustrated in Figure 4.27.

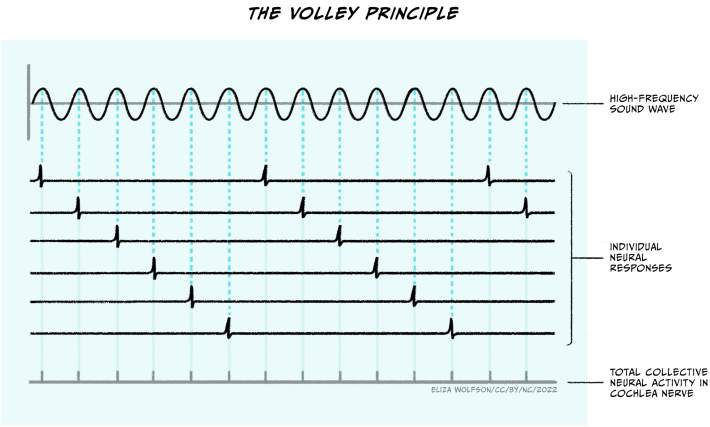

Fig 4.27. Wever and Bray suggested the volley principle where neurons work together to create an output which mimics the stimulus frequency.

The two coding mechanisms are not mutually exclusive and researchers now believe that temporal code may operate at very low frequencies (< 50 Hz) and place code may operate at higher frequencies (> 3000 Hz) with all intermediate frequencies being coded for my both mechanisms. Irrespective of which coding method is used for frequency in the cochlea, once encoded, this information is preserved throughout the auditory pathway.

Sound frequency can be considered an objective characteristic of the wave but the perceptual quality it most

closely relates to is pitch. This means that typically sounds of high frequency are perceived as having a high pitch.

The second key characteristic of sound to consider is intensity. As with frequency, intensity information is believed to be coded initially in the cochlea and then transmitted up the ascending pathway. Also in line with the coding of frequency, there are two suggested mechanisms for coding intensity. The first method suggests that intensity can be encoded according to firing rate in the auditory nerve. To understand this it is important to remember the relationship between stimulus and receptor potential which was first described in the section on touch. You should recall that the larger the stimulus, the bigger the receptor potential. In the case of sound, the more intense the stimulus, the larger the receptor potential will be, because the ion channels will be held open longer with a larger amplitude sound wave. This means that more potassium can flood into the hair cell causing greater depolarisation and subsequently greater release of glutamate. The more glutamate that is released, the greater the amount that is likely to bind to the post-synaptic neuron forming the auditory nerve. Given action potentials are all-or-none, the action potentials stay the same size but the frequency of them is increased.

The second method of encoding intensity is thought to be the number of neurons firing. Recall from Figure 4.25b that sound waves will result in a specific position of maximal displacement of the basilar membrane, and so typically only activate hair cells with the corresponding frequency which in

turn signal to specific neurons in the cochlear nerve. However, it is suggested that as a sound signal becomes more intense there will be sufficient displacement to activate hair cells either side of the characteristic frequency, albeit to a lesser extent, and therefore more neurons within the cochlear nerve may produce action potentials.

You may have noticed that the methods for coding frequency and intensity here overlap.

Considering the mechanisms described, how would you know whether an increased firing rate in the cochlear nerve is caused by a higher frequency or a greater intensity of a sound?

The short answer is that the signal will be ambiguous and you may not know straight away.

The overlapping coding mechanism can make it difficult to achieve accurate perception; indeed we know that perception of loudness, the perceptual experience that most closely correlates with sound intensity, is impacted significantly by the frequency of sound. It is likely that the combination of

multiple coding mechanisms supports our perception because of this. Furthermore, small head movements can be made which can impact on intensity of sound and therefore inform our perception of both frequency and intensity when the signal is ambiguous.

This leads us nicely onto the coding of sound location, which requires information from both ears to be considered together. For that reason sound localisation coding cannot take place in the cochlea and so happens in the ascending auditory pathway.

Which is the first structure in the pathway to receive auditory signals from both ears?

It is the superior olive in the brainstem.

The superior olive can be divided into the medial and lateral superior olive and each is thought to use a distinct mechanism for coding location of sound. Neurons within the medial superior olive receive excitatory inputs from both cochlear nuclear complexes (i.e., the one of the right and left), which allows them to act as coincidence detectors. To explain this a

little more it is helpful to think about possible positions of sound sources relative to your head. Figure 4.28 shows the two horizontal planes of sound: left to right and back to front.

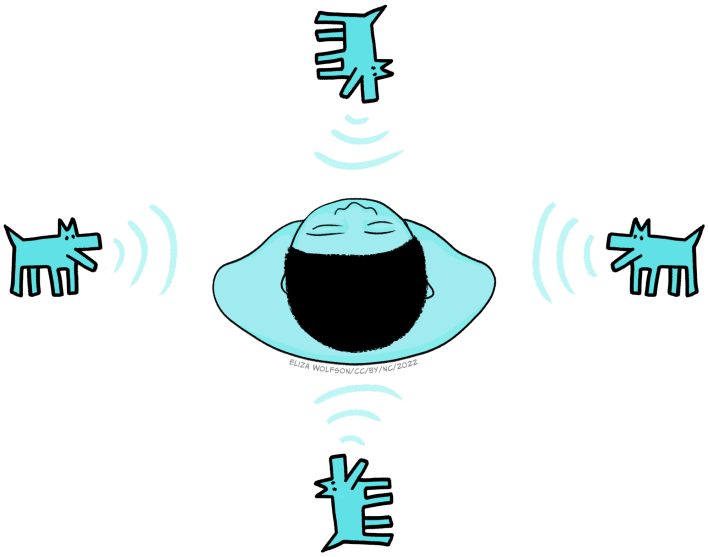

Fig 4.28. Examples of sound sources relative to the head

We will ignore stimuli falling exactly behind or exactly in front for a moment and focus on those to the left or right. Sound waves travel at a speed of 348 m/s (which you may also see written as ms^{-1}) and a sound travelling from one side of the body will reach the ear on that side ahead of the other side. The average distance between the ears is 20cm so this means that sound waves coming directly from, for example, the right

side, will hit the right ear 0.6 ms before they reach the left ear and vice versa if sound was coming from the left. Shorter delays between the sounds arriving at the left and right ear are experienced for sounds coming from less extreme right or left positions. This time delay means that neurons in the cochlear nerve closest to the sound source will fire first. This head start is maintained in the cochlear nuclear complex. Neurons in the medial superior olive are thought to be arranged such that they can detect specific time delays and thus code the origin of the sound. Figure 29 illustrates how this is possible. If a sound is coming from the left side, the signal from the left cochlear nuclear complex will reach the superior olive first and likely get all the way along to neuron C before the signal from the right cochlear nuclear complex combines with it, maximally exciting the neuron.

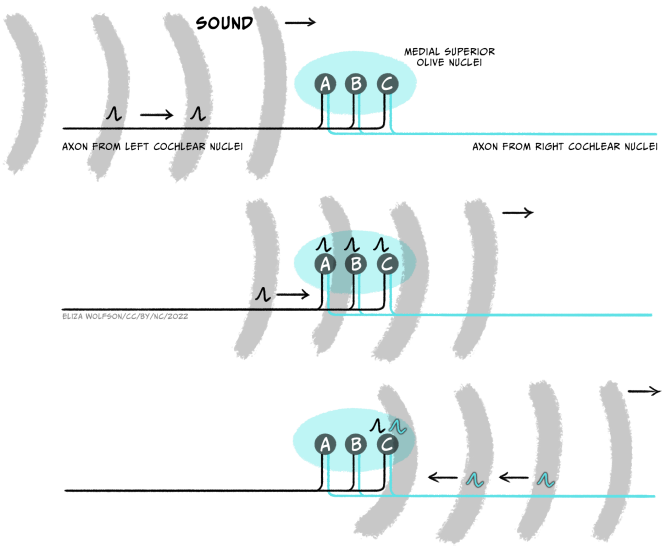

Fig 4.29. Delay lines and coincident detectors in the medial superior olive. Each of the three neurons shown (A, B, C) will fire most strongly when a signal from both ears reaches it at the same time. This will happen for neuron C when the sound wave is coming from the left, as the signal from the left cochlear nucleus has time to travel further (past neurons A and B) before the signal from the right cochlear nucleus arrives.

Using Figure 4.29, what would happen if the sound was from exactly in front or behind?

The input from the two cochlear nuclear complexes would likely combine on neuron B. Neuron B is therefore, in effect, a coincidence detector for no time delay between signals coming from the two ears. The brain can therefore deduce that the sound location is not to the left or the right – but it can't tell from these signals if the sound is in front of or behind the person.

This method, termed interaural (between the ears) time delay, is thought to be effective for lower frequencies, but for higher frequencies another method can be used by the lateral superior olive. Neurons in this area are thought to receive excitatory inputs from the ipsilateral cochlear nuclear complex and inhibitory inputs from the contralateral complex. These neurons detect the interaural intensity difference, that is the reduction in intensity caused by the sound travelling across the head. Importantly the drop of intensity as sound moves around the head is greater for higher frequency sounds. The detection of interaural time and intensity differences are therefore complementary, favouring low and high frequency sounds, respectively.

The two mechanisms outlined for perceiving location here are bottom-up methods. They rely completely on the data we receive, but there are additional cues to localisation. For

example, high frequency components of a sound diminish more than low frequency components when something is further away, so the relative amount of low and high frequencies can tell us something about the sound's location.

What would we need to know to make use of this cue?

We would need to know what properties (the intensity of different frequencies) to expect in the sound to work out if they are altered due to distance. Use of this cue therefore requires us to have some prior experience of the sound.

By combining all the information about frequency, intensity and localisation we are able to create a percept of the auditory world. However, before we move on it is important to note that whilst much of the auditory coding appears to take place in lower areas of the auditory system, this information is preserved and processed throughout the cortex. More importantly, it is also combined with top-down input and several structures will co-operate to create a perception of

complex stimuli such as music, including areas of the brain involved in memory and emotion (Warren, 2008).

Hearing loss: causes, impact and treatment

As indicated in the opening quote to this section, hearing loss can be a difficult and debilitating experience. There are several different types of hearing loss and each comes with a different prognosis. To begin with it is helpful to categorise types of hearing loss according to the location of the impairment:

- Conductive hearing loss occurs when the impairment is within the outer or middle ear, that is, the conduction of sound to the cochlea is interrupted.
- Cochlear hearing loss occurs when there is damage to the cochlea itself.
- Retrocochlear hearing loss occurs when the damage is to the cochlear nerve of areas of the brain which process sound. The latter two categories are often considered collectively under the classification of sensorineural hearing loss.

The effects of hearing loss are typically considered in terms of hearing threshold and hearing discrimination. Threshold refers to the quietest sound that someone is able to hear in a controlled environment, whilst discrimination refers to their

ability to concentrate on a sound in a noisy environment. This means that we can also categorise hearing loss by the extent of the impairment as indicated in Table 3.

Hearing Loss Classification	Hearing level (dB HL)	Impairment
Mild	20-39	Following speech is difficult esp.
Moderate	40-69	Difficulty following speech witho
Severe	70-89	Usually need to lip read or use sig hearing aid
Profound	90-120	Usually need to lip read or use sig ineffective

Table 3. Different classes of hearing loss

You should have spotted that the unit given in Table 3 is not the typical dB or dB SPL. This is a specific type of unit, dB HL or hearing level, used for hearing loss (see Box: Measuring hearing loss, below).

Measuring hearing loss

If someone is suspected of having hearing loss they

will typically undergo tests at a hearing clinic to establish the presence and extent of hearing loss. This can be done with an instrument called an audiometer, which produces sounds at different frequencies that are played to the person through through headphones (Figure 4.30).

Fig 4.30. A hearing test being conducted with an audiometer

The threshold set for the tests is that of a healthy young listener and this is considered to be 0 dB. If

someone has a hearing impairment they are unlikely to be able to hear the sound at this threshold and the intensity will have to be increased for them to hear it, which they can indicate by pressing a button. The amount by which is it increased is the dB HL level. For example, if someone must have the sound raised by 45 dB in order to detect the sound they will have moderate hearing loss because the value of 45 dB HL falls into that category (Table 3).

Conductive hearing loss typically impacts only on hearing threshold such that the threshold becomes higher, i.e., the quietest sound that someone can hear is louder than the sound someone without hearing loss can hear. Although conductive hearing loss can be caused by changes within any structure of the outer and middle ear, the most common occurrence is due to a build up of fluid in the middle ear, giving rise to a condition called otitis media with effusion, or glue ear. This condition is one of the most common illnesses found in children and the most common cause of hearing loss within this age group (Hall, Maw, Midgley, Golding, & Steer, 2014).

Why would fluid in the middle ear be problematic?

This is normally an air filled structure, and the presence of fluid would result in much of the sound being reflected back from the middle ear and so the signal will not reach the inner ear for transduction.

Glue ear typically arises in just one ear, but can occur in both. It generally only causes mild hearing loss. It is thought to be more common in children than adults because the fluid build-up arises due to the eustachian tube not draining properly. This tube connects the ear to the throat and normally drains the moisture from the air in the middle ear. In young children its function can be impacted adversely by the growth of adenoid tissue, which blocks the throat end of the tube meaning it cannot drain and fluid gradually builds up. However, several **risk factors** for glue ear have been identified.

These include iron deficiency (Akcan et al., 2019), allergies, specifically to dust mites (Norhafizah, Salina, & Goh, 2020), and exposure to second hand smoke as well as shorter duration of breast feeding (Kırıs et al., 2012; Owen et al., 1993). Social risk factors have also been identified including living in a larger family (Norhafizah et al., 2020), being part of a lower socioeconomic group (Kırıs et al., 2012) and longer hours spent in group childcare (Owen et al., 1993).

The risk factors of glue ear are possibly less important than

the potential consequences of the condition. It can result in pain and disturbed sleep which can in turn create behavioural problems, but the largest area of concern is on educational outcomes, due to delays in language development and social isolation as children struggle to interact with their peers. Studies have demonstrated poorer educational outcomes for children who experience chronic glue ear (Hall et al., 2014; Hill, Hall, Williams, & Emond, 2019) but it is likely that they can catch up over time, meaning any long lasting impact is minimal.

Despite the potential for disruption to educational outcomes, the first line of treatment for glue ear is simply to watch and wait and treat any concurrent infections. If the condition does not improve in a few months, grommets may be used. These are tiny plastic inserts put into the tympanic membrane to allow the fluid to drain. This minor surgery is not without risk because it can cause scarring of the membrane which may impact on its elasticity.

Whilst glue ear is the most common form of conductive hearing loss, the most common form of sensorineural hearing loss is Noise Induced Hearing Loss (NIHL). This type of hearing loss is caused by exposure to high intensity noises, from a range of contexts (e.g., industrial, military and recreational) and normally comes on over a period of time so gets greater with age, as hair cells are damaged or die. It is thought to affect around 5% of the population and typically results in bilateral hearing loss that affects both the hearing

threshold and discrimination. Severity can vary and its impact is frequency dependent with the biggest loss of sensitivity at higher frequencies (~4000 Hz) that coincide with many of the every day sounds we hear, including speech.

At present there is no treatment for NIHL and instead it is recommended that preventative measures should be taken, for example through the use of personal protective equipment (PPE).

What challenges can you see to this approach [using PPE]?

This assumes that PPE is readily available, which it may not be. For example, in the case of military noises, civilians in war zones are unlikely to be able to access PPE. It also assumes that PPE can be worn without impact. A musician is likely to need to hear the sounds being produced and so although use of some form of PPE may be possible, doing so may not be practical.

The impact of NIHL on an individual is substantial. For example research has demonstrated that the extent of hearing loss in adults is correlated with measures of social isolation,

distress and even suicide ideation (Akram, Nawaz, Rafi, & Akram, 2018). Other studies indicate NIHL can result in frustration, anxiety, stress, resentment, depression, and fatigue (Canton & Williams, 2012). There are also reported effects on employment with negative effects on employment opportunities and productivity (Canton & Williams, 2012; Neitzel, Swinburn, Hammer, & Eisenberg, 2017). Additionally, given NIHL will typically occur in older people, it may be harder to diagnose because they mistake it for a natural decline in hearing that occurs as people get older, meaning they may not recognise the need for preventive action if it is possible, or the need to seek help.

Looking across the senses

We have now reached the end of the section on hearing, but before we continue to look at the visual system it is helpful to spend a moment reflecting on the systems you have learnt about so far.

Exercises

1. Compare and contrast the mechanisms by which touch, pain and sound signals are transduced.

There are several similarities you could have mentioned here. For example, all these systems can include mechano-sensitive ion channels, that is, those that are opened by mechanical force. Additionally, they all involve the influx of a positively charged ion causes a depolarising receptor potential. There are also key differences as well. For example, whilst touch and hearing only use mechano-sensitive channels, pain can also use thermo-sensitive and chemo-sensitive channels. Furthermore, the ions that create the receptor potential differ. In somatosensation, the incoming ion is sodium, as is typical for depolarisation across the nervous system, whilst in hearing it is potassium due to the potassium-rich endolymph.

2. Considering the pathway to the brain, what do you notice is common to all the sensory systems discussed so far?

In all systems, the thalamus receives the signal on the way to the primary sensory cortex for that system. Additionally, there are typically, projections to a range of cortical areas after the primary sensory cortex.

3. Extracting key features of the sensory signal is important. What common features are detected across all three systems?

In all cases the intensity and location of the stimulus is encoded. Additionally, in touch and hearing, frequency of information is encoded.

Summarising hearing

Key Takeaways

- Our sense of hearing relies on the detection of a longitudinal wave created by vibration of objects in air. These waves typically vary in frequency, amplitude and phase
- The three-part structure of the ear allows us to funnel sounds inwards and amplify the signal before it reaches the fluid-filled cochlea of the inner ear where transduction takes place
- Transduction occurs in specialised hair cells which contain mechano-sensitive channels that open in response to vibration caused by sound waves. This results in an influx of potassium producing a receptor potential
- Unlike the somatosensory system, the hair cell is not a modified neuron and therefore cannot itself produce an action potential. Instead, an action potential is produced in

neurons of the cochlear nerve, when the hair cell releases glutamate which binds to AMPA receptors on these neurons. From here the signal can travel to the brain

- The ascending auditory pathway is complex, travelling through two brainstem nuclei (cochlear nuclear complex and superior olive) before ascending to the midbrain inferior colliculus, the medial geniculate nucleus of the thalamus and then the primary auditory cortex. From here it travels in dorsal and ventral pathways to the prefrontal cortex, to determine where and what the sound is, respectively
- There are also descending pathways from the primary auditory cortex which can influence all structures in the ascending pathway
- Key features are extracted from the sound wave beginning in the cochlea. There are two proposed coding mechanisms for frequency extraction: place coding and temporal coding. Place coding uses position-specific transduction in the cochlea whilst temporal coding locks transduction and subsequent cochlea nerve firing to the frequency of the

incoming sound wave. Once coded in the cochlea this information is retained throughout the auditory pathway

- Intensity coding is thought to occur either through the firing rate of the cochlea nerve or the number of neurons firing.
- Location coding requires input from both ears and therefore first occurs outside the cochlea at the level of the superior olive. Two mechanisms are proposed: interaural time delays and interaural intensity differences
- Hearing loss can be categorised according to where in the auditory system the impairment occurs. Conductive hearing loss arises when damage occurs to the outer or middle ear and sensorineural hearing loss arises when damage is in the cochlea or beyond
- Different types of hearing loss impact hearing threshold and hearing discrimination differently. The extent of hearing loss can vary as can the availability of treatments
- Hearing loss is associated with a range of risk factors and can have a significant impact on the individual including their social contact with others, occupational status and, in

children, academic development.

References

Akcan, F. A., Dündar, Y., Bayram Akcan, H., Cebeci, D., Sungur, M. A., & Ünlü, İ. (2019). The association between iron deficiency and otitis media with effusion. *The Journal of International Advanced Otology, 15*(1), 18-21. https://doi.org/10.5152/iao.2018.5394

Akram, B., Nawaz, J., Rafi, Z., & Akram, A. (2018). Social exclusion, mental health and suicidal ideation among adults with hearing loss: Protective and risk factors. *Journal of the Medical Association Pakistan, 68(3)*, 388-393. https://jpma.org.pk/article-details/8601?article_id=8601

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience, 14*(10), 693-707. https://dx.doi.org/10.1038/nrn3565

Canton, K., & Williams, W. (2012). The consequences of noise-induced hearing loss on dairy farm communities in New Zealand. *J Agromedicine, 17*(4), 354-363. https://dx.doi.org/10.1080/1059924x.2012.713840

Hall, A. J., Maw, R., Midgley, E., Golding, J., & Steer, C. (2014). Glue ear, hearing loss and IQ: an association moderated by the child's home environment. *PloS One,*

9(2), e87021. https://doi.org/10.1371/journal.pone.0087021

Hill, M., Hall, A., Williams, C., & Emond, A. M. (2019). Impact of co-occurring hearing and visual difficulties in childhood on educational outcomes: a longitudinal cohort study. *BMJ Paediatrics Open, 3*(1), e000389. http://dx.doi.org/10.1136/bmjpo-2018-000389

Kırıs, M., Muderris, T., Kara, T., Bercin, S., Cankaya, H., & Sevil, E. (2012). Prevalence and risk factors of otitis media with effusion in school children in Eastern Anatolia. *Int J Pediatr Otorhinolaryngol, 76*(7), 1030-1035. https://dx.doi.org/10.1016/j.ijporl.2012.03.027

Kryklywy, J. H., Macpherson, E. A., Greening, S. G., & Mitchell, D. G. (2013). Emotion modulates activity in the 'what' but not 'where' auditory processing pathway. *Neuroimage, 82*, 295-305. https://dx.doi.org/10.1016/j.neuroimage.2013.05.051

Neitzel, R. L., Swinburn, T. K., Hammer, M. S., & Eisenberg, D. (2017). Economic Impact of Hearing Loss and Reduction of Noise-Induced Hearing Loss in the United States. *Journal of Speech, Language, and Hearing Research, 60*(1), 182-189. https://dx.doi.org/10.1044/2016_jslhr-h-15-0365

Norhafizah, S., Salina, H., & Goh, B. S. (2020). Prevalence of allergic rhinitis in children with otitis media with effusion. *European Annals of Allergy and Clinical Immunology,*

52(3), 121-130. https://dx.doi.org/10.23822/EurAnnACI.1764-1489.119

Owen, M. J., Baldwin, C. D., Swank, P. R., Pannu, A. K., Johnson, D. L., & Howie, V. M. (1993). Relation of infant feeding practices, cigarette smoke exposure, and group child care to the onset and duration of otitis media with effusion in the first two years of life. *The Journal of Pediatrics, 123*(5), 702-711. https://dx.doi.org/10.1016/s0022-3476(05)80843-1

Tata, M. S., & Ward, L. M. (2005). Spatial attention modulates activity in a posterior "where" auditory pathway. *Neuropsychologia, 43*(4), 509-516. https://dx.doi.org/10.1016/j.neuropsychologia.2004.07.019

Terreros, G., & Delano, P. H. (2015). Corticofugal modulation of peripheral auditory responses. *Frontiers in Systems Neuroscience, 9*, 134. https://dx.doi.org/10.3389/fnsys.2015.00134

Warren, J. (2008). How does the brain process music? *Clinical Medicine, 8*(1), 32-36. https://doi.org/10.7861/clinmedicine.8-1-32

Wever, E. G., & Bray, C. W. (1930). The nature of acoustic response: The relation between sound frequency and frequency of impulses in the auditory nerve. *Journal of Experimental Psychology, 13*(5), 373. https://doi.org/10.1037/h0075820

About the author

Dr Eleanor Dommett
KING'S COLLEGE LONDON
https://twitter.com/EllieJane1980?ref_src=twsrc%5Egoogle%7Ctwcam
https://www.linkedin.com/in/eleanor-dommett-33193011a/?originalSubdomain=uk

Dr Ellie Dommett studied psychology at Sheffield University. She went on to complete an MSc Neuroscience at the Institute of Psychiatry before returning to Sheffield for her doctorate, investigating the superior colliculus, a midbrain multisensory structure. After a post-doctoral research post at Oxford University she became a lecturer at the Open University before joining King's College London, where she is now a Reader in Neuroscience. She conducts research into Attention Deficit Hyperactivity Disorder, focusing on identifying novel management approaches.

11.

CHEMICAL SENSES: TASTE AND SMELL

Dr Paloma Manguele and Dr Emiliano Merlo

Learning objectives

By the end of this chapter, you will be able to:

- identify the stimuli and sensory structures involved in taste and smell sensations
- understand the transduction mechanisms in place to transform chemical information into action potentials within each sense
- describe the neural pathways supporting gustatory and olfactory perception.

Sensing chemical compounds in the environment is the most archaic sensory mechanism in living organisms. Very early on in the history of life on Earth, unicellular organisms developed chemical detection to distinguish food from toxins, to find mates and avoid danger. All the way to the present day, motivated and emotional behaviours in human and non-human animals are greatly influenced by detection of environmental chemical signals. In this chapter, we will revise the current knowledge of chemical senses in humans, paying special attention to the signals they can detect, how the transduction from chemical to neural code takes place, and what brain regions are involved in each case.

Humans can detect chemical compounds in the environment via the olfactory (smells) and gustatory (tastes) systems. Flavours are also a product of chemical sensation, but they result from the combined perception of smells and tastes. Through these specialised senses, chemical information in the surrounding environment is captured by chemosensory receptors, located mainly in the mouth and nasal cavity. Information regarding quality and quantity of chemicals is converted into the language of the nervous system, action potentials, through **sensory transduction**, and then transmitted to the central nervous system. In the brain, this information is integrated to produce olfactory and gustatory perception that will ultimately influence decision-making and behavioural selection and action.

Even though somewhat less studied than senses such as vision and audition, the chemical senses can be organised in two systems: gustation and olfaction. The **gustatory** sense, or sense of taste, picks up on soluble chemical compounds, present in the mouth. The **olfactory** sense, or sense of smell, reacts to airborne molecules that reach the nasal cavity.

These sensory systems provide animals with key environmental information for producing adaptive behaviours. Smells serve as long- and short-range signals, whereas tastes only act in the short-range, after we ingest food or drinks. This information may be crucial for finding, selecting and consuming food, finding a potential mate or regulating social interactions with others. Even though the chemical senses are composed by standalone systems, their coordinated action can produce even more complex sensory capacities such as detecting flavours, which involves the activation of common sensory neurons within the piriform cortex – the part of the brain that first processes olfactory information (Fu, Sugai, Yoshimura, & Onoda, 2004).

Sense of taste

Anatomical overview

Our sense of taste starts with tastant molecules reaching our mouth through ingestion.

Did you know?

The lumps that you see on your tongue are mistakenly called **taste buds**, but they are actually epithelia tissues called **papillae.**

Tastants are water-soluble or lipid-soluble chemical substances, present in food or drinks, that create the sensation of taste when detected by **taste receptor cells (TRC)** within the mouth. We are quite familiar with the many little raised bumps on our tongue epithelium that can be seen by looking at the tongue with a naked eye. These lumps are called **papillae**, and it is within the walls and fissures of the papillae that we find the taste buds that contain the taste receptor cells. There are thousands of taste buds in each papillae, which are divided into three categories, depending on their location: the foliate papillae located on the sides of the posterior section of the tongue; the circumvallate papillae located at the back of the tongue; and the fungiform papillae located in the anterior part of the tongue (Figure 4.46).

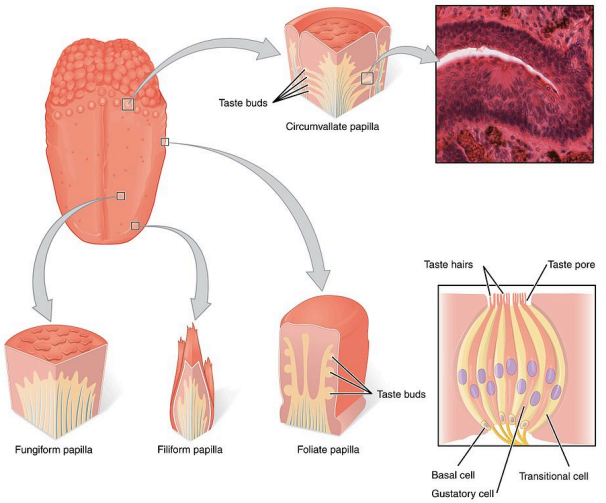

Fig 4.46. The tongue

Each taste bud contains groups of between 50 and 150 taste receptor cells, and presents an upper aperture called the **taste pore**. TRCs project fine hair-like extensions, or microvilli, out of taste pores into the buccal cavity, where they encounter the tastants.

In humans, there are three main types of taste receptor cells, according to their function. Type I cells have primarily housekeeping functions. Type II cells are sensitive to sweet, bitter, and umami tastes. Type III cells appear to mediate sour taste perception. Detection of tastant compounds by receptor cells leads to neurotransmitter release (usually ATP) and generation of action potentials in neurons at the base of these receptor cells. The axons of these neurons form the afferent nerves that transmit the information to the brain via three different cranial nerves: VII, IX and X. Different papillae areas

are innervated by different branches of the cranial nerves VII, IX, and X. The anterior two-thirds of the tongue, with the fungiform papillae, is supplied by branches of cranial nerve VII. The posterior third of the tongue is innervated by branches of nerve IX, the glossopharyngeal nerve. The posterior regions of the oesophagus and the soft palate are innervated by branches of cranial nerve X.

Nerves VII, IX and X project into the brainstem where they synapse with the rostral part of the **nucleus of the solitary tract** (NTS) which relays information to the ventral posterior medial nucleus of the thalamus. The thalamus projects to the anterior insular cortex and to a region called the **primary gustatory cortex** or insular taste cortex. Neural signals from the insular taste cortex travel to the secondary gustatory cortex, within the medial and lateral orbitofrontal cortex, and project also to structures like the amygdala, hippocampus, striatum and hypothalamus, where this sensory information can affect different stages of decision-making and behavioural output (Figure 4.48).

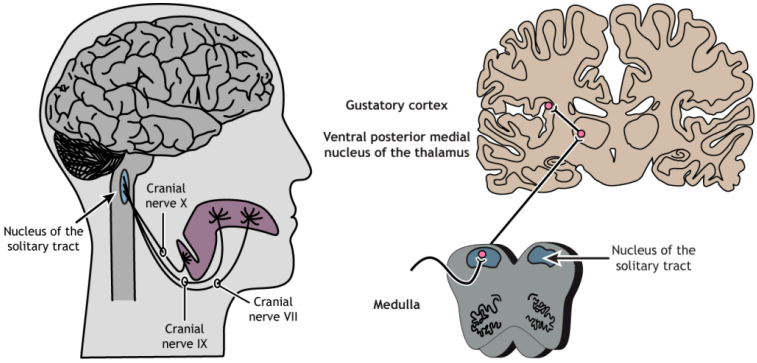

Fig 4.47. Taste information from the tongue travels through cranial nerves VII, IX, and X to the nucleus of the solitary tract in the medulla. Neurons in the brainstem project to the ventral posterior medial nucleus of the thalamus and then on to the gustatory cortex.

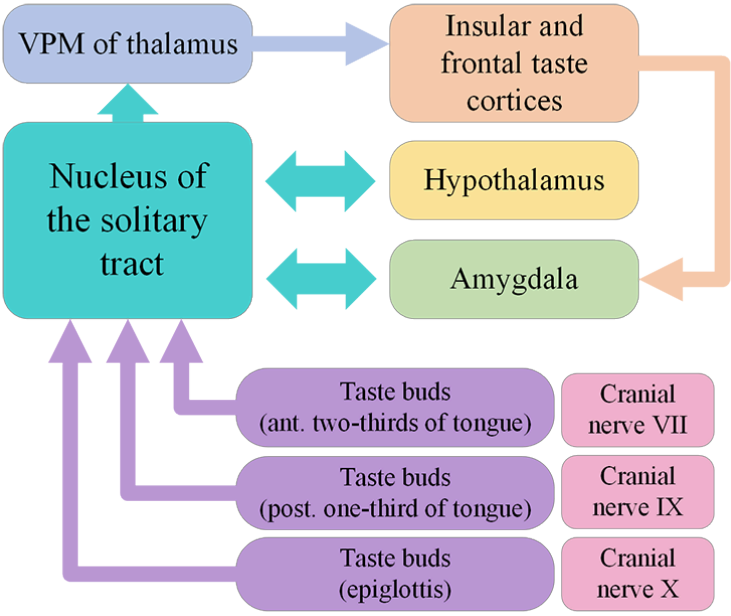

Fig. 4.48. Neuronal circuitry involved in taste perception and processing

Did you know?

Taste buds have a life span of about two weeks, allowing them to grow back even when they are

destroyed, for example when we burn our tongues. This makes them akin to skin cells, but they also share characteristics akin to neurons. For example, they have excitable membranes and release neurotransmitters.

Sensory transduction

How is the chemical information contained in the quality and quantity of specific tastants transformed into neural signals that the brain can interpret?

Tastants enter the papillae through the taste pore and induce different mechanisms in taste receptor cells. Each receptor cell has distinct mechanisms for transducing the chemical information into neural activity. Tastants are divided into salty, sour, sweet, bitter, and, umami – derived from the Japanese word meaning 'deliciousness'. The umami taste is produced by monosodium glutamate, and probably other related amino acids.

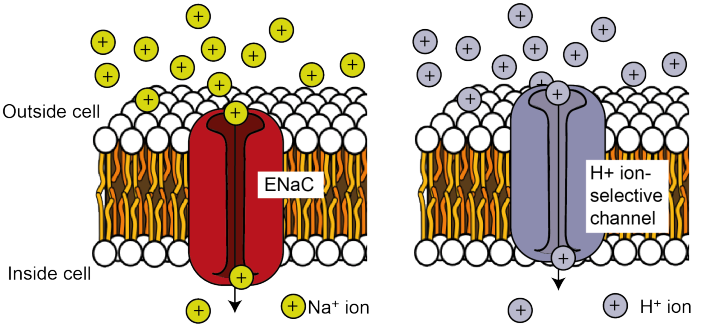

Fig 4.49a. Taste receptor cells

As we have heard, components of salty chemicals are key for survival in several animals. Na^+ ions contained in the saltiest of all salts, sodium chloride (NaCl), is key for maintaining muscle and neuronal functioning. A sub-group of Type II taste receptor cells are specialised for salt detection. These cells express receptors that detect and react to the presence of salty substances containing Na^+. Similar Type II receptors express channels that allow other free cations such as H^+ released by acid compounds into the cell (Figure 4.49a). Receptor cells expressing ion channels for Na^+ or H^+ allow these cations into the intracellular space, depolarisating the membrane, leading to release of neurotransmitter, typically ATP, and action potential firing in the neurons that make up the cranial nerves. Recent research has identified the ion channel responsible for NaCl detection in mice. Deletion of the gene that produces

the epithelial sodium channel (ENaC) in mice specifically affected a sub-group of Type II taste receptor cells. Mice lacking ENaC showed complete loss of salt attraction and sodium taste responses compared to control animals (Chandrashekar et al., 2010). This was the first evidence that salt is detected by a specific protein expressed in a distinctive type of TRCs.Other categories of tastant molecules, specifically those perceived as sweet, bitter, and umami, activate G-protein-coupled receptors (GPCRs, Figure 4.49b).

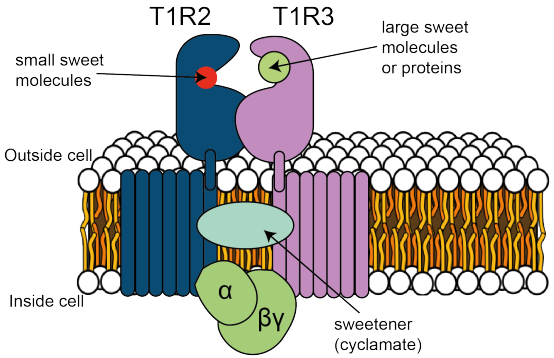

Fig 4.49b. G-protein -coupled receptors

As we have heard in earlier chapters, GPCRs are transmembrane receptors associated on their cytoplasmic side with G-proteins. They use a 'key to lock' mechanism for the transduction of the chemical information into neural activity. When a particular tastant molecule is recognised by a GPCR,

the associated G-protein is activated, dissociating into α and βγ subunits. These can activate further intracellular signalling cascades, leading to depolarisation and/or an increase in intracellular calcium concentration that ultimately results in the release of neurotransmitters, usually ATP.

In mammals, sweet and umami receptors are heteromeric GPCR named T1R2+3 and T1R1+3, respectively. These receptors are a combination of proteins from families T1R1, T1R2 or T1R3, and can detect sweet and umami taste compounds. Like the ENaC knockout mice, animals without T1R1 fail to detect umami compounds, whereas animals lacking T1R2 fail to detect sweet tastes (Zhao et al., 2003).

Exercise

Domestic cats, lions or tigers do not have the genes that codify for T1R2 receptors. This means they cannot taste sweet tastes and are unable to experience sweetness. How do you think this fact influences their strictly carnivore diet?

Topography/distribution of taste receptors

Historically, scientists had a rigid view of the topography or distribution of taste receptors, but this concept is slowly being abandoned. Nowadays, we recognise that taste zones across the surface of the tongue are not absolute, and that all zones can detect all tastes, albeit with different detection capacities. Taste sensitivity thresholds, rather than receptor distribution, vary across the surface of the tongue, with all areas showing higher or lower sensitivity to all tastants. For instance, receptors with higher sensitivity for bitter tastants tend to be distributed posteriorly in the tongue. Salty and sweet tastes are more easily detected in the tip of the tongue and are conveyed primarily by cranial nerve VII. Bitter sensations are mainly relayed by cranial nerve IX, which provides innervation to the posterior third of the tongue.

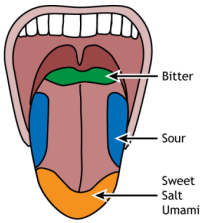

Fig 4.49. Although all tastes can be perceived across the entire tongue, sensitivity levels vary for each taste. The front of the tongue has the lowest threshold for sweet, salt, and umami tastes; the side of the tongue has the lowest threshold for sour tastes, and the back of the tongue has the

lowest threshold for bitter tastes.

Coding of information in the gustatory system

There is generally a proportional relationship between the concentration of the tastant and the firing rate of first order axons that enter the brain stem, so coding of taste intensity is based, at least in part, on frequency of action potentials.

Coding of gustatory information is also based on the topographical distribution of the taste receptor cells sensitivity. This distribution provides the foundation for **labelled-line coding** (Squire et al., 2012), meaning that information about the nature of the taste is provided by which cell has been activated. In other words, an axon that receives information from a sweet receptor is labeled as codifying sweetness. Hence, whenever this axon fires an action potential and conveys that signal into the brainstem, the received input is interpreted as sweetness. This is similar to the principles of encoding we encounter in the somatic sensory system, where the identity of the activated neuron, rather than the firing rate, indicates the quality of the signal carried by it (for example activation of a neuron innervating the finger is perceived as coming from that area, and the type of neuron activated influences what sensation is perceived).

In the case of gustation, we might recognise activity of axons as signals of the presence of sour, bitter, salty, sweet, and umami tastants. Other evidence, however, suggests that the

pattern of activity across neurons that preferentially respond to different taste characteristics is used to code for specific tastes (pattern or ensemble coding).

Within the central nervous system, tastants' identity is preserved in the relays from the nucleus of the solitary tract, into the ventral posterior medial complex of the thalamus, the gustatory cortex or insula (Doty, 2015). For some time, it was assumed that the insula would represent taste categories in a 'gustotopic map', but the empirical evidence has been elusive. Recent studies, using genetic tracing of taste receptor cells into the gustatory cortex, suggest that there are distinctive spatial patterns within the cortex, but no region is assigned to a single tastant (Accolla et al., 2007). Finally, the information from tastants reaches the orbitofrontal cortex, where it is integrated with sensory information from different modalities, suggesting that this area integrates tastes with other information to create more complex perceptual experiences.

Did you know?

There is some evidence of labelled-line coding in all of the sensory systems.

The concept refers to the idea that the line, or the pathway, from peripheral receptor into the brain, is labelled based on the presence of particular receptors that accomplish the sensory transduction process.

Sense of smell

Anatomical overview

In humans, olfaction, or the sense of smell, detects airborne molecules or odorants that enter the nasal cavity. Odorants interact with olfactory sensory neurons (OSNs) located in the olfactory epithelium that covers the dorsal and medial aspect of the nasal passageway (Figure 4.50).

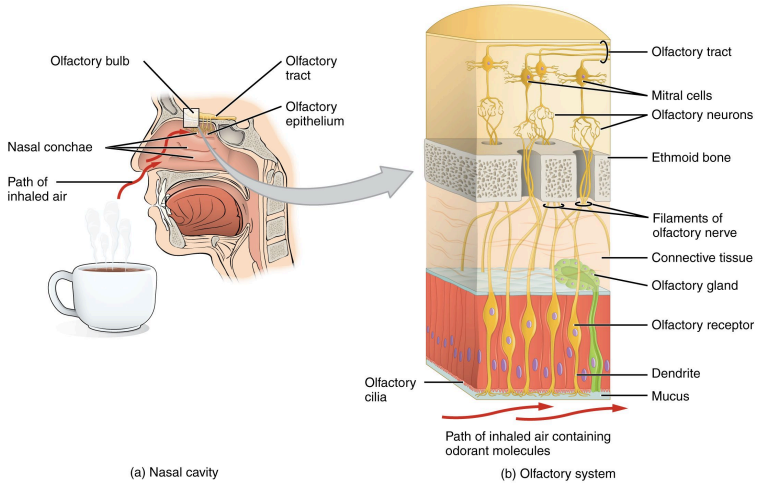

Fig 4.50. Details of the olfaction system. (a) The olfactory system begins in the peripheral structures of the nasal cavity. (b) The olfactory receptor neurons are within the olfactory epithelium.

OSNs are in charge of transducing the chemical information of odorants, encoding information about quality and quantity of smells, into action potentials that can be interpreted by the brain. Olfactory receptor cells extend their axons through the ethmoid bone, also called the cribriform plate. These axons make synaptic contact with the mitral cells within structures known as glomeruli within the olfactory bulb. Axons from the mitral cells bundle together and join the first cranial nerve, conveying olfactory information to various brain regions.

Information about the presence and quantity of smells

leaves the olfactory bulb via the **lateral olfactory tract**. On the olfactory pathway, the lateral olfactory tract connects back to the inferior and posterior parts of the frontal lobe, near the junction of the frontal lobe and the temporal lobe, which constitutes the beginnings of the **olfactory cortex** (Figure 4.51).

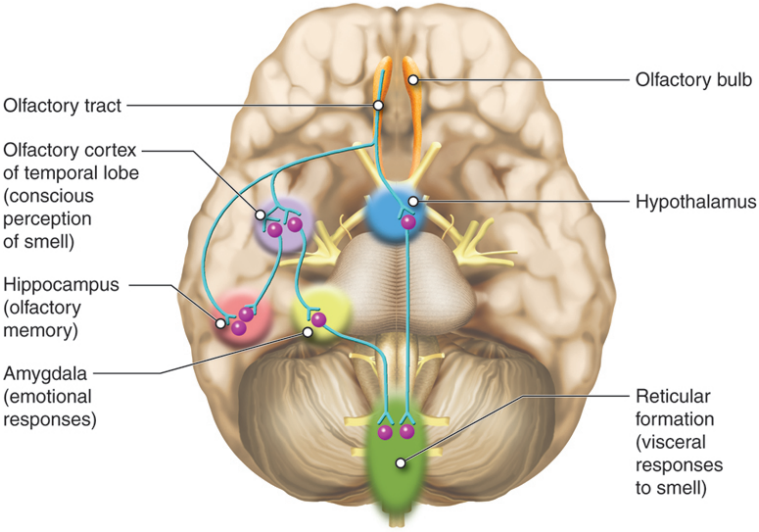

Fig 4.51

Unlike other primary sensory cortices, primary olfactory cortex comprises a number of different structures. These include subcortical structures such as the olfactory tubercle, in the ventral part of the striatum, and part of the amygdala as well as cortical regions in the medial part of the temporal lobe (entorhinal cortex) and its junction with the frontal lobe

(piriform cortex). The divisions of the olfactory cortex are interconnected, and even though there is most emphasis on the piriform cortex, the entire extended network of these regions constitute the olfactory cortex. Furthermore, these divisions of the olfactory cortex also project to other brain areas, including the thalamus, hypothalamus, hippocampus and, especially importantly, the orbital and frontal parts of the prefrontal cortex.

Unlike other sensory systems, in olfaction, there is not a thalamic relay between the peripheral sensory structures, i.e. the olfactory bulb, and the cortex (Breslin, 2019). In the olfactory system the connection with the thalamus is downstream from the cerebral cortex.

Sensory transduction and odour representation

Several types of cells are present at the olfactory epithelium. Supporting cells provide metabolic and physical support for the epithelium, but smell detection and transduction relies on mature cells called olfactory sensory neurons (OSNs). The nasal cavity is a challenging environment for living cells due to significant changes in environmental conditions such as humidity and temperature, which result in a short lifespan of OSNs. Constant mitotic divisions and maturation of basal cells replenishes the pool of OSNs, maintaining their number. In addition to the sensory and supporting cells, the epithelium

is composed of glandular cells that produce and secrete the thick mucus that covers and protects its more exposed cellular structures (Figure 4.50).

Odorant molecules that access the nasal cavity and diffuse through the mucus interact with olfactory cilia, hairlike extensions projecting from the end of the OSN dendrite. Embedded in the membrane of the olfactory cilia are the receptor proteins that bind with the odorants. Humans have around one thousand different odour receptor (OR) genes but can perceive more than a trillion different odours (Bushdid et al., 2014). In a characteristic arrangement of 'one-to-one-to-one', each OSN express only one type of OR gene, and all OSN expressing the same OR protein project their axons to the same glomeruli within the olfactory bulb (Figure 4.50). Hence, glomeruli activation recapitulates OR activation, reproducing a combinatorial code of glomeruli activity unique to each odour.

The current understanding of how odours are recognise at the neural level is explained by the shape-pattern theory, which proposes that each scent activates unique arrays of olfactory receptors in the epithelium. The molecular attributes of odours will determine how many OR can bind to them. Hence, one odour will activate a series of OR with more or less intensity, and this pattern of OR activation is what the brain recognises as a label for that particular odour molecule. Different odours will trigger different OR activation patterns, but familiar odours (i.e., sharing some molecular properties

like compounds belonging to the alcohol molecules family) will trigger more similar patterns since they may be recognised by overlapping but slightly differing OR combinations. Note that scents are usually a combination of more than one odour molecule, and scent perception is associated with a yet more complex pattern of OR activation and glomeruli representation. A graphical representation of this mechanism is presented in Figure 4.52.

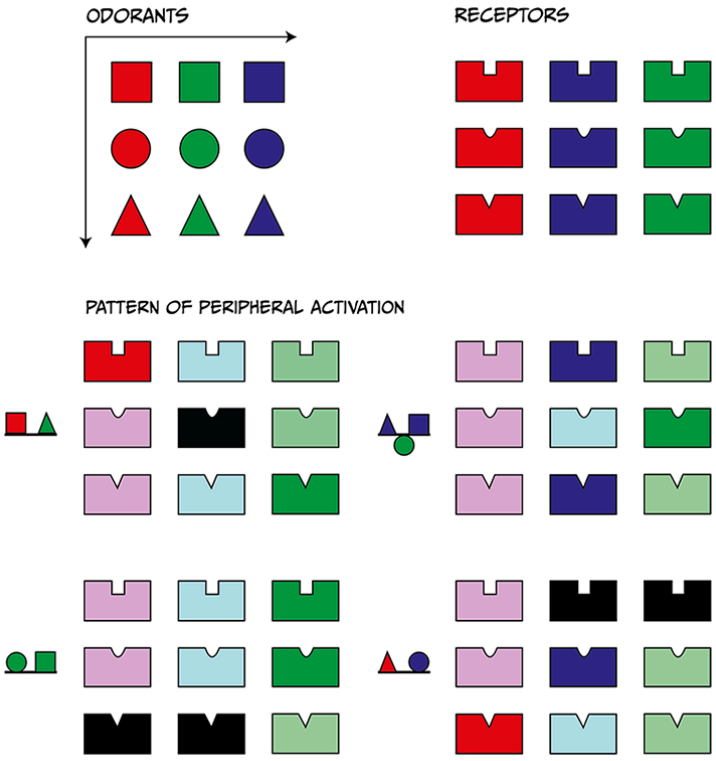

Fig 4.52. A graphical representation of scent perception

Odours are represented as geometrical shapes, and OR as a shape-fit structure. Odours will fit more or less well within specific shape-fit structures, with better fitness being associated with higher OSN activation. Specific combinations of odours (scents) will produce distinctive OR activation patterns, which will be univocally identified by the olfactory sensory brain areas.

When odorants bind the OR on a given OSN, a series of intracellular events take place, transducing the chemical information into action potentials. ORs, like rhodopsin, metabotropic glutamate receptors and some taste receptors, are GPCRs. When odours bind to their specific OR, the associated G-protein is activated and the α and $\beta\gamma$ subunits dissociate, and a second messenger pathway is activated. In this case this second messenger pathway is the activation of adenylyl cyclase and production of adenosine 3′,5′-cyclic monophosphate (cAMP) from ATP. This increase in intracellular cAMP levels opens cation selective channels, allowing calcium and sodium to enter the OSN, depolarising it and making the OSN fire action potentials (if the signal is strong enough). These action potentials are transmitted along the OSN axons out of the nasal epithelium through the olfactory nerve (cranial nerve I). At the glomeruli, OSNs make synaptic contact and activate mitral cells, which convey the chemosensory information to the brain (Schild & Restrepo, 1998). In contrast with other senses, the olfactory system lacks a topographic map of the sensory environment in the olfactory

cortex. Instead odours are associated with unique activation patterns of primary regions within the olfactory cortex, which correspond with associated activity patterns at the OSN and glomeruli levels.

Expression of OR varies from individual to individual. In humans, only a third of all OR genes present in the genome are expressed into receptor proteins, but this number is highly variable between individuals. Olfactory experience depends on which OR genes are expressed, and how many copies of a specific receptor each individual has. Two people, expressing 358 and 388 different OR, respectively, will both be 'normal', but the sensory experience associated with a given odour molecule for each one of them may be different. For instance, in a recent study, Kurz examined the perception of coriander smell and taste by different volunteers. They found people are 'lovers' and 'haters' of coriander in roughly equal parts. While 'lovers' are attracted by coriander's 'fantastically savoury' smell, 'haters' smell soap. This difference is apparently linked to the ability to detect some of the compounds present in coriander, the unsaturated aldehydes, that make 'haters' smell something like soap. 'Lovers', on the contrary, are insensitive to the unsaturated aldehydes, so do not detect a soap smell, leaving only the more pleasant characteristics of coriander to be detected by these individuals.

Key Takeaways

- Taste and smell are two senses specialised in detecting chemical compounds that reach the mouth or nose, respectively
- Taste sensory experience is the result of detection in a small number of dimensions, mainly salty, sour, sweet, bitter, and umami. Each sensory dimension is indexed by a specific type of taste receptor distributed along the tongue surface
- Smell detection is supported by a large number of odour receptor neurons, which are activated in a combinatorial fashion to give rise to molecule-specific activation patterns within the olfactory cortex
- 'Normal' smell sensation is highly variable between individuals and depends on the quality and quantity of odour receptors expressed between subjects.

References

Accolla, R., Bathellier, B., Petersen, C. C. H., & Carleton, A. (2007). Differential spatial representation of taste modalities in the rat gustatory cortex. *The Journal of Neuroscience, 27*(6), 1396-1404. https://doi.org/10.1523/JNEUROSCI.5188-06.2007

Andreou, A. P., & Edvinsson, L. (2020). Trigeminal mechanisms of nociception. In G. Lambru & M. Lanteri-Minet (Eds.), *Neuromodulation in headache and facial pain management: Principles, rationale and clinical data* (pp. 3–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-14121-9_1

Bereiter, D .A., Hargreaves, K. M., & Hu, J. W. (2008). Trigeminal mechanisms of nociception: Peripheral and brainstem organization. In C. Bushnell & A. I. Basbaum (Eds.), *The senses: A comprehensive reference: Vol. 5. Pain* (pp. 435–60). Academic Press. https://doi.org/10.1016/B978-012370880-9.00174-2

Breslin, P. A. S. (2019). *Chemical senses in feeding, belonging, and surviving: Or, are you going to eat that?* Cambridge University Press. https://doi.org/10.1017/9781108644372

Bushdid, C., Magnasco, M. O., Vosshall, L. B., & Keller, A. (2014). Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, *343*(6177), 1370–1372. https://doi.org/10.1126/science.1249168

Chandrashekar, J., Kuhn, C., Oka, Y., Yarmolinsky, D. A.,

Hummler, E., Ryba, N. J. P., & Zuker, C. S. (2010). The cells and peripheral representation of sodium taste in mice. *Nature 464*(7286), 297–301. https://doi.org/10.1038/nature08783

Doty, R. L. (2015). *Handbook of olfaction and gustation*. John Wiley & Sons. https://doi.org/10.1002/9781118971758

Fu, W., Sugai, T., Yoshimura, H., & Onoda, N. (2004). Convergence of olfactory and gustatory connections onto the endopiriform nucleus in the rat. *Neuroscience*, *126*(4), 1033-1041. https://doi.org/10.1016/j.neuroscience.2004.03.041

Hawkes, C. H., & Doty, R. L. (2009). *The neurology of olfaction*. Cambridge University Press. https://doi.org/10.1017/CBO9780511575754

Hummel, T., Iannilli, E., Frasnelli, J., Boyle, J., & Gerber, J. (2009). Central processing of trigeminal activation in humans. *Annals of the New York Academy of Sciences*, *1170*(1), 190-195. https://doi.org/10.1111/j.1749-6632.2009.03910.x

Papotto, N., Reithofer, S., Baumert, K., Carr, R., Möhrlen, F., & Frings, S. (2021). Olfactory timulation Inhibits nociceptive signal processing at the input stage of the central trigeminal system. *Neuroscience*, *479*, 35-47. https://doi.org/10.1016/j.neuroscience.2021.10.018

Price, S., & Daly, D. T. (2021). *Neuroanatomy, trigeminal nucleus*. StatPearls. https://www.statpearls.com/point-of-care/30606

Schild, D., & Restrepo, D. (1998) Transduction mechanisms in vertebrate olfactory receptor cells. *Physiological Reviews 78*(2), 429–66. https://doi.org/10.1152/physrev.1998.78.2.429

Sell, C. S. (2014). *Chemistry and the sense of smell*. John Wiley & Sons. https://doi.org/10.1002/9781118522981

Squire, L., Berg, D., Bloom, F. E., Du Lac, S., Ghosh, A., & Spitzer, N. C. (Eds.). (2012). *Fundamental neuroscience* (4th ed.). Academic Press. https://doi.org/10.1016/C2010-0-65035-8

Viana, F. (2011). Chemosensory properties of the trigeminal system. *ACS Chemical Neuroscience*, *2*(1), 38-50. https://doi.org/10.1021/cn100102c

Zhao, G. Q., Zhang, Y., Hoon, M. A., Chandrashekar, J., Erlenbach, I., Ryba, N. J., & Zuker, C. S. (2003). The receptors for mammalian sweet and umami taste. *Cell 115*(3), 255–266. https://doi.org/10.1016/s0092-8674(03)00844-4

About the authors

Dr Paloma Manguele
UNIVERSITY OF SUSSEX

Dr Paloma Manguele is a Research Fellow in the School of Psychology at the University of Sussex.

Dr Emiliano Merlo
UNIVERSITY OF SUSSEX

Dr Emiliano Merlo obtained a PhD in biology at the University of Buenos Aires, investigating the neurobiology of memory in crabs. He then moved to the University of Cambridge as a Newton International Fellow of The Royal Society and specialised in behavioural neuroscience, focusing on the effect of retrieval on memory persistence. Emiliano recently became a lecturer in the School of Psychology at the University of Sussex, where he convenes a module on the Science of Memory, and lectures on sensory and motor systems, and motivated behaviour in several undergraduate and graduate modules.

PART V
INTERACTING WITH THE WORLD

Human and non-human animals are the only organisms with brains. This unique structure allows us to not only perceive the world around but also interact with it to survive and thrive. From navigating our way to school or work, to selecting the right food to eat, or the right partner to interact with, our brain integrates sensory and internal information to produce the most appropriate behavioural responses. In this section we will analyse how the motor system is organised to execute actions, from simple reflexes to complex movements. We will then review the current understanding on how the brain integrates sensory and internal state information to produce the most adaptive behaviour given the circumstances. Later editions will also focus on how the brain integrates multi-modal internal and external sensory inputs to produce motivated behaviours such as feeding and drinking.

Learning Objectives

After reading this section you will be able to:

- recognise the components of the human motor system and the different structures involved in sensorimotor integration
- discuss how the motor system is modified by development and learning, and what is the effect of specific damages along the motor system components
- describe the participation of different brain systems in preparing and executing complex motor outputs.

12.

THE MOTOR SYSTEM

Dr Jimena Berni

Learning Objectives

After reading this chapter, you will understand:

- the organisation of the central regions and pathways involved in motor control
- the role of different regions for organising and controlling movement
- that motor systems are modified during development and by learning
- how motor systems break down when components are damaged.

Movement is key to every aspect of our lives. From breathing to walking, writing, or frowning, each behaviour is controlled by the motor system. So, understanding how movement is generated is an important step to understanding behaviour.

Despite being so 'natural', the generation of movement is a very complex task. Depending on the goal, the brain computes current and previously stored information to generate instructions and commands that are transformed into movement. This transformation is achieved at the neuromuscular junction, where a motor neuron synapses on a muscle governing its state of contraction. Therefore, to understand how purposeful movements are generated we need to understand how the nervous system is organised and how different regions communicate to control the correct sequence of contraction of hundreds of muscles that will produce the appropriate movement.

In this chapter we will discuss how studies have revealed the relationship between cortical organisation and function for the control of voluntary movement. We will look at how the spinal cord, which contains the motor neurons, is more than just a passive relay of brain information into muscle contraction. Finally, we will evaluate the function of the cerebellum and the basal ganglia for the organisation of movement. As we go along, we will discuss how the motor systems are modified during development and learning. We will also look at what happened when certain components are

damaged and how treatments or the ability of certain regions to change (plasticity) can help recovery.

Organisation of the motor system

The motor systems are used for multiple roles. They are involved in moving through and manipulating the world as well as for verbal and non-verbal (gesture) communication. They allow us to maintain posture and balance and control the contraction of the smooth muscles involved in autonomic functions like breathing and gut movements. Finally, they play a role in sensation, for example controlling the saccadic movements of the eyes as we visually track a stimulus. Despite the diversity of movements we perform, motor control is often considered as simple, probably reflecting that movements are seemingly effortless and largely unconscious. However, even simple movements require significant computations to coordinate the action of multiple muscles.

For example, imagine that you want to pick a ripe peach (Figure 5.1). This movement requires the concerted action of several regions of the nervous system, each with a specific role. Once you have decided to pick up the peach, visual information processed in the visual cortex is used to locate the fruit. This information is transmitted to the motor regions of the frontal lobe where the movement is planned, and command signals are sent. The commands are carried on to the

spinal cord, which is responsible for generating the movement through activation of motor neurons. The coordinated activity of motor neurons induces the contraction and relaxation of muscles in the arm and hand that allow the peach to be grasped. Now, a ripe peach is a very delicate fruit, and the correct amount of pressure needs to be applied to detach the fruit while avoiding bruising it. Sensory receptors in your fingers relay tactile and proprioceptive information back to the spinal cord and the somatosensory cortex. From there, the information reaches the motor cortex to confirm that you are grasping the fruit. Other areas are involved during this movement; the grasp force is judged by the basal ganglia, and the cerebellum helps in regulating the timing and accuracy of the movement.

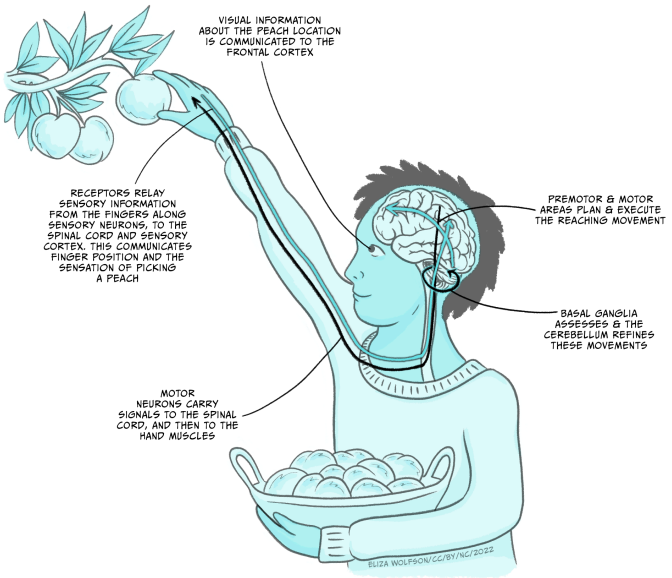

Fig 5.1. Schematic diagram of the steps and regions involved in a seemingly simple movement like picking a ripe peach

The motor hierarchy

In the diagram of interactions between different regions of the nervous system just described (Figure 5.1), each component controls a particular function. These regions are organised hierarchically.

The forebrain regions, involved in taking the decision, command lower functional areas like the spinal cord to execute the movement. Parallel processing allows us to simultaneously produce other movements, like maintaining posture while singing or walking. Finally, there is a level of independence in the function of these brain areas, which can co-ordinate complex activity in multiple muscle groups having received relatively general commands. This allows movement to happen rapidly, precisely and without conscious control.

Strategies to control movement

The concept behind how movement is controlled to be efficient has been debated for quite some time. When we execute an action, sensory information is used to inform us about the movement, the position of the body and the surrounding environment.

This information can be processed as the movement is

progressing allowing us to adjust it. This is called **feedback control,** where the output is monitored by various sensory systems and signals are relayed into the **CNS** to inform regions that generate motor outputs.

However, this model is limited to slow movements and sequential actions, since the processing of sensory feedback is relatively slow. For example, when catching a ball it may take 700 milliseconds to respond to visual clues, but the movement only takes between 150 and 200ms. This means that another motor control mechanism must be used for fast, ballistic movements.

In **feedforward control**, the optimal movement is predicted from current sensory conditions and from memory of past strategies. For example, if you open your front door and see snow and ice you will walk differently to how you would on a sunny day; you will take small steps, walk slowly, and hold your arms out for balance, because you know that there is a risk of slipping and falling. If we return to our ball example, knowing the initial conditions of the arm and hand and being able to predict the ball trajectory are used to choose a stored motor programme to catch the ball. A general feature of feedforward control is that it improves with learning.

Feedback and feedforward controls are not mutually exclusive and are combined to optimally generate coordinated movements.

To understand how the regions of the central nervous system work together to plan and command movements, we

will now analyse the role of the main regions, starting with the forebrain.

Key Takeaways

- Motor systems are used for multiple roles
- Motor systems consist of several regions that are hierarchically organised
- Motor and sensory systems work together to generate effective movement.

The forebrain and initiation of movement

In the frontal lobes of the brain, specific regions such as the **prefrontal cortex**, **premotor cortex**, and **primary motor cortex** contribute to movements in unique ways.

The prefrontal cortex is critical for making the decision to execute a particular action. For example, if you decide to grab your mobile phone to call a friend, it is the frontal cortex that reacts to that goal and instructs the motor system to initiate

movement. The premotor cortex receives information from the prefrontal cortex and prepares the required motor sequences, selecting the movements that are most appropriate for the action in the current circumstances. In this case, you will need to retrieve the phone and unlock it with a passcode, moving your fingers from one number to another in an organised sequence and following a specific memory. The information about the motor sequence to be executed is conveyed to the primary motor cortex that produces the required movements by muscle contraction and relaxation. Sensory input from the posterior parietal cortex, for example about where the phone and your fingers are, also shapes this process.

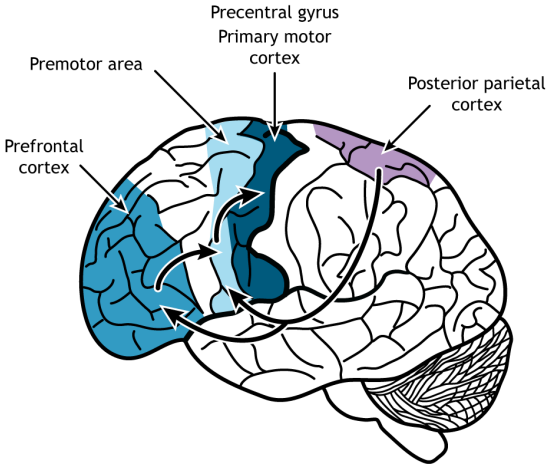

Fig 5.2. Sequence of activation of cortical regions that initiates a movement. The posterior parietal cortex is a global sensory integration centre and informs motor planning to take into consideration the current environmental conditions. The prefrontal cortex plans movement, the premotor cortex organises the movement sequence, and the motor cortex send the commands to generate movement.

The motor cortex

Evidence on the organisation of the motor cortex has been very influential in thinking about its function. Wilder Penfield was a neurologist who pioneered neurosurgery for the treatment of epilepsy that could not be controlled with medication. Through surgical interventions, he removed regions of the brain from which the seizure originated. To avoid catastrophic consequences, during surgery he electrically stimulated local regions of the nervous system in

awake patients and recorded the results. He found that different parts of the primary motor cortex controlled different muscles (Figure 5.3) (Penfield & Boldrey, 1937).

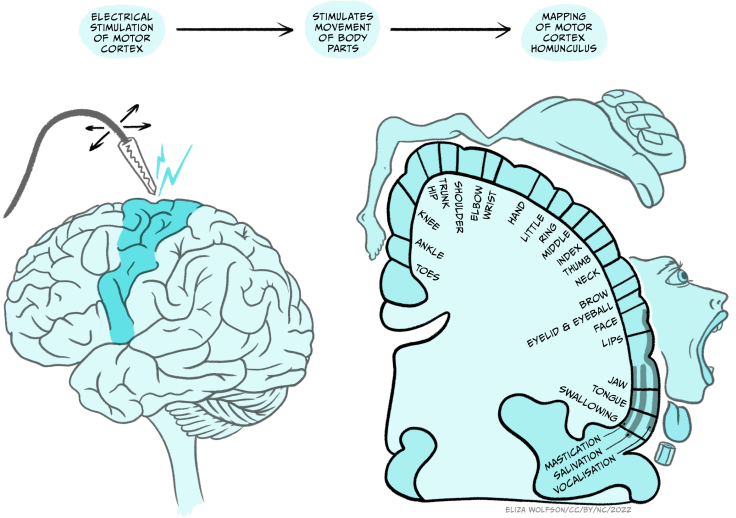

Fig 5.3. Wilder Penfield built a homunculus of movement located in the primary motor cortex. By stimulating specific regions of the motor cortex he found an organisation of the primary motor cortex that showed that the most ventral areas control head movement; more dorsal regions control the arms, trunk, and the legs. Body parts are not represented equally and muscle groups that need greater control like fingertips and tongue are over-represented.

This led to the drawing of a homunculus, which is a

topographical representation of how the primary motor cortex contains a motor map of the body. As with sensory maps, body parts are not equally represented. Areas that need greater motor control – hands, fingertips, lips, and tongue – are controlled by disproportionately larger regions of the motor cortex compared to other body parts.

The homunculus is a simplification and Penfield himself noted that the facial, arm/trunk and leg regions overlap. He attributed this to variability in brain size and the lack of precise stimulation, but more recent analyses have shown a fractured somatotopic organisation that sees neurons controlling movement of facial, arm/trunk and leg movements intermingled. This has generated controversy: does the primary cortex control muscles, or movement?

Modelling movement

A more detailed analysis of the relationship between the primary motor cortical areas and the movement they generate has helped in making sense of how motor cortex works.

From an anatomical point of view, there is evidence that single cortical neurons make direct connections with motor neurons that innervate multiple muscles that work together (they are synergistic) to produce a particular movement.

Furthermore, finger representation is found in several regions of the cortex. This suggests that the fingers, which are

involved in so many actions, can be linked to particular tasks and activated independently in different contexts.

These observations point to an organisation of the primary motor cortex to control movement, rather than the contraction of individual muscles. Different groups of neurons are grouped together, providing 'libraries' of muscle synergies that can be used for different movements or parts of movements. For example, a region of this cortex will be involved in activating the muscles required for picking up a marble between thumb and index finger.

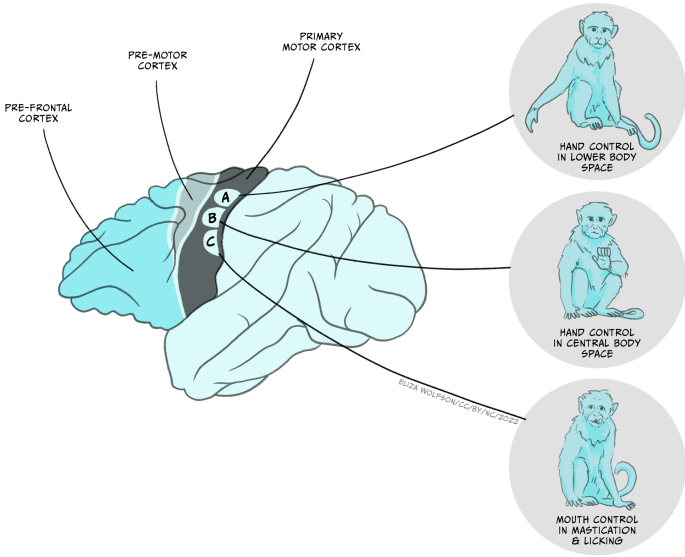

Fig 5.4. Action map in the Macaque motor cortex. Intracortical long stimulations (500ms) evoke ethologically meaningful actions.

Recent discoveries have shown there is substantial complexity in the movements that can be controlled by the motor cortex. Using longer electrical stimulations of the motor cortex (half a second) in macaque monkeys, Michael Graziano and colleagues have shown that long (half a second) electrical stimulations of motor cortex in macaque monkeys can evoke complex actions (Figure 5.4) (Graziano et al, 2016). These actions represent movements usually used by the monkey (**ethologically** relevant). For example, stimulating one area of the motor cortex repeatedly and reliably induced hand-to-mouth action (E). They also found sites evoking apparent defensive movements (F) or reach to grasp (G). Each ethologically relevant type of action is organised in zones, and ablations to these zones affect the ability to generate the corresponding movements. This zonal organisation of complex movements has been termed an action map.

Plasticity in the motor cortex

The cortical areas involved in the control of movement (prefrontal cortex, premotor cortex, and primary motor cortex) show an amazing plasticity. This means that the connections between neurons and their strength can change, new ones being made and old ones broken.

This is particularly obvious during development, when the nervous system is highly malleable, allowing for the maturation of new behaviours like walking for a toddler. In

humans, changes in the motor map also occur with the **acquisition of skilled movement**, like writing or playing the violin. The effects have been studied in detail in animals. At the beginning the motor map is absent but as the skill is learned the map is refined and becomes more precise. The changes are centred in regions that control the muscles involved in the learnt skill: each finger is controlled by a very defined region in the violinist primary motor cortex (Elbert et al., 1995).

This plasticity has also profound implications when the motor areas in the cortex are damaged. If a monkey damages a cortical motor area controlling its paw and does not undergo rehabilitation, this paw becomes paralysed. After a few months, an analysis of the motor cortex in that animal shows that the area controlling the monkey's paw (wrist and digits) has become smaller, while the lateral areas controlling the elbow and shoulder have enlarged. If animals are not allowed to use their good hand, by use of a cast for example, they are forced to use their bad hand. This is a form of rehabilitation as the areas that control the hand and digits then retain their size and the monkey retains some ability to move its hand (Nudo et al., 1996).

These experiments, performed in animals, have permitted the development of new rehabilitation treatments for humans. Amongst them, **constraint-induced movement therapy** helps improve the deficit that results from different types of substantial damage to the central nervous system

(CNS), such as stroke, traumatic brain injury, multiple sclerosis, cerebral palsy, and certain paediatric motor disorders (Taub, 2012). For example, in stroke patients, transcranial magnetic stimulation has been used to stimulate the damaged motor cortex or to inhibit the intact motor cortex in the opposite hemisphere and improve function (Ziemann, 2005).

The corticospinal tract

The main afferent route from the primary motor cortex to the brainstem and spinal cord is via the corticospinal tract. Most of the axons originate from pyramidal neurons in layer V of the cortex, but also include tracts from the premotor cortex and sensory cortex. The axon bundle descends into the brainstem where it sends several collaterals to brainstem nuclei and divides into two main branches. The opposite-side lateral tract controls movement of limbs and digits on the opposite side of the body. The same-side tract controls movements closer to the midline on the same side of the body, in particular movements of the trunk and shoulders that influence body orientation (Figure 5.5).

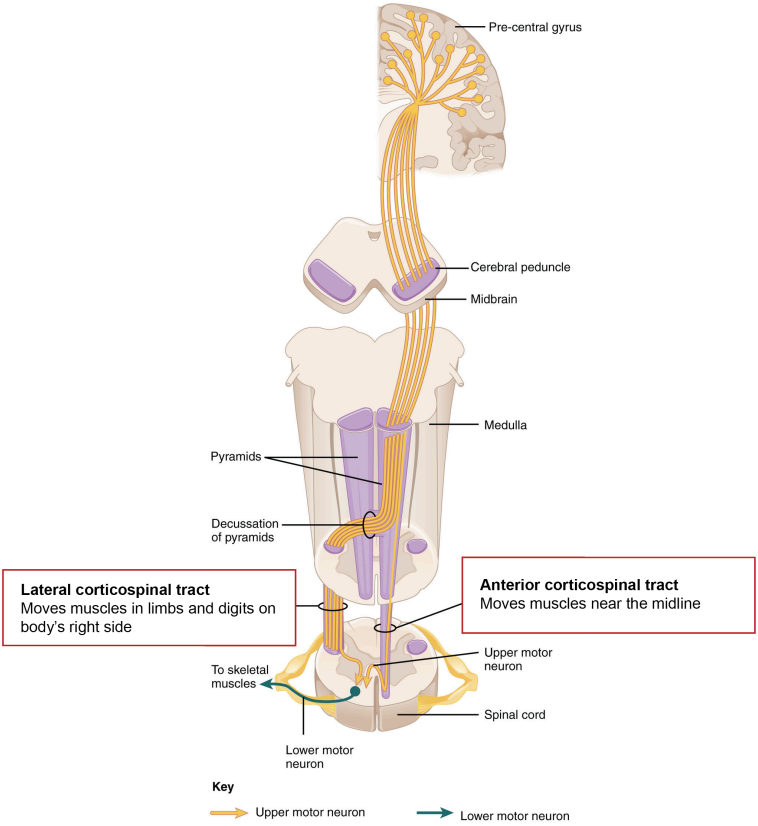

Fig. 5.5. Corticospinal tract of the left hemisphere. Nerve fibres descend from the cortex to the brainstem, where the tract branches. The lateral tract crosses the brainstem's midline, descending into the right side of the spinal cord to move limb and digit muscles on the body's right side. The anterior tract remains on the left side to move muscles at the body's midline.

Key Takeaways

- The forebrain organises the initiation of movement: the prefrontal cortex plans, the premotor cortex organises and the motor cortex sends commands to produce movement
- The primary motor cortex contains a motor map of the body: the homunculus
- Motor cortical organisation represents simple and ethologically relevant movements
- Plasticity is fundamental for learning new motor skills and for rehabilitation
- Descending corticospinal tract conveys inputs to the executive circuits in the brainstem and spinal cord.

The spinal cord

The spinal cord plays a fundamental role for the execution of movement. It contains the motor neurons responsible for

muscle contraction. It receives descending input from higher brain regions and the sensory feedback from muscles and from touch receptors. It generates the simplest movement: the reflex contraction. It also contains the circuits that control the generation of rhythmic movements, like walking or chewing. When it is lesioned, voluntary movement is impossible below the level of the damage.

A cross-section of the spinal cord [Figure 5.6a] reveals the outer white matter that contains the axon tract and the central grey matter where the nuclei of neurons from the spinal cord are located. The grey matter is divided into the dorsal horn that relays sensory inputs to the spinal cord and the brain, and the ventral horn that contains the motor neurons. In the intermediate grey matter, the interneurons that relay inputs to motor neurons are found. The spinal cord is divided into four sections: cervical, thoracic, lumbar and sacral, each comprising several segments (Figure 5.6b, left image). Limb muscles are supplied by nerves from several segments, reflecting the complexity of the movement generated.

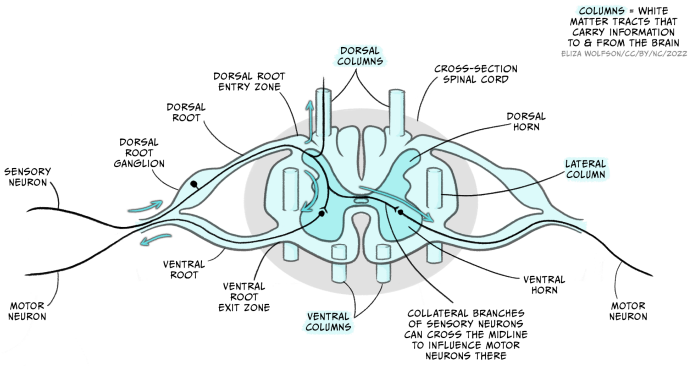

Fig 5.6a. Organisation of the spinal cord. A cross-section of the spinal cord showing the white matter (light blue) and grey matter (dark blue). The spinal cord connects the sensory and motor systems throughout the CNS. The grey matter contains most of the neuronal somas which makes it look darker. Sensory nerves enter via the dorsal horn and the motor neuron nerves exit from the ventral horn. The white matter contains many nerve tracts that receive and send information to other regions of the spinal cord or higher brain regions.

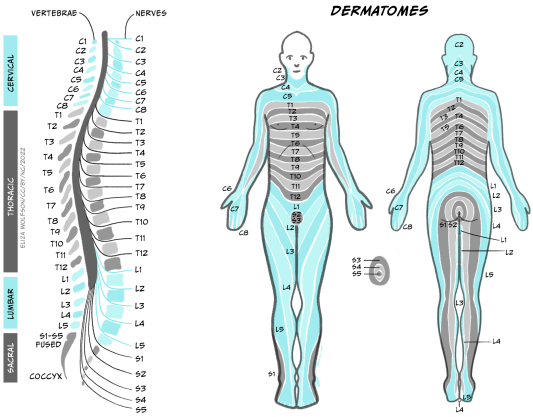

Fig 5.6b. The spinal vertebrae and nerves (left) and the correspo nding areas of the body sending touch informati on into the spinal cord (right).

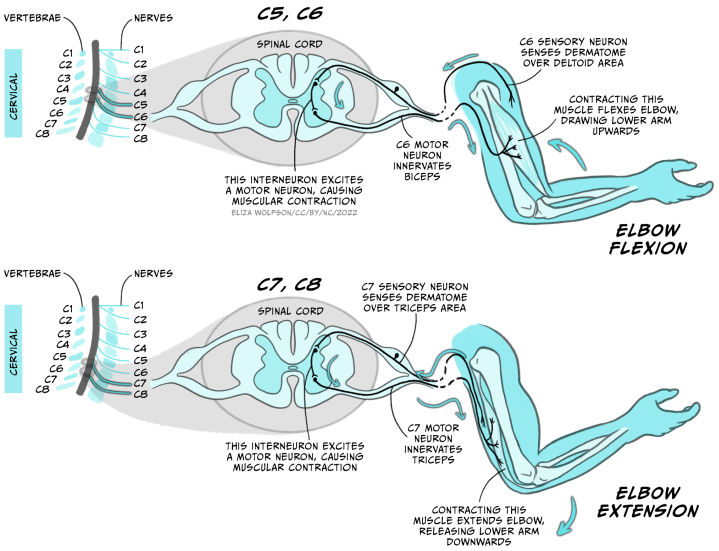

Fig 5.7a, b. Elbow flexion (a) and extension (b) are driven by motor neurons originating in different segments of the spinal cord.

The arm moves thanks to the coordinated stimulation of motor neurons that drive the contraction of extensor and flexor muscles.

For example, elbow flexion is mediated by cervical segments C5 and C6, while its extension is mediated by C7 and C8 (Figure 5.7). Sensory inputs from single strip of skin are supplied by individual spinal nerves, reflecting the importance of localised sensation.

The motor neurons

The motor neurons are the final output elements of the motor system. Each motor neuron innervates as many as 150 fibres of a single muscle (Figure 5.8a). This collection of fibres innervated by a single motor neuron constitutes the smallest unit of contraction, and was named the 'motor unit' by Sir Charles Sherrington (Nobel Lecture, December 12, 1932). Most muscles comprise hundreds of motor units.

By controlling the activity of each motor neuron and the number and type of motor units recruited, the type of movement and muscle force can be adjusted (Figure 5.8b).

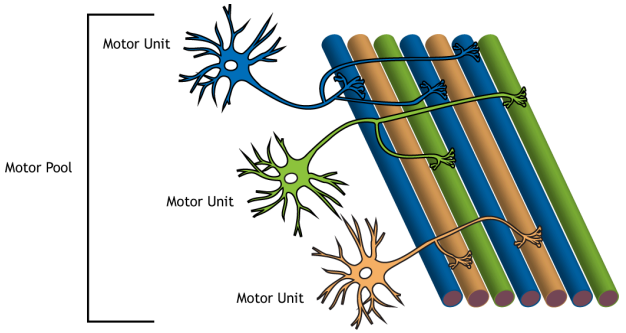

Fig 5.8a. Motor neurons can innervate more than one muscle fibre within a muscle. The motor neuron and the fibres it innervates are a motor unit. Three motor units are shown in the image: one blue, one green, one orange. Those three motor units innervate all the muscles fibres in the muscle and are the motor pool for that muscle.

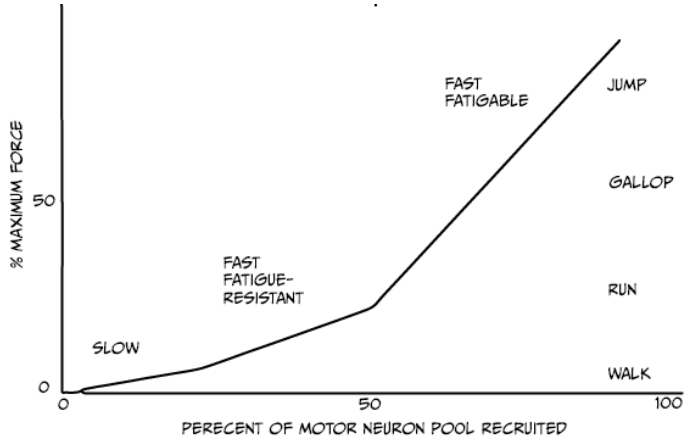

Fig 5.8b. Sequential recruitment of motor neurons as a function of the force exerted

Three types of motor units exist:

- **Slow motor neurons** generate a low and sustained tension, and are recruited first. They provide enough strength for standing or slow movements.
- Fast units generate more strength and are recruited for more intense activity. The **fast fatigue-resistant** units provide force for intermediate activity like walking or running.
- Finally, when intense movements are done like jumping, the **fast fatigable units** will be recruited.

The strength of contraction of each motor unit can also be modulated by changing the firing frequency of the motor neurons at the neuromuscular junction (NMJ).

The neuromuscular junction

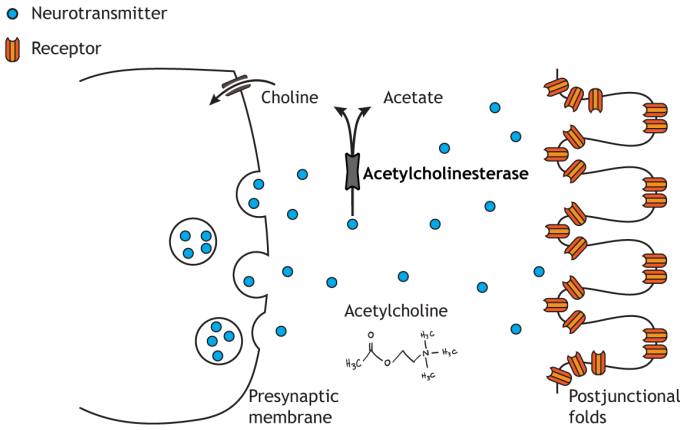

Fig 5.9. The neuromuscular junction (NMJ) is the synapse between a motor neuron and a muscle fibre. Acetylcholine is released at the NMJ and acts on nicotinic acetylcholine receptors located in the postjunctional folds of the muscle fibre. Neurotransmitter action is terminated by breakdown by acetylcholinesterase.

The neuromuscular junction (NMJ) is the chemical synaptic connection between the terminal end of a motor neuron and a muscle (Figure 5.9). It allows the motor neuron

to transmit a signal to the muscle fibre, resulting in muscle contraction. It begins when an action potential reaches the axon terminal of the motor neuron. In vertebrates, the neurotransmitter acetylcholine (ACh) is released from the axon terminal and diffuses across the synaptic cleft, where it binds to the nicotinic acetylcholine receptors (nAChRs) on the post-synaptic site on the muscle fibre. nAChRs are ligand-gated ion channels. ACh-binding opens the ion channel allowing Na ions into the muscle cell, depolarising the membrane At the muscle, this depolarisation is termed the 'endplate potential' (contrasting to the EPSP at a neuron to neuron synapse). This endplate potential causes an action potential in the muscle fibre that eventually results in muscle contraction. To prevent sustained contraction of the muscle, ACh is degraded in the NMJ by acetylcholinesterase.

The NMJ is the site of many diseases that affect the way messages are transmitted from the nerves to the muscles. For example, in congenital myasthenic syndrome, proteins required for synaptic transmission at the NMJ are mutated so an action potential in a motor neuron is less able to cause muscle contraction. This condition produces muscle weakness and impacts on mobility to different degrees, depending on the type of genetic mutation. Symptoms range from drooping eyelids and fatigue, to affecting breathing and other essential functions in the life-threatening forms of the disease. How to modulate the efficacy of NMJ transmission is

a very active area of research to help patients with this syndrome.

Generation of rhythmic movements

As we already mentioned, the spinal cord is not only a relay site from the brain to the muscles, but also plays a fundamental role in the generation of rhythmic patterns of movement, like walking or running. This means that circuits located in the spinal cord are capable of coordinating the concerted actions of several muscles. More than one hundred years ago, Charles Sherrington (1910) and Graham Brown (1911) performed the first experiments that showed that the spinal cord, disconnected from the brain, could produce the rhythmic movement of stepping in cats. After years of controversy and experimentation in many species, it is now accepted that the spinal cord contains circuits that generate rhythmic movements like chewing or walking independently of the inputs it receives. These circuits of interneurons are called Central Pattern Generators (CPG) and they ensure the coordinated action of muscles, so that extensors and flexors work in concert to produce fluid movements (Figure 5.7).

While the CPGs can generate rhythmic movements, they require pre-motor inputs that select and coordinate the types of motor neurons needed. For example, walking and running

require the contraction of muscles in the legs at different phases and with different intensities. During walking the duration of the stance is longer and the legs are less bent in comparison to running (Figure 5.10). The activity of the CPG that controls the timing and coordination of muscle contraction is influenced by descending inputs from higher centres (mostly primary motor cortex) that send the signal to select between gaits. Additionally, sensory feedback from the muscles (proprioception) and the environment shape the correct execution of movements (Figure 5.10b, bottom).

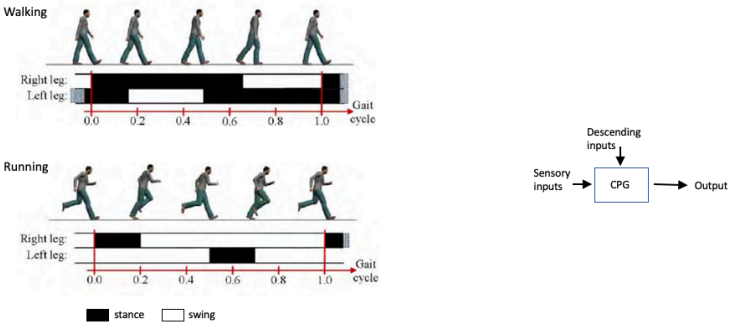

Fig 5.10. Top: Schematic of a walking man. The stance is more upright when walking and the legs are less bent. Bottom: The locomotion CPG is influenced by descending inputs from higher centres and by sensory information.

Spinal cord injury

The understanding of the importance of the circuits located in the spinal cord is used to help patients with spinal cord injury. When the spinal cord is severed, the circuit below the lesion site cannot be activated. When the lesion occurs at lumbar level (C4-C6), the arms and legs are paralysed, resulting in quadriplegia. If the lesion is at thoracic level, the legs are paralysed, resulting in paraplegia (refer to Figure 5.6c).

However, it is possible to improve the recovery of locomotion by training. During **step training** (Figure 5.11)**,** a patient's body weight is supported by a harness over a treadmill. Therapists and technicians move the legs and joints of the patient to simulate normal walking. As the patient walks, sensory inputs from the legs, the sole of the foot and the trunk are repetitively sent to the spinal cord. This trains the spinal cord circuit, and walking and standing are slowly relearned. After several weeks of training, most patients can generate spontaneous walking when placed on the treadmill with support. This enhances health and well-being. When patients have incomplete spinal cord injury, it can be the beginning of recovery since it also stimulates the rewiring of descending inputs from the brain.

Fig 5.11. Patient undergoing Step Training at the Christopher and Dana Reeve Foundation, which is dedicated to finding treatments and cures for paralysis caused by spinal cord injury and other neurological disorders.

See also Locomotor Training video on YouTube: https://www.youtube.com/watch?v=diZLK32DUts

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://openpress.sussex.ac.uk/introductiontobiologicalpsychology/?p=636#oembed-1

Key Takeaways

- The spinal cord has an organised structure
- The connection between a motor neuron and several muscle fibres comprises a motor unit – the smallest unit of motor output. Each muscle contains many motor units
- The activity and recruitment of motor units influences the motor output
- The neuromuscular junction is the cholinergic

synapse between the motor neuron and the muscle
- The spinal cord has neural circuits capable of generating rhythmic movements like walking and chewing
- Training the spinal circuits has a beneficial effect for the treatment of spinal cord injury patients.

The cerebellum and the control of skilled movement

The cerebellum comprises between 10 and 20% of the brain volume, but it contains 50% of its neurons. This disparity is possible because the cerebellum is a highly organised structure that allows the dense packing of neurons. It is located on the back of the brain and just above the brain stem. The cerebellum is divided into several regions, each with specific functions and connections to different parts of the brain (Figure 5.12).

The cerebellum contains sensory and motor components, but it is not necessary for the direct execution of movement. Rather it plays a role in the **coordination and planning of**

movement, which are affected in patients with cerebellar lesions.

The first insight into the role of the cerebellum was obtained by the neurologist Gordon Holmes (Holmes, 2022). After World War 1, he analysed the behaviour of soldiers that had been wounded by bullets and and presented with localised damage to the cerebellum. He observed that despite not presenting sensory loss, the movement of the patients was affected: they presented **cerebellar ataxia** (lack of coordination).

The patients presented weakness (hypotonia), showed inappropriate displacements like overreaching (dysmetria) and struggled to make rapid alternating movements (dysdiadochokinesis). Their movements seemed to be decomposed, with lack of coordination of different joints.

All these defects pointed to a role of the cerebellum in the construction of movement, contributing to coordination, scaling, timing and precision.

Interestingly, one of Holmes' patients described that 'the cerebellar lesion meant that it was as if each movement was being performed for the first time' (Holmes, 1922). This and other observations led to the current view that the cerebellum enables predictive motor commands to be made. This means that over repeated reiterations of a movement – for example, hitting a tennis ball with a racquet – an internal model of the movement is learnt: a motor programme. The next time you want to hit the ball, this cerebellar representation is used to

generate and construct the appropriate movements in response to the sensory inputs received, making your slide each time more accurate and 'automated' (remember feedforward control of movement).

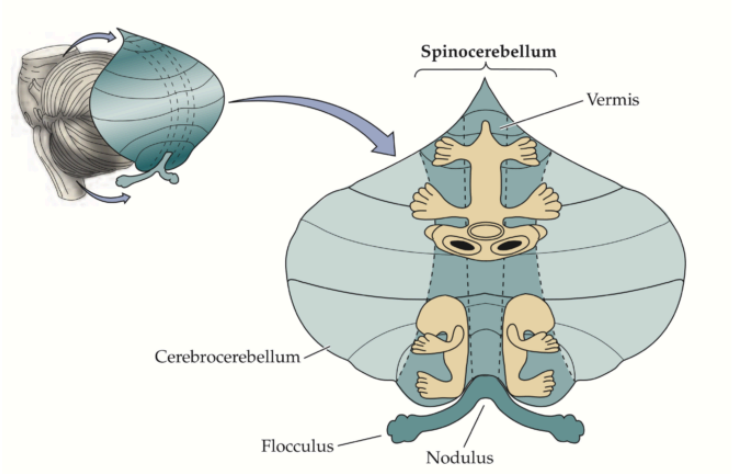

Fig 5.12. The projections in the cerebellum are largely organised somatotopically (the representation in the cerebellum parallels the position in the body) creating a somatotopic map.

The coordination of movement by the cerebellum is possible thanks to its high interconnectivity. It receives inputs about planned movements from the motor cortex, and sensory feedback on the actual movement. **This allows the comparison between planned and actual performance of the movement.** It produces a precise computation that uses

sensory information to adjust the ongoing movement as a part of a feedforward predictive control system.

Key Takeaways

- The cerebellum plays a role in construction of movement. Cerebellar lesions dramatically affect movement, because the timing, scaling and pattern of muscle contractions is inappropriate
- The cerebellum is important in translating 'sensory' signals into 'motor' coordinates, as part of a feedforward predictive control system
- It also influences motor learning, contributing to the automatisation of movements.

Basal ganglia

The basal ganglia are structures that modulate the motor function at the highest levels. They receive extensive connections from the neocortex and feedback to the motor

cortex. The basal ganglia participate in a wide range of functions, including action selection, association and habit learning, motivation, emotions and motor control. In this chapter we will look into their functional organisation and focus on the mechanism by which they allow the selection of movement and modulate movement force.

The basal ganglia are five interconnected nuclei within the forebrain located below the cerebral cortex. The main nuclei are the **striatum** (which means 'with stripes') formed by the **putamen**, the **caudate nuclei**, and the **globus pallidus**. And two midbrain nuclei, the **substantia nigra** and the **subthalamic nucleus** (Figure 5.13).

The ganglia receive inputs from all areas of the neocortex, comprising the motor cortex, as well as inputs from the limbic areas, which are involved in emotions, like fear. The nuclei project back to the motor cortex via relays in the thalamus influencing the descending commands from the primary motor cortex. There are no direct connections from the basal ganglia with the spinal cord.

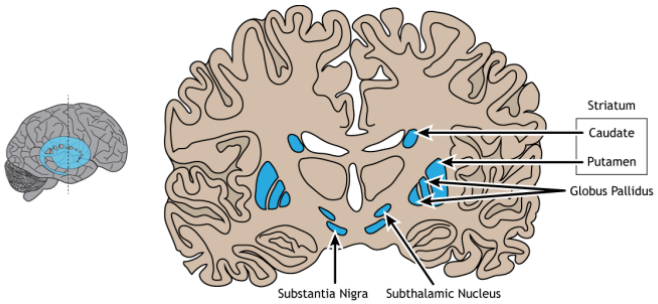

Fig 5.13. Anatomy of the basal ganglia that are composed of five interconnected nuclei within the forebrain. They are comprised of the caudate and putamen, which both make up the striatum, as well as the globus pallidus, substantia nigra, and subthalamic nucleus.

Functional network organisation: the volume hypothesis

In the **volume control theory** the globus pallidus acts like a volume dial. It projects indirectly to the motor cortex via the thalamus. The globus pallidus is inhibitory: this means that it inhibits the thalamus when activated. If this happens, the thalamus, which is excitatory, does not activate the motor cortex and this results in less movement. On the other hand, if the internal globus pallidus is inhibited, inhibition on the thalamus is released and movement can occur. This model suggests that it is through this 'volume control' that we make choices and select appropriate goals while rejecting less optimal options.

In this schema of the functional organisation of the basal

ganglia, the pathways towards the internal globus pallidus are critical in setting its output. There are direct and indirect pathways.

The direct pathway

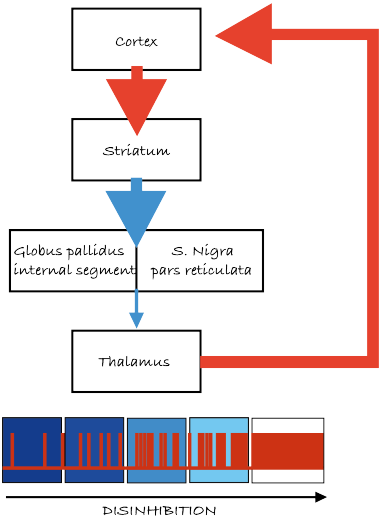

Fig 5.14a. Activation of the direct pathway facilitates movement

In the direct pathway (Figure 5.14a), the striatum (caudate/putamen) is directly connected to the internal globus pallidus and the substantia nigra. If the direct pathway is activated, it inhibits the internal globus pallidus, thus removing the inhibition of the thalamus. This facilitates movement by increasing thalamic excitation of the motor cortex.

Blue is inhibitory, red is excitatory, the thickness of the line indicates the strength of the connections.

The indirect pathway

In the indirect pathway (Figure 5.14b), the striatum projects to the external globus pallidus and subthalamic nucleus.

The striatum inhibits the external globus pallidus. This disinhibits the subthalamic nucleus which excites the internal globus pallidus. This results in less motor cortex excitation.

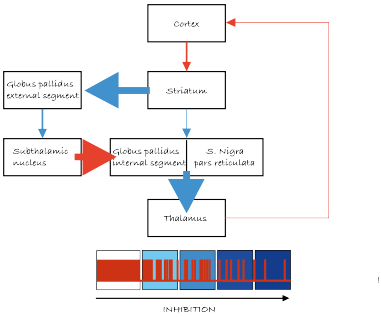

Fig 5.14b. Activation of the indirect pathway results in less movement

Dopamine also plays a role in the modulation of movements. Dopaminergic inputs to the basal ganglia from the substantia nigra pars compacta facilitate movement via both the direct and indirect pathways. In the direct pathway, activation of D1 dopamine receptors on neurons in the striatum enhances striatal inhibition of the internal globus pallidus, disinhibiting the thalamus and facilitating motor outputs. Conversely, in the indirect pathway, dopamine activates D2 dopamine receptors in the external globus pallidus to reduce its

inhibition. The external globus pallidus can therefore more strongly inhibit the subthalamic nucleus, reducing excitation of the internal globus pallidus and decreasing inhibition of the thalamus, further facilitating motor outputs.

Overall, the balance between the direct and indirect pathways controls the 'volume dial' that determines the strength of the basal ganglia output to the thalamus, thus acting to modulate the excitatory input received by the motor cortex to select and regulate movement.

Diseases of the basal ganglia

Damage to the basal ganglia can produce two main types of motor symptoms:

- **Hyperkinetic symptoms,** where there is excessive involuntary movement, as seen in **Huntington's chorea.**
- **Hypokinetic symptoms,** where there is a paucity of movement, as seen in **Parkinson's disease.**

Huntington's disease is a genetic disorder characterised by uncontrolled movements (chorea). The symptoms are excessive spontaneous movements, irregularly timed, randomly distributed and abrupt in character. It is followed by dementia and ultimately death.

There is evidence that the motor symptoms are originated

by neuronal death that can reach up to 90% in the striatum (caudate/putamen). This primarily disrupts the indirect pathway, where inhibition of the external globus pallidus is lost, producing a tonic inhibition of the subthalamic nucleus. This in turn reduces the inhibitory output to the thalamus, thus producing excessive movement.

The symptoms can be treated with antipsychotics that block dopamine transmission (e.g. clozapine) and decrease motor activity; as well as anxiolytics or anticonvulsants that increase inhibition via GABA (e.g. clonazepam).

Parkinson's disease is a slow progressive disorder that affects movement, muscle control and balance. It has three main symptoms: resting tremor, stiffened muscles, and slowness of movement that results in small shuffling steps.

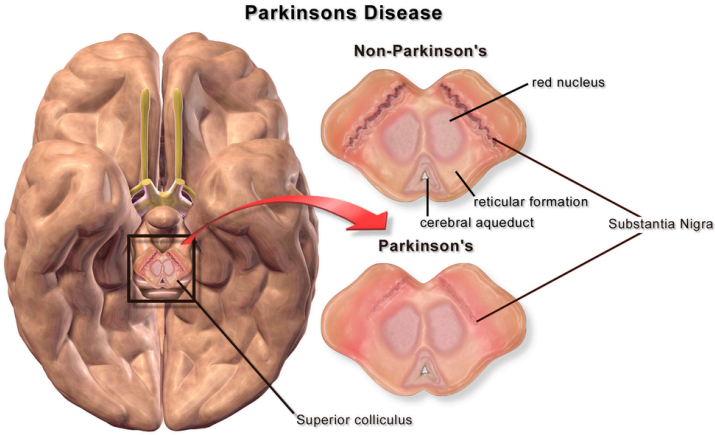

Fig. 5.15. A diminished substantia nigra (black substance) indicates the loss of dopaminergic neurons that results in Parkinson's disease.

It is produced by a loss of dopaminergic neurons in the substantia nigra and the levels of dopamine in its output regions are dramatically reduced (Figure 5.15).

Dopamine normally facilitates movement. When the levels are decreased, both the direct and indirect pathway are affected, increasing the inhibitory output of the basal ganglia and reducing motor activity.

Pharmacological treatment of Parkinson's disease largely focuses on restoring dopamine levels. Dopamine cannot be administered directly since it does not cross the blood-brain barrier, so does not reach the brain when systemically administered. Instead the dopamine precursor L-DOPA is

used, which is taken up by the brain and becomes active upon conversion to dopamine by dopadecarboxylase. Dopamine receptor agonists, or inhibitors of dopamine breakdown have also been used. These treatments are beneficial, but require gradual increases in dose over time, which can generate many side effects.

Alternatively, stimulation of the subthalamic nucleus or internal globus pallidus through implanted electrodes ('deep brain stimulation') has been introduced as a treatment for Parkinson's disease. This treatment can help relieve symptoms of Parkinson's disease, but it is not clear whether this is by inhibiting, exciting or more broadly disrupting abnormal information flow through the direct and indirect pathways (Chiken and Nambu, 2016).

See YouTube video 'Medtronics Deep Brain Stimulation Patient': https://www.youtube.com/watch?v=_tkmSn2m0Ck

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://openpress.sussex.ac.uk/introductiontobiologicalpsychology/?p=636#oembed-2

Key Takeaways

- The basal ganglia contribute to high level motor control
- Inputs to the basal ganglia arise from many regions of the cerebral cortex, outputs are directed to the frontal lobe
- Disorders of the basal ganglia involve limited or excessive movement as exemplified by Parkinsonism and chorea, respectively
- The basal ganglia also have important non-motor functions.

References

Brown, T. G. (1911). The intrinsic factors in the act of progression in the mammal. *The Proceedings of the Royal Society B*, *84*(572), 308-319. https://doi.org/10.1098/rspb.1911.0077

Chiken, S., & Nambu, A. (2016). Mechanism of deep brain stimulation: Inhibition, excitation, or disruption? *The*

Neuroscientist, *22*(3), 313-322. https://doi.org/10.1177/1073858415581986

Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., & Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science*, *270*(5234), 305-307. https://doi.org/ 10.1126/science.270.5234.30

Filh P., & Thomas B. (2010). Recognizing human gait types. In Ude, A. (Ed.), *Robot Vision* (pp. 183-208). InTech Open. https://doi.org/10.5772/9293

Graziano, M. S. A. (2016). Ethological action maps: a paradigm shift for the motor cortex. *Trends in Cognitive Sciences*, *20*(2), 121-132. https://doi.org/10.1016/j.tics.2015.10.008

Holmes, G. (1922). The Croonian lectures on the clinical symptoms of cerebellar disease and their interpretation. Lecture I. *The Lancet*, *199*(5155), 1177-1182. https://doi.org/10.1016/S0140-6736(00)55081-8

Holmes, G. (1922). The Croonian lectures on the clinical symptoms of cerebellar disease and their interpretation. Lecture II. *The Lancet*, *199*(5156), 1231-1237. https://doi.org/10.1016/S0140-6736(01)33076-3

Kandel E., Schwartz, J., & Jessell, T. (2012) *Principles of neural science* (5th Ed.). McGraw-Hill Education.

Kleim, J. A., Hogg, T. M., VandenBerg, P. M., Cooper, N. R., Bruneau, R., & Remple, M. (2004). Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning. *Journal of Neuroscience*,

24(3), 628-633. https://doi.org/10.1523/JNEUROSCI.3440-03.2004

Kolb, B., Whishaw, I. Q., & Teskey, G. C. (2019). *An introduction to brain and behavior* (6th ed.). Macmillan Learning.

Nudo, R. J., Wise, B. M., SiFuentes, F., & Milliken, G. W. (1996). Neural substrates for the effects of rehabilitative training on motor recovery after ischemic infarct. *Science*, *272*(5269), 1791-1794. https://doi.org/10.1126/science.272.5269.1791

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*(4), 389-443. https://doi.org/10.1093/brain/60.4.389

Sherrington, C. S. (1910). Flexion-reflex of the limb, crossed extension-reflex, and reflex stepping and standing. *The Journal of Physiology*, *40*(1-2), 28-121. https://doi.org/10.1113/jphysiol.1910.sp001362

Sherrington, C. S. (1932). Inhibition as a coordinative factor. *Nobel Lecture*. https://www.nobelprize.org/prizes/medicine/1932/sherrington/lecture/

Taub, E. (2012). The behavior-analytic origins of constraint-induced movement therapy: An example of behavioral neurorehabilitation. *The Behavior Analyst*, *35*, 155–178. https://doi.org/10.1007/BF03392276

Ziemann, U. (2005). Improving disability in stroke with

RTMS. *The Lancet Neurology*, *4*(8), 454-455. https://doi.org/10.1016/S1474-4422(05)70126-5

About the author

Dr Jimena Berni
UNIVERSITY OF SUSSEX

Dr Jimena Berni is a Senior Researcher at the Brighton and Sussex Medical School, University of Sussex. Her laboratory investigates the relation between neuronal circuits and behaviour with an emphasis on the diversification of circuits and the role of genes in specifying different neuronal networks and their assembly during development.

13.

SENSORIMOTOR INTEGRATION

Dr Emiliano Merlo

Learning Objectives

After reading this chapter you will be able to understand:

- the different levels of sensorimotor integration
- the involvement of different brain systems in preparing, executing and evaluating a behavioural action based on external and internal sensory information.

Animals are the only branch of living organisms that have brains and almost all of them do. (Note: There are a few exceptions. Sponges are simple animals that survive on the sea floor by taking nutrients into their porous bodies, and they have no brain or nervous tissue of any kind.) Current theories suggest that one of the main advantages of having a brain is to allow its carrier to move around and interact with the environment. Let's analyse an illustrative example: the sea squirt, a marine invertebrate animal, which has a very peculiar cycle of life (Figure 5.16).

In its juvenile form, the sea squirt swims around, looking for a suitable rock on which to attach itself. To do so, it uses a rudimentary central nervous system of around 200 neurons. Once attached, the animal becomes sessile (immobile), and eats its brain, a rich source of energy. For the rest of its life the sea squirt will remain immobile, so there is no longer any need for a brain. This fascinating example offers a strong support for the necessity of brains to generate adaptive behaviour by coordinating sensory information into motor action.

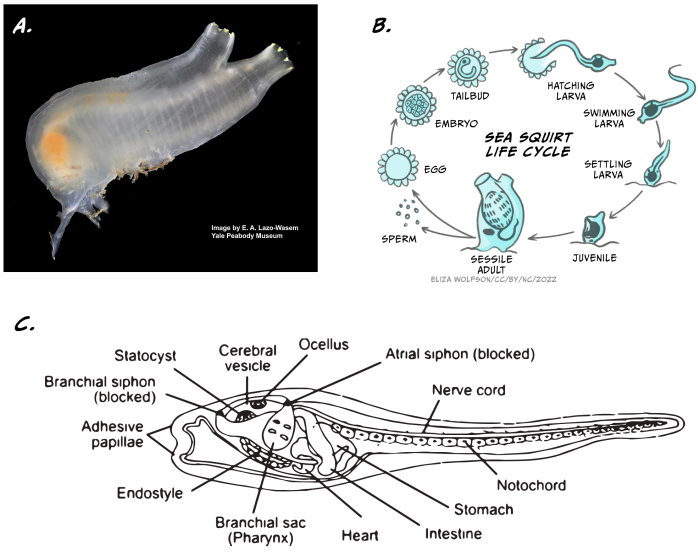

Fig 5.16. A: The adult sea squirt; B: Sea squirt life cycle; C: Anatomy of a larval sea squirt

In this chapter, we will explore how brains produce adaptive behaviour by acquiring information from the environment through the senses. We will start by analysing the simplest sensorimotor integration mechanism, the spinal monosynaptic reflex, and escalate in complexity all the way to explain the generation of a complex behaviour such as hitting a tennis ball with a racquet during a match.

Intuitively, becoming a World Chess Champion is a much more complicated task than moving a pawn one square forward in a chess game. Reality suggests otherwise. We humans have been able to build, after a lot of effort and

investment, a computer that is capable of consistently winning chess games against human World Champions. Nevertheless, we are still at the infancy of designing and building machines that have the manual dexterity of a small child when picking up a pawn and gently moving it one square forward. How is this possible?

In the game of chess there are a finite number of rules and movement possibilities, all of which are known to us. So, programming a computer that had sufficient computing power to calculate all movement possibilities and outcomes during a chess game was simply a technical challenge: considerable, but feasible. In 1997, an IBM computer called Deep Blue was the first machine capable of defeating the best human chess player of the time, the World Chess Champion Gary Kasparov. This was certainly an outstanding achievement, but Deep Blue was not physically moving the chess pieces but instead deciding on the next move given the current state of play. A human helper was required to physically move the pieces following the computer instruction.

Robots that can do physical actions mimicking humans are much more complicated to build and program than Deep Blue. Building and programming a robotic arm that can uncap a water bottle and pour a glass of water takes considerable amounts of money and the brains of lots of intelligent engineers. But the same robotic arm is unable to do other simple tasks for humans, such as tying shoelaces or breaking an egg to prepare a meal. Why is this the case?

Probably because robot designers are still not able to incorporate most of the fundamental rules that the nervous system uses to coordinate such tasks, based on real-time integration and analysis of noisy and multi-dimensional sensory information. In the following sections we will revise what is known about the basic principles ruling the sensorimotor integration, focusing on how and where sensory information is computed by the brain to produce adaptive and flexible behavioural outputs.

Sensorimotor integration: the minimal unit

One of the simplest structures to produce sensorimotor integration in humans is the monosynaptic spinal reflex (Figure 5.17).

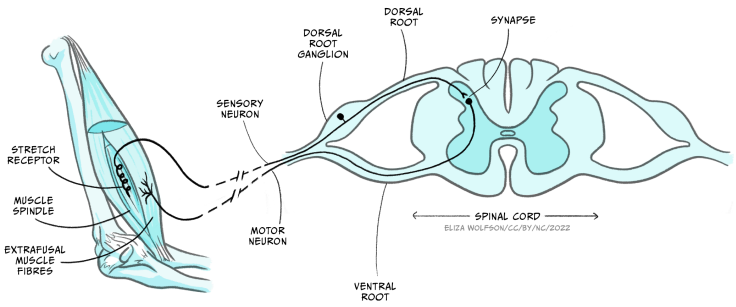

Fig 5.17. The monosynaptic spinal reflex

These reflex arches comprise one sensory neuron, originating in a target muscle, and one motor neuron, originating in the spinal cord and making a synaptic contact with the same target muscle. The sensory neuron collects information about the stretch status of the muscle fibres via stretch sensitive sensory terminals. When the muscle stretches beyond a certain threshold the sensory neuron is activated, firing action potentials that travel to its axon terminal. The sensory neuron releases neurotransmitters that activate the motor neuron, which in turn fires action potentials that travel to its axon terminal located in the target muscle. Activation of the motor neuron axon releases the neurotransmitter **acetylcholine** (ACh), which, via ACh receptor channels, results in muscular fibre contraction. This very simple circuit is an example of sensorimotor integration. In this case, the sensory information is rather simple: the sensory neuron is activated or not, depending on the muscle stretch surpassing the predetermined threshold. Also, the outcome is all or none: either the motor neuron fires action potentials, and contracts the target muscle, or not. This very simple sensorimotor integration system serves a very specific function of preventing overstretching of the target muscle, which can damage permanently the muscle fibres. The behavioural outcome is also simple, producing a muscle contraction to prevent injury, but well-suited for its biological function.

Since we claimed above that the reason for having a brain is to produce behaviours such as swimming or looking for a

suitable place to attach, you may be puzzled by the fact that spinal reflexes produce behaviours without involving the brain. One explanation for this apparent discrepancy is the biological function served by these reflexes. This becomes clear in the following home-based experiment on the knee patellar reflex in humans from Backyard Brains.

Applying a gentle hit to the patellar tendon (connecting the quadriceps muscle with the tibia) of a human volunteer, the quadriceps of the same leg contracts within 20 to 30 milliseconds. In contrast, if we instruct the volunteer (who is blindfolded so they cannot anticipate the upcoming hit to the knee) to contract the quadriceps of the other leg every time they detect the gentle hit on the target knee, the delay between the hit and the contraction increases to around 200 milliseconds. The contraction of the same leg that receives the hit is governed by a spinal reflex, whereas the contraction of the contralateral leg is controlled by the participant's voluntary decision to move it, a decision that involves the participant's brain activity. Given the biological function of the spinal reflexes, bypassing the brain allows for a faster response with higher chances of preserving tissue integrity.

However, most human behaviours taking place in our daily activities are a consequence of complex interactions of sensory information, internal state and response possibilities, which requires the computational power of the brain to maximise the benefits of action selection in real time in an ever-changing environment.

Behaviour in an ever-changing world

Let's explore the following scenario. One morning, you are ironing your clothes before attending a job interview. In a split-second distraction one of your fingers touches the hot iron, and you immediately and rapidly retrieve your affected hand and arm from contact with the iron's surface.

This everyday life example illustrates the function of another type of spinal reflex called the **polysynaptic reflex**. In this case, the sensory and motor neurons characteristic of the monosynaptic reflexes presented above are complemented by an interneuron making a synaptic bridge between them (Figure 5.18a).

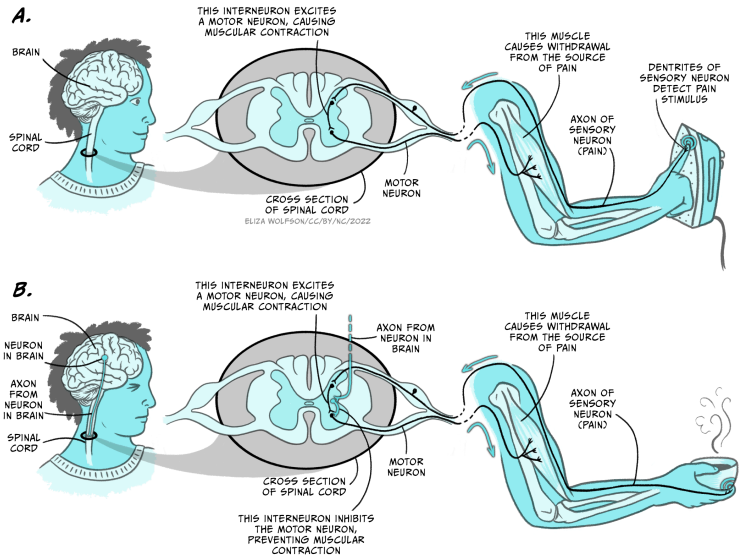

Fig 5.18a, b. The polysynaptic reflex

The function of this reflex is to prevent damage to the target body part and does not recruit or require brain activity.

But let's now imagine that after ironing your clothes, you prepare some coffee for breakfast. When you remove the recently-heated cup of coffee from the microwave you realise that you overheated the coffee, and the cup is too hot for you to handle it all the way to the table. As in the previous example, the corresponding spinal reflex will be activated by the heat to prevent damage to your hand, but you do not drop the hot cup. Instead, you look for a nearby surface on which to place the cup down without spilling its precious contents.

If your brain were not involved in this scenario you would

end up with a broken cup of coffee scattered all over the floor, but instead you managed to find a better solution and saved your hand and the coffee. This is an example of sensorimotor integration where action selection by the brain is key to produce an adaptive behavioural response. The polysynaptic spinal reflex that saved your hand from the hot iron needed to be inhibited in this case, and this was possible by the activation of additional interneurons descending from the brain (Figure 5.18b).

Your intention of drinking that coffee, and a plausible prediction of contacting a hot object when you were reaching for the cup inside the microwave, influenced your action selection and overrode the polysynaptic spinal reflex that was sensory-activated. This example illustrates why the brain is essential for this type of behavioural selection, since it can integrate sensory information from different sources, along with the internal state of the subject, to produce a more accurate and advantageous action at the right moment and time.

Neural pathways and structures involved in voluntary actions

Up until now we have revised how reflexes can control actions, and how the integration of multiple sources of information can alter predetermined reflex actions to produce more adaptive behaviours. Nevertheless, most of our daily actions

and behaviours are produced voluntarily, without an apparent involvement of mono or polysynaptic reflexes. Seemingly trivial movements like hitting a tennis ball with a racquet require a complex integration of sensory perception and analysis of internal state, including posture and muscle status, action selection and execution. Such a complex task is entirely up to the brain and involves detecting the looming ball through visual and auditory information, estimating the ball speed and area of bounce, approaching the target area and preparing the strike, and finally striking the ball with the centre of the racquet.

In this section we will revise the sensory and neural pathways, and body structures, necessary to produce such voluntary action. You have heard about many of these in previous chapters into sensory and motor pathways, but here we will consider in more detail how they work together to generate behaviour. We will also discuss some of the basic principles that the brain uses to produce the best possible solution for the problem, hitting the tennis ball back to the other side of the tennis court, as well as how the consequences of our actions can sculpt more refined sensorimotor integration processes, producing better actions.

Tracking the ball: audition and vision in action

If you ever played a tennis match, you may recognise that there

are two main sources of information when we are trying to track a looming tennis ball. Clearly, this task is mainly solved by the visual system, but the auditory system also plays a part.

For more experienced players, the sound of the ball being hit by the opponent is an early indication of the rough course the ball might follow. If the sound volume is very low or high, the probabilities of the ball not hitting the permitted section of are our side of the court are high. Also, the quality or frequency of the sound may also indicate if the ball may be worth tracking and preparing to return it. In those cases, we might even decide the ball is not worth tracking at all and we prepare for the next point. As discussed in the chapter **Perceiving sound**, both the volume and pitch of auditory stimulus are perceived in our inner ear by the structure called the **cochlea**.

Inner hair cells distributed along the basilar membrane can detect specific sound frequencies and codify the intensity of such frequencies by their action potential firing rate. The auditory information is translated into the language of the nervous system, action potentials, and it reaches the brain via the auditory nerve. Based on previous experience, the brain may be able to determine when the volume and pitch of a sound from the ball being hit by the opponent is more likely associated with a ball missing the target area of the court. In those cases, the motor command activated by the brain will be to prepare for the next point rather than tracking and striking the ball. But, if the sound is about right, then the visual system

and a more complex sensorimotor integration mechanism takes place.

The eye is responsible for translating visual information into action potentials. The moving tennis ball travelling at speed towards our side of the court constitutes a looming object that occupies gradually more space on the retina surface. Both eyes will detect the ball, and the visual information will be integrated in the brain to estimate not only the direction of the ball but also its speed. As the ball moves towards the near side of the court, the eyes will move aiming to maintain the object in focus within the **fovea**. Keeping the image of the ball within the fovea will give the player the best visual resolution in daylight, maximising the capacity to detect the ball as it travels in a luminous environment. Photons bouncing on the tennis ball that arrive to the fovea will excite a collection of photoreceptors. These specialised cells will translate the visual information into electrical information via the activation of the photopigment and specific ion channels. Changes in the photoreceptor membrane potential lead to activation of the bipolar and ganglion cells, which convey the visual information, now converted into action potentials, to the brain.

The visual information arriving at the brain will play different roles in different motor outputs during action selection and execution. Early visual processing will be required for tracking the moving ball. Empirical research has determined which are the neural pathways involved in object

tracking, and how this information is used to produce motor commands for eye movement. Rapid eye movements, called **saccades**, are used to track the ball and acquire information about the environment. This will be particularly important as we start approaching the area where the ball may bounce, since we not only require keeping an eye on the ball, but also moving safely and effectively within the court.

In the laboratory, visual attention of healthy volunteers can be traced by tracking the position of the eyes in real time. Using **electroencephalography** or brain imaging techniques in combination with eye tracking, we have learnt that saccades are mainly controlled by the oculomotor loop involving the cerebral cortex, the basal ganglia, and the thalamus (Figure 5.19).

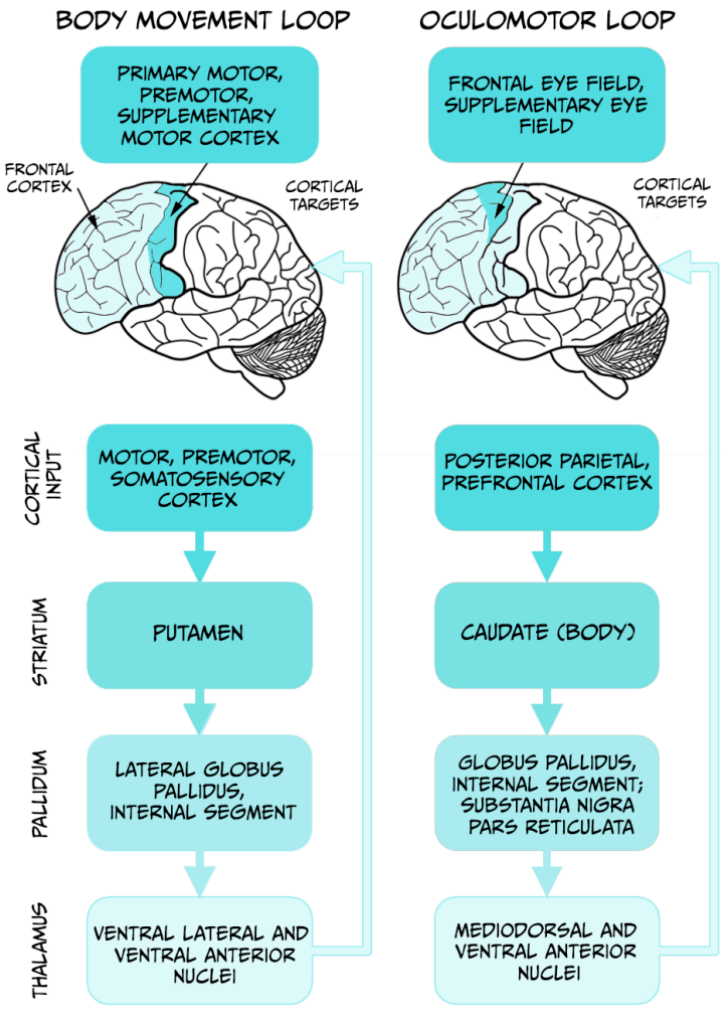

Fig 5.19. The body movement and oculomotor loops

Neurons in the posterior parietal cortex, an associative region that receives visual and motor information, increase their firing

rate just before a saccade is observed. Lesions affecting this region destroy the capacity to perform saccades and produce a condition called 'spatial neglect' in humans, characterised by attentional deficit in the visual field contralateral to the injured brain hemisphere.

The posterior parietal cortex sends connections to several nuclei of the basal ganglia controlling eye movements. One of these nuclei, the **superior colliculus**, contains visual fixation neurons. These cells are activated immediately after a saccade takes place, and keep firing during eye fixation, inhibiting eye movement away from the target location.

Hence, the concerted action of several key brain regions is responsible for tracking down the tennis ball approaching to our side of the court. This information is valuable, but simply tracking a moving tennis ball does not mean we will be able to hit it back with a racquet. How can we use this information to prepare our action of returning the ball to the other side?

In the following section, we will analyse how the different sensory information streams are integrated to coordinate this action.

Integration of visual information: navigating the court towards the ball

The visual information carried out by the optic nerve follows

parallel pathways for the analysis of different attributes of the visual sensory experience. Two main pathways are distinguished by the involvement of the **primary visual cortex V1**.

- In the **geniculostriate** pathway, the visual information from the optic nerve arrives to the lateral geniculate nucleus of the thalamus and then follows into the primary visual cortex V1.

The geniculostriate pathway divides the visual information into the **dorsal** and **ventral** streams. In our example, the dorsal stream will be responsible for perceiving the motion of the ball and the spatial relationship between the ball and myself (the so-called *how* information). The ventral stream will be responsible for determining the contrast, contour, and colour of the tennis ball (the so-called *what* information).

- In the **tectopulvinar** pathway, the visual information carried by the optic nerve is relayed into the superior colliculus, a region of the midbrain, and then follows into the pulvinar nucleus of the thalamus.

The tectopulvinar pathway determines the spatial location of objects in the environment, allowing us to navigate without hitting stationary objects. This visual pathway is independent of V1, and allows for an effective navigation of the tennis court

avoiding stepping on stationary balls or other potentially dangerous objects.

As usual in neuroscience research, analysis of brain lesions and their consequences are key for understanding brain functioning. Some individuals who suffer a stroke affecting the primary visual cortex V1 are technically blind. They fail all tests for detecting objects or recognising others and places. This is due to the disruption of the geniculostriate pathway, that analyses the *how* and *what* of the visual experience, and supports the conscious experience of seeing.

Nevertheless, these patients can solve a visual navigation test (watch a video example of a visual navigation test). If they are left alone to walk down a corridor with different objects scattered along the path, the patients manage to navigate on their own without tripping, even if they do not experience conscious visual perception. This condition is known as blindsight (see also Box 9, Lighting the world: our sense of vision) and the remarkable behavioural observation is explained by the functioning of the tectopulvinar pathway, which does not use V1 for determining the position of objects in the environment. This fascinating observation is a good example of how neuroscientific research reveals brain functioning by analysing the effect of focal lesions in different brain regions.

But let's go back to the moving tennis ball and how the brain uses this information to produce an action.

With all the visual information flowing through the

different pathways, together with the interoceptive information regarding our internal state and the position of our legs and arms, the brain is making a continuous integration and selecting the right action for the right time. The rules the brain follows to make such decisions are currently a matter of intense focus in basic neuroscience research. One hypothesis is that the brain is constantly producing a **Bayesian analysis** of the world based on sensory information (Körding et al., 2007), using prior experience modulated by ongoing information to calculate the most probable outcome. The brain possesses prior information on where the ball is likely to bounce, which is generated from previous experience playing the game. For instance, very good tennis players aim for the ball to bounce near the court lines, which makes it more difficult for the adversary to return it. This prior information (the likelihood that the ball will be bouncing close to the court line) is combined with the live sensory information of where we estimate the ball is going to bounce. Hence, our estimation of the actual likelihood of the ball bouncing at a particular location on our side of the court is a product of overlapping the prior information with the present information. Our brain will then produce a prediction of where the ball is likely to bounce, and we will approach that position to prepare for the action. As the ball gets closer to the floor, the prediction based on sensory information becomes more accurate and so does the selection of the right behavioural action for those set of conditions.

Action!

Now that the brain has integrated the available sensory information and predicted where the ball is going to touch the floor, it is time to execute the motor command of hitting the ball after it bounces. Execution and control of voluntary motor sequences are performed by motor loops involving different regions of the cerebral cortex and the basal ganglia.

In our example, the two main motor loops involved are the oculomotor and body movement loops (see Figure 5.19, above).

As we mentioned in the section **Tracking the ball**, the oculomotor loop receives sensory and interoceptive information to control eye movement necessary for following the moving ball. The body movement loop controls the hundreds of muscles necessary for performing actions, and involves a serial connection of motor, premotor and somatosensory cortices with areas within the basal ganglia and the thalamus. The thalamus sends feedback connections into the early regions of the cortex. This neuronal circuit is key for constant monitoring of the current action and allows for modification of actions while they are being executed. The striatum and globus pallidus within the basal ganglia are important for action selection, initiation, and termination of motor actions (as seen in the example of eye saccades), and for

relating actions with their consequences. According to tennis instructors, to hit a tennis ball properly requires a refined coordination between the position of the ball and the movement of the racquet. To achieve this goal, it is important to keep the eyes tracking the ball at all times, even when we are hitting the ball. To achieve this goal, the **cerebellum** needs to get involved (Miall et al., 2001). This brain region is activated during tasks that require high coordination between eye tracking and hand movements.

In addition, the execution of any of the movements mentioned so far, as well as the action of hitting the ball, will require the activation of the motor homunculus maps of the motor cortex. All the regions controlling the movement of the participating limbs and muscles will be recruited by the motor command during the whole exercise.

Lastly, activation of motor neurons by these motor commands will produce the firing of action potentials that will travel to the axon terminals. As we heard in the previous chapter, the synaptic contact between the motor neuron axons and the muscle fibres is a specific type of synapse called the **neuromuscular junction**. When the axon is activated by the arrival of one or more action potentials the internal concentration of the Ca^{2+} ion increases, increasing the probability of release of synaptic vesicles containing the neurotransmitter acetylcholine. Release of ACh into the synaptic cleft will activate ACh receptor channels expressed in

the muscle fibre membrane, driving the depolarisation of the cell membrane and contraction of the muscle fibres.

The coordinated contraction of specific muscle and muscle groups are the outcome of a complex sensorimotor integration and coordination system. Even after the action is executed, the senses and the brain will continue to monitor the environment analysing its consequences.

Behavioural outcome and prediction error

Whether or not we were able to hit the ball, and depending on the outcome of that action, the brain will integrate this information through the reward system dependent on the neurotransmitter **dopamine** (Figure 5.20).

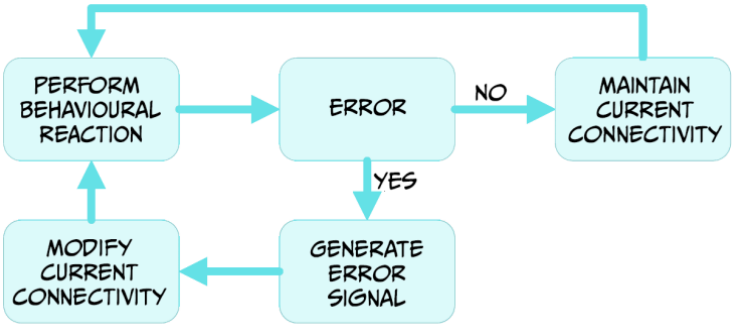

Fig 5.20. Dopamine reward system

When the outcome matches our expectations, there is no error signal. The brain has mechanisms to maintain the neural relationships responsible for that behaviour as an adaptive response for similar scenarios. If the outcome differs from the expected results because we miss the ball completely, or it hit the net or went flying past the bottom court line, a prediction error signal is generated in several regions of the brain, by the release of dopamine (Schultz, 2000). This dopamine signal will affect the way different regions of the brain connect to each other, allowing for the modification of the action of approaching or hitting the tennis ball in future encounters. The reward system and the prediction of specific outcomes allow for the sensorimotor integration mechanism to learn from its own performance, allowing improvement of actions.

Key Takeaways

- Sensorimotor systems have evolved in animals to generate adaptive behavioural responses to environmental and internal stimuli.
- A simple action of a tennis ball during a match

requires coordinated activity of a myriad of brain systems.

- Sensory information, relayed in real time to the brain, is key for selection of most appropriate motor actions for task performance.
- Memory, as prior knowledge, is also key for motor action selection.
- Producing an action involves several cortex-basal ganglia loops (oculomotor and skeletomotor), as well as the cerebellum.
- The dopaminergic system is often key for coordinating the maintenance or modification of motor actions.

References

Körding, K. (2007). Decision theory: What 'should' the nervous system do? *Science 318*(5850), 606–10. https://doi.org/10.1126/science.1142998

Miall, R. C., Reckess, G. Z., & Imamizu, H. (2001). The

cerebellum coordinates eye and hand tracking movements. *Nature Neuroscience* *4*(6), 638–44. https://doi.org/10.1038/88465

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience* *1*(3), 199–207. https://doi.org/10.1038/35044563

About the author

Dr Emiliano Merlo
UNIVERSITY OF SUSSEX

Dr Emiliano Merlo obtained a PhD in biology at the University of Buenos Aires, investigating the neurobiology of memory in crabs. He then moved to the University of Cambridge as a Newton International Fellow of The Royal Society and specialised in behavioural neuroscience, focusing on the effect of retrieval on memory persistence. Emiliano recently became a lecturer in the School of Psychology at the University of Sussex, where he convenes a module on the Science of Memory, and lectures on sensory and motor systems, and motivated behaviour in several undergraduate and graduate modules.

14.

MOTIVATED BEHAVIOUR: NUTRITION AND FEEDING

Dr Kyriaki Nikolaou and Professor Hans Crombag

Learning Objectives

By the end of the chapter you will understand the processes involved in:

- basic mechanics of homeostatic system, including temperature regulation
- mechanisms in brain, body including gut, involved in drinking and feeding

- how homeostatic mechanisms that regulate physiological variables around a set-point, can deregulate to vary away from a set-point, including learning mechanisms, and more complex ones involving desire and hedonics.

Why do we get up in the morning? This may be an oft-heard question about the causation of our action, but why *do* we get up in the morning?

A complex causal chain of events drives the 'getting up' behaviour: basic physiological regulatory mechanisms involving brain stem nuclei and hypothalamic photosensors involved in regulating our circadian wake-sleep rhythm; gut hormonal mechanisms that interact again at the level of hypothalamic nuclei signalling hunger and our need or desire to eat, maternal mechanisms involving e.g. oxytocin that drive our instinct to nurse our infant child, dopaminergic forebrain mechanisms that regulate our desire for earning rewards at work (i.e., a salary), as well as more diffuse and long-term cognitive expectations about what the day, week, year and career may bring us, and so forth.

The study of motivation cuts across psychological domains to understand principal mechanisms that cause our behaviours, whether they are basic, essential regulatory

behaviours such as drinking, eating/feeding, fighting and desire for sex, or more complex, psychologically-driven behaviours. We will focus principally on feeding, though will touch briefly on temperature regulation as a model regulatory physiological system that underlies motivated action, and drinking as a motivated behaviour aimed at maintaining hydration levels to ensure optimal physiological, neuronal and psychological function.

Motivation can be defined as an internal state that explains why we behave or why we learn to behave, and the study of motivation focuses on understanding what causes, drives and energises behaviour. We can therefore use terms like motivational states, motivational drives, and motivational desires to describe motivation. There are broadly two main classes of motivated behaviours: those that are 'regulatory' in nature and those that are 'non-regulatory' in nature.

We find so-called **homeostatic regulatory** mechanisms at the foundation of those motivated behaviours essential for basic survival needs; mechanisms that regulate our thirst (and thereby levels of (de)hydration), our sense of hunger and satiety (and in doing so, our levels of nutrients, including carbohydrates, fats, vitamins, and protein). These mechanisms encompass complex physiological mechanisms by which nutrients, water and salts are absorbed, distributed, released and excreted, but also behavioural consummatory mechanisms namely drinking and eating, and appetitive behaviours that

direct us to approach locations in the environment, and then make us work for water and nutrients to consume.

But regulatory homeostatic mechanisms, as essential as they are, are only part of the story of motivation; we do not just eat because we lack nutrition, nor do we only drink when we are dehydrated. We do not just have sex to procreate, or run when we are scared. Human motivation is often not regulatory in nature, requiring explanations beyond homeostatic mechanisms. We will explore these 'higher motivations' that rely on a complex set of brain structures forming a ring around the thalamus in the human and non-human forebrain, described early in the history of neuroscience by Paul Broca (known also for identifying Broca's speech production temporal lobe area), and built on throughout the years by Papez, McClean, and more recently the likes of Frederik Toates and Kent Berridge.

Considering basic regulatory mechanisms of feeding and thirst, preceded by a very brief consideration how our body (and brain) and the thermostats in our houses and offices regulate temperature requires us to consider research from early and mid last century, mostly involving American, British and western European scientists, but the field has burgeoned over the decades into a diverse and inclusive scientific community. In Box 1 we consider B. F. Skinner. A typical scientist of his generation, because he was an American (Caucasian) male professor at Harvard; but he was also atypical – or unexpected – because he was a vocal opponent

of the study of motivational mechanisms and states as genuine targets of scientific inquiry by psychologists.

Motivation as a homeostatic negative-feedback mechanism

Box 1: Behaviourism and the study of motivation

One common tool used by experimental psychologists for studying motivational processes (especially in non-human animals) is the operant or instrumental conditioning chamber and was (somewhat ironically, given his opposition to the study of motivation) designed and developed by the American scientist B.F. Skinner (1904-1990), who used the chamber for his seminal studies on the experimental analysis of behaviour to show how 'behaviour is shaped and maintained by its consequence'. If, for example, the consequence of a behaviour is generally positive because it leads to a rewarding outcome (e.g. the delivery of food for a

hungry animal), or because it leads to avoiding an unpleasant outcome (e.g. the avoidance of a loud sound), then the animal will learn to repeat that behaviour, i.e. the behaviour is reinforced and the outcome of the behaviour is considered a reinforcer of the behaviour. Interestingly, however, Skinner also believed that trying to understand any internal states that may make an animal seek the reinforcing outcome more in some cases than others (e.g. seeking food when hungry vs. when satiated) distracts from understanding the effect of the reinforcing outcome or the reinforcer on behaviour, writing that 'Mentalistic terms associated with reinforcers and with the states in which reinforcers are effective make it difficult to spot functional relations' (B.F. Skinner, *About Behaviorism*, 1974).

Nevertheless, the methods that Skinner developed to study how the outcome of a behaviour affects whether the behaviour will be learned and repeated are also used today to delineate the motivation behind the execution of a behaviour.

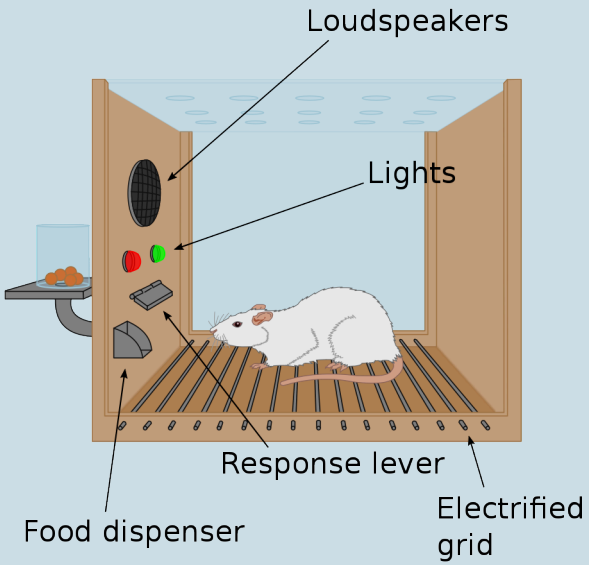

Fig 5.21. Illustration of Operant Conditioning chamber, or Skinner box

Operant conditioning chambers (or Skinner boxes), typically contain a lever and a food/sugar-pellet dispenser cup on one side of the chamber. They also typically contain signal lights of different colours and speakers through which sound tones can be played. They may also contain an electric grid through which mild electric shocks (negative stimuli) can be

delivered. Experiments involve animals learning that simple presses of the lever result in the delivery of food rewards. In other experiments, animals may learn that a tone or a particular light may signal that food becomes available in the dispenser cup.

Motivation researchers may explore the parts of the brain that are involved in the behaviour of pressing the lever for food when the animal is hungry and must alleviate this internal motivational state of hunger (i.e. regulatory motivated behaviour). In other experiments, motivation researchers may want to explore how other factors, such as learned associations between a light or a sound tone and food, may instigate pressing the lever for food even in an animal that is full, and thus explore brain regions involved in motivated behaviour that is not regulatory in nature.

Clark Hull (1884-1952) proposed that motivated behaviour is principally determined or driven by the need to alleviate an internal state of deprivation. Said simply, food reinforces a feeding behaviour if and **because** it alleviates a hunger state. Thus, the state of hunger is the internal state of deprivation and therefore the motivation to eat. Hull's Drive Reduction theory (1943) emphasised the importance of maintaining

homeostasis as the drive or motivation behind behaviour, and suggested that if this homeostasis, or balance in the internal environment of the organism, was taken away, this would lead to increases in arousal that would initiate action to bring back the balance. Thus the goal for an organism is to remain in homeostasis and to reduce any drives or motivations that arise from an imbalance in the system, so to reduce the arousal.

The organism can restore balance in its internal environment by acting to minimise the difference between the current state of an organism and a set point, which is the point that the organism wants to be at in order to be at equilibrium, and function optimally. In order for our body to work properly, certain variables in our body must be maintained within narrow limits. As humans, we have optimum set points for body temperature (36.5-37.5 degrees Celsius; °C). We also have optimum set points for levels of hydration and levels of different nutrients. The systems in our body that control body temperature, hydration and levels of nutrients, are homeostatic systems that bring the system towards equilibrium at particular set points. If the body's state deviates from these set points, the homeostatic processes that control them become active so that actions and behaviours (e.g. putting a jacket on when it's cold and body temperature drops or moving to shaded or cooler spots when it's hot and body temperature increases), or physiological mechanisms outside our conscious control (e.g. immune responses) can be activated to restore equilibrium.

These early drive reduction views about what motivates behaviour rested on the idea of negative feedback proposed by Walter Cannon (1871 – 1945). He was a doctor and a medical researcher in the First World War, and proposed that homeostatic systems maintain balance via negative feedback. Negative feedback is a process by which the effect produced by an action serves to diminish or terminate that action. Negative feedback mechanisms are the primary way by which homeostatic systems can reduce the difference between a point the body is actually at and the ideal set point that the system wants to be at. Let's have a look at how negative feedback processes work:

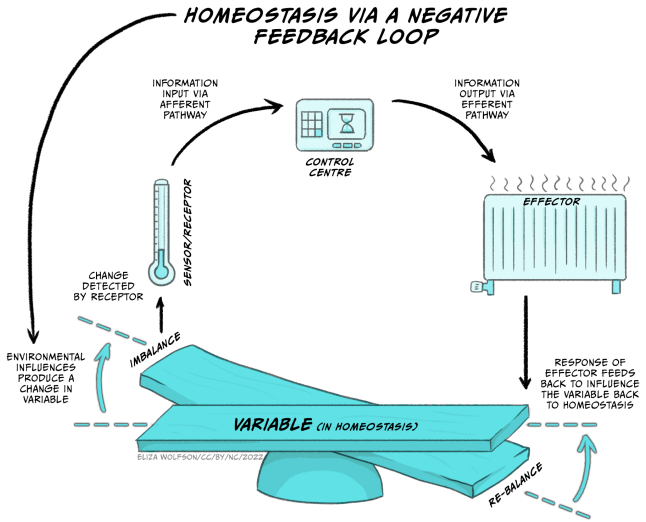

Fig 5.22. Homeostatic systems rely on negative feedback processes to maintain equilibrium.

Figure 5.22 demonstrates how negative feedback loops work to maintain homeostasis.

At the bottom of the image you can see the physiological variable that must remain within narrow limits of the set point so that the balance does not tip to either one side. If we use the example of body temperature, the set point is within 36.5°C and 37.5°C, because that is normal body temperature for humans, at which our physiological processes work optimally. The system also consists of sensors or receptors. These measure what the **actual** body temperature is. Information about actual body temperature is typically sent to a control system that can monitor deviations from the set point. If there are deviations and the balance tips one way or the other, the control system will send this information to the effector part of the system so that correctional behaviours or physiological responses can be initiated in order to restore body temperature within the narrow margins of normal body temperature. For the control of temperature these responses may involve behaviours such as putting on or removing clothing, and physiological responses such as sweating and vasodilation of peripheral blood vessels to cool down or peripheral vasoconstriction and shivering to increase core body temperature.

This is also how the thermostat in our homes works. If the ideal/target temperature on the thermostat has been set to 21°C, and it happens to be a cold windy day, the thermostat

will display the target temperature (i.e. 21°C) and an actual temperature (e.g. 18°C). Since the actual temperature on the sensor/thermostat deviates from the target, the boiler (effector) will start working, turn the radiators on, and restore the home to 21°C, at which point the boiler will switch off. This is conceptually how we think of physiological homeostatic systems in our body that use negative feedback mechanisms to restore and maintain balance to the system.

Thus, a homeostatic system, or a physiological system that depends on homeostasis, requires a **system variable** that is controlled by the system (e.g., temperature, hydration, nutrients), which must remain within narrow bounds of a **set-point** for the system to work well (e.g. 36.5-37.5°C for body temperature). **Sensors (receptors)** measure the actual value of the system variable, and transmit this information to a **control centre**, which can detect deviations from the set point. If deviations are detected, the control centre transmits this information to the **effector system** which initiates the necessary behavioural/physiological processes to change the system variable and restore homeostasis.

Motivation to eat/stop-eating to maintain homeostasis

Box 2: How does the body use energy, and how does it extract energy from food?

The body uses energy for three primary reasons.

The largest amount of energy that we take in through our food is used to maintain **basal metabolism rates** (BMR). Thus 55% of energy usage is to maintain body heat and other basic bodily functions (e.g. breathing, blood circulation). The proportion of energy used to maintain BMR varies as a function of body size. Elephants, for example, consume more energy to maintain basic functions than mice. Of this 55%, the liver uses 27% and the brain uses 19% (this includes the energy for neuronal signalling as well as basic housekeeping processes).

The **digestion** of food and the processes involved in

extracting nutrients from food use 33% of the energy that comes in through food.

Finally, 12-13% of the energy that we take in is used as energy for **active behaviour**, and this percentage varies depending on the level of exercise/activity that we do. If we go to the gym, for example, we will use more than the 13%. Since only a fraction of the energy that we consume is utilised for active behaviour, while exercise is a good way to lose weight, reductions in intake are usually also necessary for weight loss. Energy which is not used for BMR maintenance, digestion, or activity, will be stored as energy reserves either in the liver (short term storage) or in fatty tissue (long term storage).

Glucose is the primary fuel or form of energy that the body uses. Glucose is derived from three main sources in our diet:

- carbohydrates (sugars),
- amino acids (building blocks of proteins), and
- lipids (fat).

Carbohydrates are broken down and converted into glucose as soon as they are taken in by the body. Glucose in turn is used as the main energy source to fuel the brain, muscles and the rest of the body.

Excess glucose is stored in the form of **glycogen** in the liver. This is a short term storage of energy that we can use when needed through a process involving the pancreas. On detecting an increase in blood glucose, the pancreas releases insulin, which converts excess glucose into glycogen that is then stored in the liver for short term storage. If we need this energy, the pancreas secretes glucagon to convert the glycogen back to glucose so that it is then used by the body and the brain. Carbohydrates are not the only possible source of energy because proteins and fats can also be broken down to form glucose. This is the basis for many low carb diets, whereby by reducing the intake of carbohydrates, the individual will need to get their glucose from amino acids, and especially fats.

Amino acids derive from proteins, and provide the basic building blocks that cells use to make new proteins to perform the different specialised or general jobs within a given cell. However amino acids are also a source of glucose, as they can be converted to **glycerol** which in turn can be converted to glucose. Out of the twenty different amino acids, nine are essential, i.e. we cannot produce them in our bodies and need to take them in through our diet. For example, **tryptophan** is an

essential amino acid and is found in oats, bananas, dried prunes, milk, tuna fish, cheese, bread, chicken, turkey, peanuts, and chocolate. It is the sole precursor of the neurotransmitter **serotonin**. The ability to change the rates of serotonin synthesis through the manipulation of levels of tryptophan in the body is the foundation of a large body of research examining the relationship between serotonin dysregulation and mood, behaviour, and cognition (Richard et al., 2009 for review).

Finally, lipids or fats can also be converted to glucose, but also constitute essential building blocks for our cells, forming the lipid bilayer that forms the cell membrane. Glucose can also be stored long term in fatty tissue, or adipose tissue in our body. Fats are stored in fatty or adipose tissue, or converted either into fatty acids or glycerol. Glycerol can in turn be converted into glucose for energy.

Carbohydrates are non-essential but amino acids and lipids are essential from a building block perspective, as are minerals and vitamins. Minerals and vitamins must also be taken in through our diets or via supplements; they are essential for normal body functioning, but they are not a source of energy.

If the motivation to eat or stop eating results from a need to alleviate a negative state of hunger or of feeling full respectively, it begs the questions:

- What is the system variable that needs to remain in homeostasis?
- Which sensors or receptors measure the variable?
- Is there an effector mechanism that either changes metabolic processes or that initiates or terminates the feeding behaviour so that equilibrium is restored in the system?
- If so, is the effector mechanism located in a particular part of the body, or the brain?

Since glucose is the main source of energy in the body, it would make sense that we should have a homeostatic system that regulates the amount of glucose in the body.

The notion that glucose metabolism plays a key role in the control of hunger, satiety and the regulation of body energy balance, was first proposed by Anton Julius Carlson (1916), but was later formalised into the glucostatic theory of food control by Jean Mayer (1954;1955). According to this theory, the system variable that should be maintained within narrow limits is the level of glucose concentration in the blood. Campfield and Smith (2003) recorded blood glucose concentration changes in rats over time, and found that a fall in blood glucose was correlated with meal initiation. Thus,

when blood glucose concentrations decreased, the animal would begin feeding, which would result in the rise of blood glucose concentrations.

While the **glucostatic** theory of food control proposed that short-term appetite control or starting/stopping eating is mediated by deviations from a hypothetical blood glucose level set point, other proposals included glucose concentrations in the brain as being the key set point. In terms of long term regulation of weight, which is different from a glucose-mediated short-term control of appetite, the **lipostatic** theory suggested that in the long term the body is trying to maintain an optimum body fat level. These theories are not mutually exclusive, as they deal with short and long term appetite control, and might be complementary: It may be that our body is regulating multiple variables in a homeostatic way.

If deviations from optimum blood or brain concentrations of glucose elicit regulatory motivational drives to eat or to stop eating, what part of the brain or body constitutes the effector mechanism?

According to the dual centre model (Stellar, 1954), two areas in the hypothalamus, the lateral hypothalamus and the ventromedial hypothalamus, were thought to be the dedicated hunger and satiety centres, or start and stop eating centres. The lateral hypothalamus is a group of cells in the hypothalamus that are located away from the midline of the brain, while the ventromedial hypothalamus is a group of cells that are near the

midline (medial) and towards the bottom (ventral) part of the hypothalamus.

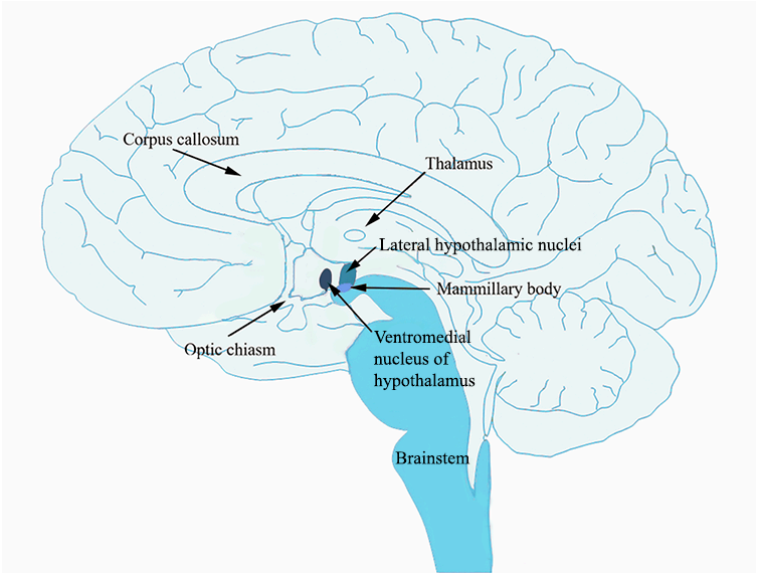

Fig 5.23. Lateral view of the human brain showing the ventromedial and lateral hypothalamus in humans

The model was based on findings from lesion studies. Bilateral lesions of the ventromedial hypothalamus resulted in the animal starting to eat and put on weight (Hetherington and Ranson, 1942; Brobeck, Tepperman and Long, 1943). Thus, it was reasoned that if removal of this area results in initiating feeding behaviour, then this area must be responsible for stopping feeding. Conversely, bilateral lesions of the lateral hypothalamus resulted in the animal eating less and losing

weight compared to control animals without the lesion (Hetherington and Ranson, 1940; Anand and Brobeck, 1951). Thus, if removal of the lateral hypothalamus results in less feeding, then this area must be responsible for starting to eat. More recent experiments using optogenetics to stimulate the lateral hypothalamus have shown that animals initiate eating upon stimulation of the lateral hypothalamus (Urstadt et al., 2020). The following video shows you this effect (https://www.youtube.com/watch?v=lBhYmBkqj4o):

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://openpress.sussex.ac.uk/introductiontobiologicalpsychology/?p=361#oembed-1

Thus the motivation to eat/stop eating from a homeostatic perspective could be governed by these effector mechanisms located in the lateral and ventromedial hypothalamus, with the lateral hypothalamus being responsible for initiating the processes that make the animal start to eat when the animal is hungry, and the ventromedial hypothalamus being responsible for the processes that make the animal stop eating when the animal is satiated. In support of this hypothesis, research has identified receptors in the lateral hypothalamus and the liver

(i.e. sensors) that measure levels of glucose so could use changes in glucose level to drive feeding.

However, further research suggested that the dual centre model may not reflect the full picture.

Research conducted by James Olds and Elliott Valenstein contradicted the idea that the lateral and ventromedial nuclei of the hypothalamus are **dedicated** starting and stopping eating centres. They placed animals in operant conditioning chambers and attached an electrode to their lateral hypothalamus. A lever was placed on one side of the chamber. When the animal accidentally pressed the lever, they received electric stimulation to their lateral hypothalamus. Thus, pressing the lever would result in the animal self-stimulating their lateral hypothalamus (a method known as 'self-stimulation reward'). They observed that the animals would readily self-stimulate the lateral hypothalamus, and often to exhaustion. In some of their experimental set ups, the animals would run across a chamber where mild electric shocks were given in order to reach the lever that would allow it to self-stimulate the lateral hypothalamus. If the lateral hypothalamus is the hunger or start eating centre, why would the animals repeatedly press for stimulation that produces a hunger-like state?

In follow-up experiments, Elliot Valenstein changed the design of the studies so that only the researcher was able to administer the stimulation. They observed that similar to the optogenetics experiment mentioned above, the animal would

eat upon stimulation when food was available. However, when water was available then the animal would drink. If there was an intruder in the chamber (e.g. another male rat), the animal would fight, and if there was a receptive female in the chamber the animal would mount the female. This suggested that the effects of lateral hypothalamic stimulation depend on the situation, and that therefore the lateral hypothalamus is not a dedicated hunger center, but more generally involved in motivated behaviours (including feeding).

The idea that the lateral and ventromedial hypothalamic nuclei are involved in hunger and satiety has not been rejected, but where research has moved is that maybe there are not dedicated parts of the brain but maybe there are dedicated receptors, or dedicated hormones that act on receptors. Maybe there are dedicated hunger or satiety hormones that play the role of the effector mechanism and lead to feeding or stopping feeding?

Thus, the idea of dedicated locations in the brain for hunger and satiety was revisited, following the discovery of various peptide hormones that are released predominantly in the periphery, by the gut, intestines or adipose tissue, that seemed to signal hunger or satiety.

Two peptide hormones, ghrelin and orexin, secreted from the gut (ghrelin), and from adipose tissue as well as from within the hypothalamus (orexin) not only stimulate food intake, but are also involved in wider motivational and also body clock regulatory processes.

The second set of hormones that were discovered were cholecystokinin (CCK) and Peptide YY. CCK is released from the intestines in response to the intake of fat. If hungry rats are administered CCK, feeding is inhibited. Peptide YY is released in the gut (stomach and intestines), and similarly, injections of PYY inhibit eating in hungry animals. Furthermore, there is some evidence that PYY may be abnormally low in individuals who are obese, suggesting these individuals may be less able to inhibit eating due to lack of PYY.

Leptin was discovered by researchers in Jeffrey Friedman's lab in 1994 (Zhang et. al., 1994). The discovery of leptin was preceded by the accidental discovery of a genetic strain of mice (ob^{-}/ob^{-} mice) which grew to become obese, had decreased rates of basic metabolism and low physical activity. It was later concluded that their genetic mutation resulted in reduced circulation of leptin (see Figure 5.23 below). We now know that leptin is produced and released from adipose tissue (fatty tissue) and we know that it acts on several different receptors, some of which are located within the ventromedial hypothalamus to signal stopping eating. Lack of these receptors in ob^{-}/ob^{-} mice therefore results in overeating. Cases of congenital human leptin deficiency however are extremely rare, and while some clinical work in humans has shown that delivery of leptin in obese individuals allows them to lose weight, the clinical picture is more complicated as there is also evidence of leptin resistance (leptin doesn't work well

enough), as most obese individuals have plenty of leptin but do not respond to leptin by stopping eating.

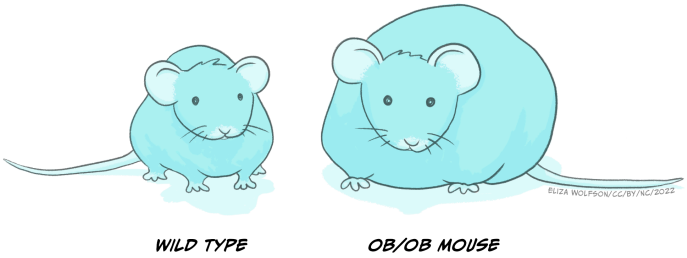

Fig 5.24. Genetic obese mouse (right), compared with normal control (left)

Non-regulatory motivated behaviours: motivation not for homeostasis

We know that motivation to eat or not does not only result from the need to maintain homeostasis of nutrients in the body.

We can stimulate eating through tastes and smells even in

animals that are full, and we can stimulate eating and stopping eating through learned associations. Motivation in these cases does not depend on homeostatic mechanisms. Thus, conditioned motivational drives can cause changes in appetite.

Prior to three months, babies feed to maintain homeostasis: they take large breastfeeds first thing in the morning to relieve hunger when they wake up. However, at around 3-6 months, they switch to a large feed last thing at night. This large meal anticipates the relative difficulty of obtaining night-feeds. So this is not to relieve hunger, but in anticipation of possible hunger.

This anticipatory eating behaviour has also been observed in rats (Strubbe and Woods, 2004). Rats are nocturnal animals: they eat and drink when it's dark, and sleep during day time.

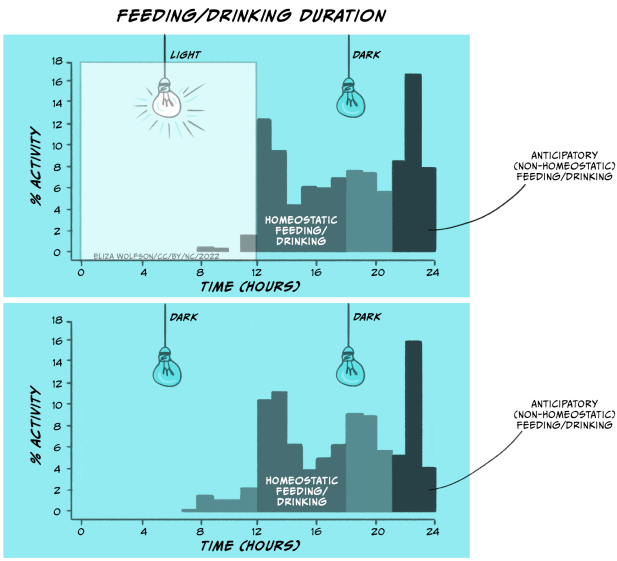

Fig 5.25. Homeostatic and anticipatory eating and drinking in rats

In Figure 5.25 (top panel), you can see the distribution of eating and drinking behaviours in rats when the lights in their chambers were turned on (when rats go to sleep) and when the lights in their chambers were off (when rats wake up). In the top panel, rats engage in homeostatic feeding and drinking as soon as the lights go off. Thus, rats will, similar to babies, increase their intake of food and drink as soon as the lights go off (when they wake up and start feeding/drinking). This is regulatory/homeostatic feeding and drinking. However, rats also increase their food and drink before the lights go on (when

they go to sleep). This is anticipatory feeding and drinking. This behaviour, however, is not directly related to whether the lights are on or off, but rather relates to their own internal body clock (which has been entrained over time by the light and dark cycles). You can see this in the lower panel when the lights are kept switched off. Eating and drinking increases in the same way even when the rats are always in the dark.

Motivation to eat/stop eating as a result of conditioned responses

Conditioning or learning can drive feeding even when the animal is full. This is known as cue-potentiated feeding, and it has been shown in rats and in humans.

Peter Holland and researchers in his lab taught hungry rats to associate a tone with the delivery of food. This association was achieved through simple pairing of the tone with the delivery of food, similar to the experiments carried out by Ivan Pavlov. (In the classical experiments, Pavlov presented a neutral stimulus – in the original experiments, a metronome rather than a bell – immediately prior to the delivery of food [the unconditioned stimulus, UCS] to a dog. It was found that following multiple pairings of the neutral stimulus and the food, the dog eventually displayed digestive responses [salivation and gastric secretions] in the presence of the stimulus alone [conditioned stimulus, CS], before the food

was delivered. Pavlovian conditioning is also briefly described in the Introduction, below Figure 1.8).

A second control cue (a different tone) was not paired with food. The researchers then allowed the rats to eat until they stopped, presumably because they were full. This ensured that the homeostatic drive to eat did not apply during the subsequent experiment. The rats were full and therefore were not motivated to eat to relieve a hunger state. The researchers then presented the tone that had been associated with the delivery of food or the tone that had been associated with no food. They found that when the rats heard the tone associated with the delivery of food they ate – this was termed cue-potentiated feeding. When rats heard the tone associated with no food they consumed less food. The researchers found that this mechanism depended on the amygdala and, more importantly, on the connection between the amygdala and the lateral hypothalamus. Severing the connection between the amygdala and the lateral hypothalamus stopped cue-potentiated feeding.

A similar experiment was undertaken with preschool students (Birch et. al., 1989). Over several training days, the researchers presented students with a rotating red light and music followed by the presentation of different snacks that the students preferred so that the students learned to associate certain light conditions and music with favourite snacks (peanut butter, hot dogs etc.). On the test day, students were allowed to eat as much as they wanted. Then the light and

music changed to the lighting conditions and music associated with the training sessions. The researchers found that not only did the students consume food again, even though they were full, but also that when the light and music were the same ones as when their favourite food was available, they began eating sooner than when the light and music presented in the cafeteria had not been paired with their favourite food. Thus humans, too, exhibit cue-potentiated feeding, eating depending on the environmental cues even in the absence of a homeostatic drive to eat.

Motivation to eat/stop-eating as a result of 'liking' vs. 'wanting'

Earlier in the chapter we described a series of experiments by James Olds and Elliott Valenstein which used the method of 'self-stimulation reward' in which rats readily pressed a lever to self-administer electric stimulation to their lateral hypothalamus. We also saw that in subsequent experiments, researcher-elicited stimulation resulted in the animals engaging in various motivated behaviours depending on the situation (eating, drinking etc.). These findings cast doubt on the prevalent idea that behaviour is motivated by drive reduction because if the drive reduction theory were true, stimulation in the same region that elicited hunger should have resulted in the animal experiencing the state of hunger, finding this state aversive and becoming motivated to behave in a way to reduce

this state of deprivation. Instead, the animals self-stimulated the same region of the brain. The researchers concluded that the rats were motivated to self-stimulate because they found the self-stimulation rewarding (Valenstein et al., 1970).

Subsequent work by psychobiologists Robert C. Bolles, Dalbir Bindra, and Frederick Toates in the 1970s and 1980s allowed psychologists to abandon 'drive reduction' views of motivation, and paved the way for the concept of 'incentive motivation'. Incentive motivation theories propose that behaviour is motivated by the prospect of an external reward or incentive. Thus incentive motivation is mediated by learning (consciously or unconsciously) about the availability of rewards in our environment. If a particular behaviour is expected to lead to a rewarding outcome, then we will be motivated to repeat this behaviour in order to obtain the goal of the reward (e.g. if pressing a lever will provide a sugary treat to a rat, the rat will increase the rate of pressing the lever, i.e. its motivation to execute the behaviour has increased in order to obtain the sugar reward).

Similarly, if a stimulus in the environment is expected to lead to a reward (e.g. a Pavlovian conditioned association where the sound of a bell predicts that food will be available), then motivation will increase for seeking out the stimulus that predicts the reward. Interestingly, in the Bindra-Toates model of motivation, physiological states were proposed to moderate incentive motivation, so that the value of the incentive/reward and thus also the value of the stimulus that may predict the

reward can change depending on the physiological state of the animal. Thus the motivation to take a hot bath on a hot day if we are feeling cold will be higher, and the hot bath will be perceived as more pleasant and rewarding than it usually would on a hot day.

The Bindra–Toates incentive motivation model additionally suggested that rewards and incentives are liked and wanted. In addition, the learned Pavlovian stimuli that predict them also become both 'liked' and 'wanted' as a consequence of the learned association with the reward. Liking and wanting were proposed to be synonymous in the Bindra-Toates model.

However, Terry Robinson and Kent Berridge, in their incentive salience model, proposed that the incentive motivational processes of '*liking*' and '*wanting*' should be considered separately because these two components of reward are mediated by different brain mechanisms.

In a series of influential experiments, they dissociated the processes of 'liking' a reward and 'wanting' (or working for) a reward. 'Liking' was linked to the hedonic pleasure that was associated with the reward (e.g. they observed that as in babies, rats will also lick their mouth upon receiving a sweet taste). 'Wanting' on the other hand, or what they termed 'incentive salience' is the motivational value of the reward or of a stimulus that may predict the reward, and while in some cases pleasure/hedonic impact/'liking' and 'wanting'/motivational value of the reward may coincide to motivate behaviour (e.g.

eating a cold ice cream on a hot day), in some cases they do not (e.g. eating a cold ice cream on a very cold day – here the liking will be the same, but eating a cold ice cream on a hot day to cool down might be wanted more).

Reward has long been associated with dopamine release in the mesocorticolimbic dopamine system which projects from the ventral tegmental area to the nucleus accumbens and to parts of the prefrontal cortex (see Figure 5.26).

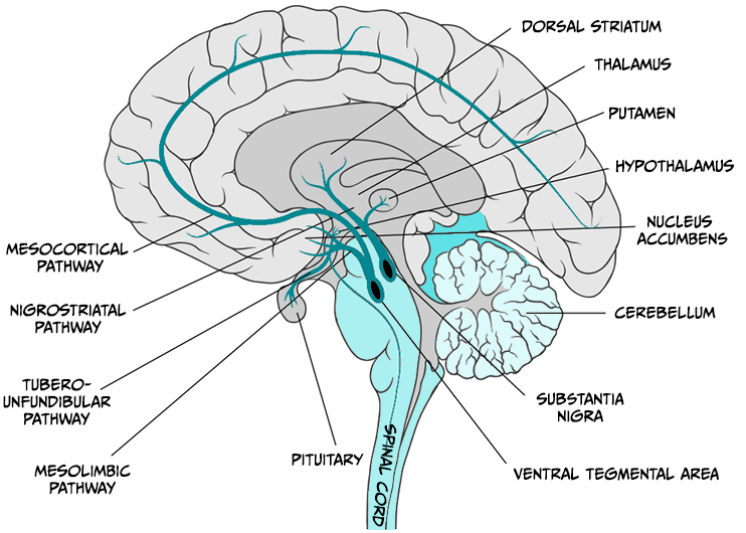

Fig 5.26. Dopaminergic pathways

As a result, Kent Berridge and Terry Robinson hypothesised that if dopamine in the nucleus accumbens were depleted (through selective lesioning of dopamine neurons), then rats would not seek out a reward (no 'wanting'). This is indeed what they found. Hungry rats that lacked dopamine became

aphagic and adipsic (they did not eat or drink). However if they were forced to eat something sweet, they did show licking responses associated with 'liking'. In follow-up experiments with genetically modified mice which had high levels of dopamine in their nucleus accumbens, the researchers found that the mice would work more than control, wild type mice to obtain sucrose, but 'liking' responses did not differ compared to mice without the mutation. These experiments suggested that the incentives of 'liking' and 'wanting' were indeed dissociable and that 'wanting' was mediated, at least in part, by dopamine release in the nucleus accumbens (see Berridge and Robinson, 2016 for review).

The influential work by Ann Kelley and her colleagues in the 1990s corroborated the idea that liking and wanting incentives are likely mediated by separate systems by showing that liking may be in part be mediated by opioid receptors in the nucleus accumbens, as opioid receptor stimulation of the nucleus accumbens resulted in the enhancement of intake not of food in general but specifically in the enhancement of intake of palatable sweet or high fat foods more than other foods (Kelley et al., 1996; Zhang et al., 1998; Zhang and Kelley, 1997).

The current understanding is that motivation due to 'liking' is mediated by opioid, GABA and cannabinoid neurotransmitter systems in the nucleus accumbens and that motivation due to 'wanting' is mediated by dopamine in the nucleus accumbens.

Limbic structures involved in non-regulatory motivation

Emotions also influence motivated behaviour. We are inclined to avoid fearful situations or environments and approach situations and environments that can make us feel happy.

The amygdala, a region within the limbic system of the brain, has long been associated with both emotion and motivation, ever since it was observed that amygdala lesions in monkeys resulted in the animals showing no behavioural responses to ordinarily-threatening stimuli but that they increased exploration of familiar stimuli (as if they were unfamiliar), elicited feeding towards inedible objects such as rocks and increased sexual behaviours towards inappropriate partners such as human experimenters. This behavioural pattern is termed Klüver-Bucy syndrome and can occur in humans with medial temporal lobe damage including the amygdala.

Research is still ongoing to delineate which precise regions of the amygdala are involved in motivated behaviours but the amygdala is thought to be involved in motivational processes that involve conditioned and learned associations between environmental cues and rewarding or aversive outcomes.

Key Takeaways

- Basic physiology of motivation involves homeostatic (negative feedback) mechanisms that maintain temperature, hydration, nutrient levels around set-point
- Hypothalamic areas critical in homeostatic regulation
- Non-homeostatic influences through learning, emotion etc. dependent on limbic systems.

References

Anand, B. K., & Brobeck, J. R. (1951). Hypothalamic control of food intake in rats and cats. *The Yale Journal of Biology and Medicine*, *24*(2), 123-140. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2599116/

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction.

American Psychologist, *71*(8), 670–679. https://doi.org/10.1037/amp0000059

Birch, L. L., McPhee, L., Sullivan, S., & Johnson, S. (1989). Conditioned meal initiation in young children. *Appetite*, *13*(2), 105–113. https://doi.org/10.1016/0195-6663(89)90108-6

Brobeck, J. R., Tepperman, J., & Long, C. N. H. (1943). Experimental hypothalamic hyperphagia in the albino rat. *The Yale Journal of Biology and Medicine*, *15*(6), 831. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2601393/

Campfield, L. A., & Smith, F. J. (2003). Blood glucose dynamics and control of meal initiation: A pattern detection and recognition theory. *Physiological Reviews*, *83*(1), 25–58. https://doi.org/10.1152/PHYSREV.00019.2002

Carlson, A. J. (1916). *The control of hunger in health and disease.* (1st ed.). Chicago: University of Chicago Press.

Hetherington, A. W., & Ranson, S. W. (1940). Hypothalamic lesions and adiposity in the rat. *The Anatomical Record*, *78*(2), 149–172. https://doi.org/10.1002/AR.1090780203

Hetherington, A. W., & Ranson, S. W. (1942). The spontaneous activity and food intake of rats with hypothalamic lesions. *136*(4), 609–617. https://doi.org/10.1152/AJPLEGACY.1942.136.4.609

Hull, C. L. (1943). *Principles of behavior, an introduction to behavior theory.* Appleton-Century-Crofts.

Kelley, A. E., Bless, E. P., & Swanson, C. J. (1996). Investigation of the effects of opiate antagonists infused into the nucleus accumbens on feeding and sucrose drinking in rats. *Journal of Pharmacology and Experimental Therapeutics*, *278*(3), 1499–1507.

Mayer, J. (1953). Glucostatic mechanism of regulation of food intake. *The New England Journal of Medicine*, *249*(1), 13–16. https://doi.org/10.1056/NEJM195307022490104

Mayer, J. (1955). Regulation of energy intake and the body weight: the glucostatic theory and the lipostatic hypothesis. *Annals of the New York Academy of Sciences*, *63*(1), 15–43. https://doi.org/10.1111/J.1749-6632.1955.TB36543.X

Richard, D. M., Dawes, M. A., Mathias, C. W., Acheson, A., Hill-Kapturczak, N., & Dougherty, D. M. (2009). L-Tryptophan: Basic metabolic functions, behavioral research and therapeutic indications. *International Journal of Trytophan Research*, *2*(1), 45. https://doi.org/10.4137/IJTR.S2129

Skinner, B. F. (1974). *About behaviorism* (1st ed.). New York: Knopf.

Stellar, E. (1954). The physiology of motivation. *Psychological Review*, *61*(1), 5–22. https://doi.org/10.1037/H0060347

Strubbe, J. H., & Woods, S. C. (2004). The timing of meals. *Psychological Review*, *111*(1), 128–141. https://doi.org/10.1037/0033-295X.111.1.128

Urstadt, K. R., & Berridge, K. C. (2020). Optogenetic mapping of feeding and self-stimulation within the lateral

hypothalamus of the rat. *PLOS ONE*, *15*(1), e0224301. https://doi.org/10.1371/journal.pone.0224301

Valenstein, E. S., Cox, V. C., & Kakolewski, J. W. (1970). Reexamination of the role of the hypothalamus in motivation. *Psychological Review*, *77*(1), 16–31. https://doi.org/10.1037/H0028581

Zhang, M., Gosnell, B. A., & Kelley, A. E. (1998). Intake of high-fat food is selectively enhanced by Mu opioid receptor stimulation within the nucleus accumbens. *Journal of Pharmacology and Experimental Therapeutics*, *285*(2), 908–914.

Zhang, M., & Kelley, A. E. (1997). Opiate agonists microinjected into the nucleus accumbens enhance sucrose drinking in rats. *Psychopharmacology*, *132*(4), 350–360. https://doi.org/10.1007/S002130050355

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L., & Friedman, J. M. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature*, *372*(6505), 425–432. https://doi.org/10.1038/372425a0

About the authors

Dr Kyriaki Nikolaou
UNIVERSITY OF SUSSEX

Dr Nikolaou completed her PhD at Goldsmiths University of London before completing postdoctoral work in the School of Psychology at the University of Sussex, the Department of Developmental Psychology at the University of Amsterdam, and the Institute of Psychiatry at Kings University of London. She moved to the University of Sussex where she is currently a lecturer in Psychology.

Her work has focused on understanding the acute effects of various drugs of abuse on executive and cognitive functioning, as well as on how drug-related cues are processed in the brain and elicit biased behavioural and cognitive responses.

Professor Hans Crombag
UNIVERSITY OF SUSSEX

Professor Hans Crombag is an internationally recognised expert in behavioural and neurosciences with a PhD in Biological Psychology. His research has primarily focused on

mental health, biological/environmental interactions, and substance abuse.

Since 2007, he has been employed by the University of Sussex, developing and overseeing innovative and interdisciplinary scientific research programmes, spanning and integrating multiple health-related and public/social justice fields. He has influenced thinking around neurolaw, justice and public policy as Co-Director of the Sussex Crime Research Centre.

Previously, at the Department of Psychological & Brain Sciences at the John Hopkins University, he worked on research in the areas of neurogenetics of eating/ eating disorders (and obesity) and substance abuse/addiction. He is a member of the Society for Neuroscience, European Behavioural Pharmacology Society and International Neuroethics Society.

He has a long-standing interest in mental health and wellbeing and public health policy.

PART VI
DYSFUNCTION OF THE NERVOUS SYSTEM

Studying the function of the nervous system enables us to advance our understanding of basic biology and behavioural processes. Such understanding can help gain insight when things go wrong and biological function/behaviour becomes disrupted.

In this section you will learn about the biological basis and psychological presentation of major psychiatric and neurological disorders including addiction, schizophrenia, affective disorders and the main causes of dementia. We will also consider how the ageing process and placebos are able to modulate our biological and psychological functioning.

15.

ADDICTION

Dr Andrew Young

Learning Objectives

- Appreciate the features of reward and reinforcement
- Understand how the effects of drugs on reinforced behaviours points to a critical role of dopamine
- Understand how abused drugs affect mesolimbic dopamine transmission
- Appreciate the link between drug self-administration and drug dependence (addiction).

What is addiction?

Drug addiction, or to give it its more scientific term, **dependence**, is the taking of a chemical substance (the drug) for non-nutritional and non-medical reasons, where the drug-taking behaviour is compulsive. An addict feels they have no control over taking the drug, but instead feels driven to take it. Their lives often become centred around acquiring and consuming the drug, to the detriment of behaviours necessary for survival – for example, eating, drinking water – and they often engage in risky or illegal behaviour in order to feed their drug habit. Addicts often develop a tolerance to the drug, such that they need more of the drug to produce the 'high'. Drug dependence must be distinguished from drug use and drug abuse. Drug use is where the substance is taken in small quantities, relatively infrequently, and importantly with no damage to relationships or daily function. For example, people often enjoy a glass of wine with a meal, or a drink with friends. If drug use escalates to frequent and/or excessive taking of the substance, causing disruption to daily functioning or relationships, but without the compulsivity, this would be termed drug abuse. Drug dependence, as stated above, is similar to drug abuse, except that the drug-taking is compulsive, with the addict feeling they have no control over whether to take the drug.

There are four major stages of drug addiction: initiation, maintenance, abstinence and relapse, each of which are likely

to be driven by different mechanisms. **Initiation** is the first stage, where a person takes the drug for the first time. The main factors which influence initiation are: the 'pleasant' feeling (hedonic impact) from taking the drug; overcoming stress; peer pressure and the desire to conform to a group; or simply to experiment. For many people, drug taking never progresses beyond this stage, and whether or not they take a drug is entirely a conscious decision.

However, where a person becomes dependent, or addicted, this moves on to the second stage: **maintenance**. Here the person no longer feels in control of the decision as to whether to take a drug, but rather feels a compulsion to take it. The maintenance stage can be long-lasting, and is very often accompanied by an increasing motivational drive to take the drug, which is driven by a process called sensitisation (see 'Tolerance and sensitisation' box below). However, there is rarely an accompanying increase in hedonic impact from taking the drug: indeed very often hedonic impact decreases, and the drugs may even become aversive.

Tolerance and sensitisation

These are forms of neuroadaptation which are important in many aspects of neuronal function, and are particularly important in understanding processes of drug addiction.

Tolerance refers to the process where a drug becomes less effective, that is, it produces a weaker response after repeated administration.
Sensitisation, on the other hand, is where the drug becomes more effective over repeated administration.

Both processes are mediated through changes in cellular function, including:

- changes in neurotransmitter synthesis, storage and release,
- changes in receptor density,
- changes in reuptake and metabolism of transmitters,
- changes in second messenger signalling,

many of which probably involve upregulation or downregulation of specific gene expression.

However, at present we do not fully understand how these mechanisms are controlled. Interestingly, the processes involved in sensitisation are very long-lasting, and some have even suggested that they may be irreversible, accounting for the enduring changes that underlie maintenance of addiction. A link with mechanisms of learning has also been suggested by the observation that antagonists at NMDA-type glutamate receptors, given into VTA, prevent sensitisation. NMDA-receptors are known to be critically involved in neuroplasticity mechanisms of learning, and this evidence suggests that similar processes involving NMDA receptors may underlie drug-induced sensitisation.

Once a person has become dependent, it is likely that the addiction remains with them for the rest of their life: there is little evidence for true recovery. Therefore when an addict refrains from taking a drug, they are not normally considered to be 'cured' or 'recovered', but rather they are considered to be **abstinent**. This reflects the view that the addiction is still present, but is not expressed because the person no longer

takes the drug. However, the motivational drive to take the drug – that is, the craving – may still be strong. This is underlined by evidence showing that physiological and neurochemical changes occurring in the brain with the development of dependence are largely irreversible, as we will see later. Therefore, abstinent addicts are very prone to restarting their drug taking, termed **relapse**. A single intake of the drug can reinstate the maintenance phase in an addict who may have been abstinent for many years, hence the requirement, in treatment programs for dependence, that the addict never take the drug. Relapse is driven by cravings in the individual, which may be brought about by stress or by exposure to people, items or situations associated to previous drug taking, emphasising the link between classical conditioning and drug taking.

When an addicted person stops taking a drug, they often experience withdrawal symptoms. These are behavioural changes, often opposite to the effects elicited by the drug, and can be very aversive. In the early stages of abstinence these withdrawal symptoms are particularly strong and can be extremely unpleasant. Thus, avoiding withdrawal symptoms provides a strong motivation to take the drug and can lead to relapse. However, as the period of abstinence increases the withdrawal symptoms subside, so reducing this as a motivation for reinstatement.

Brain motivation circuits and

addiction

Many studies have been undertaken in experimental animals, particularly rats and mice, but also primates, to investigate the neural circuitry underlying addiction. These mostly focus on pathways controlling reinforcement and motivation often termed the reward pathway.

Reward or reinforcement?

Reward is a term widely used in the discussion of dopamine signalling in the mesolimbic pathway. Indeed, many publications refer to the mesolimbic pathway as the 'reward pathway'. However, the term 'reward' has several problems in the scientific context. Reward is a pleasurable experience, and is subjective: what is pleasurable for one person may not be for another. It also raises problems of assessing pleasure in experimental animals: how do we know that an animal is enjoying an experience? Third, the term does not necessarily imply that it will change behaviour.

In scientific terms, for empirical research, we need to avoid subjective measures: far better to have objective measures. The term **reinforcement** refers to the ability of a stimulus, situation, or outcome, to elicit a behaviour. Reinforcement strengthens an animal's future behaviour on exposure to the stimulus. It is an objective measure: we can simply measure changes in the behaviour, for example the number of operant lever presses. Importantly, also, reinforcement does not imply pleasure, so in measuring the behaviour, we don't have to worry about whether the animal is enjoying the experience.

The discovery by Olds and Milner (1956) that rats would work by pressing a lever to receive mild electrical stimulation to a specific area of the brain, fuelled major research programs looking at this and related pathways in controlling motivation. They saw that the rats were motivated to stimulate areas of the mesolimbic pathway electrically: this pathway projects from cell bodies in the ventral tegmental area (VTA) along the axons of the mesolimbic pathway, to terminals located primarily in the nucleus accumbens.

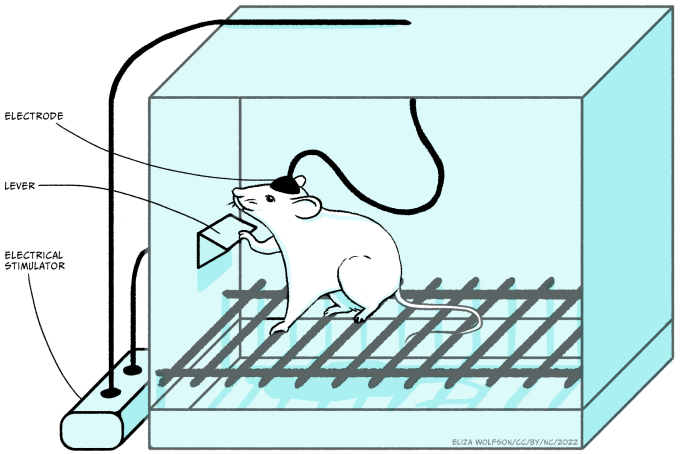

Fig 6.1. A rat pressing a lever to receive mild electric stimulation

Subsequent experiments have characterised the mechanism promoting the level pressing response in greater detail, but, importantly, they have emphasised the critical role of **dopamine** in driving the behaviour effect. Thus, amphetamine or cocaine, which enhance dopamine signalling, increase the lever press rate, whereas giving a dopamine antagonist reduces the lever press rate, indicating that the reinforcement signal driving the lever pressing behaviour is mediated through dopamine. It is important to note that the anatomical location of brain regions which support self-

stimulation is very specific: if the electrodes are located outside these localised regions, animals will not self-stimulate. An interesting point here, in relation to the phenomena of addiction, is that in self-stimulation experiments such as these, animals will repeatedly press the lever to receive the stimulation rather than eating or drinking, indicating that the electrical stimulation is a very strong motivational drive which suppresses the drive to carry out behaviours critical for survival: addicts often neglect normal nutrition and self-care in order to maintain their drug-taking.

In a similar procedure, rats and mice will also press a lever in order to receive injections of certain drugs. In its simplest form the drugs are administered intravenously (i.v.), through an indwelling cannula in a blood vessel: therefore lever pressing gives an i.v. injection of the drug. In a modification of the design, drugs can be administered via a microinjection into local brain areas. There are a number of drugs which animals will administer intravenously, including amphetamine, cocaine, nicotine, morphine, heroin and ethanol, and they will also administer amphetamine or cocaine into the nucleus accumbens and morphine into the VTA. It should be emphasised that animals will only self-administer certain drugs: the vast majority of drugs do not support self-administration. Similarly, the brain regions where animals will self-administer the drugs, VTA and nucleus accumbens, are very specific and indicates the importance of the mesolimbic pathway.

The importance of the mesolimbic dopamine system can be confirmed with lesion experiments, using 6-hydroxydopamine (6-OHDA), a drug which specifically kills catecholamine (dopamine and noradrenaline) containing cells. Following lesions to the mesolimbic pathway, but not to other pathways, animals will no longer self-administer drugs. Furthermore, enhancing dopamine function by giving local injections of cocaine or amphetamine into nucleus accumbens also increases the lever pressing, supporting the role of dopamine in the lever-pressing response. Paradoxically, many experiments have shown that dopamine antagonists also increase the lever press rate. However, it was subsequently found that the dose was critical: at low doses the lever-press rate is increased, but at higher doses it is entirely abolished. This dose-dependence can be explained by considering that at low antagonist doses not all receptors are occupied and therefore increasing the amount of drug self-administered will overcome the effect of the antagonist, whereas at high antagonist doses, the receptors are completely blocked, and so no matter how much more drug is administered the effect of the antagonist cannot be overcome. So it is concluded that the motivational effects of these drugs is mediated via dopamine neurones in the mesolimbic pathway.

You will recall that stress is a major contributory factor to development of dependence and relapse in abstinent addicts, and the influence of stress can be seen in animals trained to self-administer. If rats that had previously been trained to

lever-press to receive self-administration of cocaine are left drug-free for several weeks, they no longer press the lever when cocaine is once more available, modelling abstinence. If they then receive a foot shock, they do start pressing the lever again, reinstating the compulsive self-administration and paralleling the effect of stress on relapse in people. In the context of the role of stress in reinstatement, it is notable that this reinstatement is prevented by corticotropin-releasing hormone antagonists, emphasising the role of the hypothalamus-pituitary-adrenal axis mediated stress response to the process.

Measuring rodents' operant behaviour, such as lever pressing, is a good way of measuring their level of motivation, and as we saw from the experiments above, they show strong motivation to receive stimuli which activate the dopaminergic mesolimbic pathway. However these are clearly very artificial behaviours: they do not mimic activities which the animals undertake naturally. But we just need to look in the wild at how much effort animals will put in to getting food, be it a predator chasing down a prey, or annual migrations to new feeding grounds. Animals have an innate motivation to pursue behaviours which are beneficial to survival, for example eating, drinking and reproducing. As such, motivational systems in the brain are highly evolved to reinforce behaviours which enhance animals' ability to perform these actions. Stimuli associated with these action become strong predictors of outcome, and strong motivational cues to perform behaviours leading to consumption.

In the laboratory, too, motivation to pursue beneficial behaviours can be demonstrated: initial observations by B.F. Skinner in the 1940s, opened the way for many subsequent operant experiments where rats or mice pressed a lever in order to receive food or water or even a sexually receptive mate. Moreover, as with self-stimulation and self-administration, lesion and pharmacology experiments have shown the importance of dopamine in the mesolimbic pathway for controlling this behaviour. Thus, lever pressing to receive a natural reward (food, water) is abolished in animals with 6-OHDA lesions of the mesolimbic pathway, or by the application of dopamine receptor antagonists, and is enhanced by the administration of amphetamine or cocaine. So the mesolimbic pathway is clearly involved in motivation to undertake behaviours vital for survival, and self-stimulation and self-administration tap into this mechanism by promoting activity in the pathway either electrophysiologically or pharmacologically.

Direct neurochemical measurement in localised brain areas, primarily using brain microdialysis or fast-scan cyclic voltammetry (FSCV) (see 'Measuring neurotransmitter release in the brain' box, below) have shown that dopamine release in nucleus accumbens, but not in other dopaminergic terminal regions in the brain, is increased during appetitive behaviours. These behaviours include eating and drinking, and during electrical stimulation of the VTA, similar to that used in self-stimulation, therefore largely confirming the importance of

dopamine in motivation processes. Importantly, also, drugs which support self-administration also increase dopamine release in nucleus accumbens preferentially over other regions. The mechanisms by which the different drugs increase mesolimbic dopamine function varies across the different drug types. Some, such as nicotine, morphine, heroin and alcohol activate the pathway by either direct or indirect actions on the dendrites and cell body in the VTA, while others, such as amphetamine and cocaine affect the reuptake of released dopamine in the terminal regions, including nucleus accumbens.

Considering the site of action of the different drugs accounts for why animals will self-administer morphine into the VTA and amphetamine and cocaine into the nucleus accumbens, as these are the regions where the respective drugs activate mesolimbic function. Therefore, although addictive drugs exhibit very different primary pharmacology, with only amphetamine and cocaine acting directly on the dopamine system, and also have very different primary behavioural effects (Table 1), they all share the ability to increase dopamine function selectively in the mesolimbic pathway and it is this action that is believed to underlie their motivational effects. The drugs 'hijack' the neural pathway in the brain which controls the animals' motivation to pursue behaviours essential for survival, and instead motivate the individual to perform behaviours related to taking the drugs.

Measuring neurotransmitter release in the brain

Measurement of neurotransmitter release in localised brain areas during behaviour and/or in response to drugs is really important in understanding underlying neurotransmitter actions. Over the last few decades, two main methods have been employed, both of which can be used in awake, freely moving experimental animals.

Brain **microdialysis** involves implanting a small length of dialysis membrane into the brain, and perfusing it continuously with artificial cerebrospinal fluid (aCSF). Dissolved substances in the brain extracellular fluid pass through the membrane, by dialysis, into the aCSF, and can be measured typically by high performance liquid chromatography (HPLC). Second, **fast-scan cyclic voltammetry (FSCV)** measures the oxidation of chemicals when a voltage is applied to a carbon fibre microelectrode. Although it is possible to measure some other neuroactive compounds, the most widespread use of FSCV is to

measure dopamine. Microdialysis has the advantage that many different compounds can be measured in a single sample, whereas FSCV is mainly restricted to a single compound, normally dopamine. However, microdialysis probes are comparatively large (typically 1 to 2 mm long, 0,5 mm diameter) and so have a relatively poor spatial resolution. FSCV, on the other hand, uses carbon fibre microelectrodes which are much smaller (typically 100 µm long, 10 µm diameter) which give a much higher spatial resolution and allow targeting of smaller brain sub-regions. Microdialysis also has relatively poor temporal resolution, as it requires collection of enough sample to be able to analyse: most studies use sample collection times of 1 to 10 minutes, although some have managed less than a minute. In contrast, FSCV typically makes 10 measurements per second. Therefore FSCV is able to pick up fast transient changes in response to specific stimuli, whereas microdialysis can only pick up slower more sustained changes.

Recent developments with genetic markers are opening the way to novel approaches to measuring many aspects of neurotransmitter function with high chemical specificity, spatial resolution and temporal resolution.

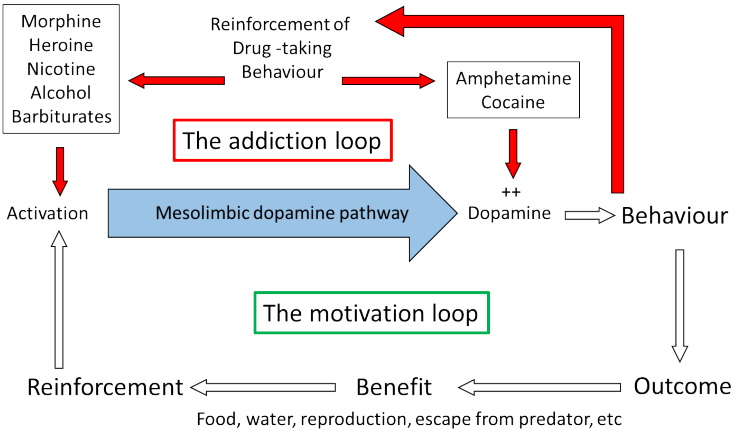

Fig 6.2. Diagrammatic representation of the central role of the mesolimbic pathway in motivation in natural situations beneficial for survival (the motivation loop) and how addictive drugs interact with this system (the addiction loop).

Conditioned place preference tests an animal's preference for an environment which is associated with a reinforcer, and can be assessed using a two compartment testing box. Animals are trained over repeated sessions that one compartment contains a reinforcer (typically food, sucrose or water), whereas the other compartment does not. After several training trials, the place preference is tested in the absence of any reinforcer (that is, both compartments are empty). The animal is placed back in the testing box, and the time spent in each compartment is recorded. Animals spend more time in the previously

reinforced compartment than in the control compartment, even though at test there is no reinforcer present.

This shows that the animal has learned which compartment of the test box contained the reinforcer, and it is motivated to visit that compartment in preference to the control compartment, even when the reinforcer is no longer present. This effect is:

1. abolished by 6-OHDA lesions of the mesolimbic pathway;
2. enhanced by drugs which increase dopamine, such as amphetamine and cocaine; and
3. attenuated by dopamine receptor antagonists.

Therefore, the mesolimbic dopamine pathway is critical for expression of conditioned place preference.

In a variant of the procedure, instead of a natural reinforcer, drugs can be used. In this case, during training, the animal is given an injection of a drug and placed in one compartment or a saline injection and placed in the other compartment. At test, with no drug present, they show a preference for the compartment in which they have previously received the drug. The drugs which will induce place preference, including amphetamine, cocaine, morphine, heroin and alcohol, are the same ones which animals will self-administer, and all are potentially addictive drugs in people. One important aspect that the place preference experiments demonstrates is that

animals learn about the environment in which they receive reinforcing stimuli, be it natural reinforcers or reinforcing drugs, and, given the choice, they return to that environment, even when the reinforcer is not present. This indicates that associative learning, or conditioning, is occurring between the reinforcer and the environment. Similar learning can also be demonstrated to specific cues: in the lever press experiments described above, if a neutral stimulus (e.g. a light) is presented immediately before the lever is made available, the animals will approach the lever when the light stimulus is presented alone, and they will try to press the lever, even when it is still retracted and unavailable.

Microdialysis and FSCV experiments have shown that, once this learning has taken place, dopamine release in nucleus accumbens is increased during the presentation of the light stimulus, even long after the withdrawal of the reinforcer. Therefore, animals learn to associate specific cues and environment to the reinforcer, such that they can evoke the both release on dopamine in nucleus accumbens and reinforced behaviour, even in the absence of the reinforcer. These behaviours in experimental animals strongly resemble behaviours seen in drug addicts, where cues associated with drug taking (e.g. an empty vodka bottle to an alcoholic or a needle to a heroine addict), or the environment they associate with the drug taking can be very strong motivational drivers, or cravings, to take the drugs. Interestingly these conditioned effects can long outlast the period of association in both

experimental animals and in addicts: so cues and/or environment can trigger cravings in abstinent addicts even years after they last took the drug.

In the early stages, drug taking is not addictive: that is, people take the drugs through choice. A number of psychological factors originating in prefrontal cortex, including impulsivity and inhibitory self-control, have been shown to be vulnerability factors for drug taking. Thus, dysregulation of prefrontal control over mesolimbic circuits may underlie impaired inhibitory self-control exhibited by many addicts, while, high impulsivity, mediated through abnormalities of the orbitofrontal area of prefrontal cortex, may explain people's choice of short-term gratification of drug taking over long-term benefits of abstinence.

However, at some stage there is a change from use or abuse to the compulsive drug use characteristic of dependence. As mentioned earlier, this change includes neuro-adaptive processes involving sensitisation of dopamine systems controlling motivation and seems to be largely irreversible, accounting for enduring cravings even after long periods of abstinence. Moreover, evidence has shown that there is a learned component to this process: experimental animals are more likely to show sensitisation (e.g. sensitisation of locomotor activity, mediated through mesolimbic dopamine activity) if tested in the same environment where the initial drug administration took place. In addition, previous sensitisation enhances the acquisition of self-administration

and place preference, effects mediated through mesolimbic dopamine. With repeated drug administration, drugs may acquire greater and greater incentive value and become increasingly able to control behaviour. This may parallel the observation in drug addicts where places, acts or objects associated with drug-taking become especially powerful incentives.

Addictive drugs produce long-lasting changes in brain organisation. The brain systems that are sensitised include the dopaminergic mesolimbic pathway, responsible for the incentive salience ('wanting') of the drug or drug-associated cues. Systems mediating the pleasurable or euphoric effects of the drug ('liking') are not sensitised. Animal studies have looked at the mechanisms of sensitisation to repeated drug taking. Psychostimulants, including amphetamine and cocaine, cause increased locomotor activity in rodents, an effect which is mediated through the mesolimbic pathway: lesions of the mesolimbic pathway abolish it. On repeated systemic administration (i.e. intravenous, intraperitoneal or subcutaneous: where the drug accesses the whole brain), the hyperlocomotion induced by the drug increases, showing a sensitised response. The precise mechanism of the sensitisation is not certain, but it probably involves long-term and enduring neuro-adaptive changes in the cell body region in the VTA (see box below).

Localisation of neuroadaptation underlying sensitisation

Rats were given repeated local injections of amphetamine into either the cell body region of the mesolimbic pathway in the VTA, or the terminal region in the nucleus accumbens. A third control group received no injections.

Animals injected into the nucleus accumbens showed a hyperlocomotor response, which did not increase over repeated injections: that is, there was no sensitisation. Animals given injections into the VTA showed no behavioural response. This is not surprising, as the pharmacological effect of amphetamine is at the terminals, it increases release and blocks reuptake: therefore it is likely to be most effective in the terminal region. After these repeated local injections, animals were left for a week drug-free, then given a challenge dose of amphetamine systemically. All animals showed hyperlocomotion. Animals which had received repeated drug injections into nucleus accumbens showed a similar level of

hyperlocomotion to the non-injected controls: that is, there was no sensitisation. However, animals which had received drug into the VTA showed an augmented response compared to the other groups, indicating that sensitisation had taken place.

Therefore, repeated injection into nucleus accumbens evoked a behavioural response, but not sensitisation, whereas repeated injection into VTA produced no behavioural response, but did cause sensitisation, providing evidence for the critical role of VTA in sensitisation.

In summary, there is strong evidence from studies in experimental animals that:

1. dopamine signalling in the mesolimbic pathway drives motivational systems to promote behaviours critical for survival;
2. addictive drugs impact on this system to promote behaviours associated with drug-seeking and drug taking;
3. neuro-adaptation in this pathway accounts for the long-term, enduring nature of dependence and
4. activity in this pathway driven by conditioned associations can cause cravings to take the drug even after long periods of abstinence.

Therefore activity in this pathway can account for many of the phenomena associated with addiction in people.

Models of addiction

Several models have been proposed to account for the features of addiction, prominent amongst which is the incentive sensitisation model, proposed by Robinson and Berridge in 1993. This develops ideas taken from two other prominent models, the opponent process model and the aberrant learning model, and it is worth considering these two models briefly first.

Aberrant learning model

According to the aberrant learning model, abnormally strong learning is associated with drug taking, through two distinct components of learning. First, explicit learning where the association between action (drug taking) and outcome (drug effect) is abnormally strengthened leading to drug taking because of an expectation of the hedonic impact, even when the drug no longer produces that effect. Second, implicit learning where the action-outcome relationships (as above) change to more automatic stimulus-response relationship (habit), meaning that the stimulus evokes the response irrespective of any conscious expectations about the outcome.

While this theory accounts for the motivational drive from

stimuli associated with drug-taking and the ability of these stimuli to promote cravings, it does not account for the fact that most addicts do not report expectation of a positive hedonic effect. Therefore it seems unlikely that this could be the motivation for their drug seeking and taking. Similarly, it does not explain the compulsive nature of addiction – it implies that drug seeking and taking are purely automatic behaviours, whereas in fact they appear more as a motivational compulsion. Finally, it does not explain the behavioural flexibility shown by addicts. The theory would predict that if the normal route to drug taking were prevented, the addict would not be able to adapt behaviour in order to seek the drug from a different source or via a different process, whereas in fact addicts do show substantial behavioural flexibility in these circumstances.

Opponent process model

The opponent process model is well founded in neuroscience, as a mechanism for homeostatic control of many functions. It posits two processes, the A-process and the B-process which oppose each other: the A-process is activated by an external stimulus, leading to a change in functioning, and the B-process is the body's reaction to the change brought about by the A-process to return to the set point level. In the context of drug taking, the A-process represents the direct effect of the drug, which triggers the B-process, the opponent process, which

aims to restore the homeostatic state. The A-process leads to the hedonic state ('high') associated with taking a drug, while the B-process leads to the aversion from not taking the drug, for example the withdrawal symptoms. Over repeated drug taking, tolerance builds up to the A-process, accounting for the reduced hedonic impact of the drugs, while the B-process is strengthened, leading to withdrawal symptoms, which can only be eliminated by taking more of the drug. Thus the driving force for drug taking is to prevent the aversive withdrawal symptoms which occur when the A-process diminishes, but the B-does not. Thus, people who initially take drugs to gain a positive hedonic state, are subsequently motivated to continue drug taking to avoid a negative hedonic state.

This accounts for the drive to take the drug to achieve a homeostatic state, but does not account for evidence showing that avoiding the negative hedonic state of withdrawal is not a major motivator for drug taking. Indeed, many addictive drugs do not evoke strong withdrawal symptoms. Also, withdrawal symptoms are maximal in the days following abstinence, yet cravings for the drug, and reinstatement, even after a small dose, can last for years – in alcoholics who have been abstinent for years, a single alcoholic drink can reinstate the addictive behaviour.

Incentive sensitisation model

The incentive sensitisation model derives certain aspects from the above models, but puts them into a motivational framework. It delineates two distinct components of reinforcement – hedonic impact ('liking') and incentive salience ('wanting'), which are dissociable behaviourally and physiologically. Robinson & Berridge use the terms 'liking' and 'wanting' (in quotation marks) to represent these very clearly defined behavioural parameters. Thus, when given in quotation marks, they represent much more specific scientific terms than the everyday usage of the two words. Incentive learning, both explicit and implicit, which forms the core of the aberrant learning model, provide the route through which stimuli associated with the behaviour acquire incentive salience – they become salient, attractive and wanted – and guide behaviour.

The incentive sensitisation model focusses on how drug cues trigger excessive motivation for drugs, which drives drug seeking and drug taking behaviour. The subjective pleasure derived from taking the drug, the hedonic impact or 'liking', is due the direct psychopharmacological action of the drug in producing a 'high', reducing social anxiety and/or, increasing socialisation. Incentive salience, or 'wanting' on the other hand, represents the motivational importance of stimuli, making otherwise unimportant stimuli able to attract attention, making them attractive and 'wanted'. The critical

feature of the model is the dissociation between these two processes, both behaviourally and physiologically, and that the incentive salience, or 'wanting' is sensitised over repeated drug taking, so increasing the driving force to take drugs, whereas the hedonic impact, or 'liking', is unaffected, or may even reduce, through tolerance. Under normal conditions of natural reward the two processes work together to motivate behaviours which are beneficial to survival, it is only in unnatural situations such as taking of addictive drugs that there is a dissociation between the actions of the two systems, such that drugs can become exceptionally strong motivators of drug-seeking and drug-taking behaviour. This dissociation of the two components accounts for the observation that addicts continue to seek and take drugs, even when they derive little or no pleasure from it, and when they are fully aware of the physical, emotional and social damage it is causing. Importantly, the dissociation between 'wanting' and 'liking' has also been demonstrated experimentally, indicating that it is not simply a theoretical concept, but does actually occur.

The dissociation between 'wanting' and 'liking'

In a series of experiments designed to test whether a dissociation between 'wanting' and 'liking' could be demonstrated experimentally, Berridge and co-workers devised a scoring scheme for measuring facial expressions related to palatability across several species, through which they assessed 'liking' (Figure 6.3). Amphetamine was shown to have no effect on 'liking', and may have increased aversion.

In order to assess 'wanting', a lever press experiment was used in the same animals. Animals were first trained to press a lever for sucrose reward, then trained that one auditory stimulus signalled that the a lever press will deliver sucrose (CS+) but that a different auditory stimulus signalled that the level press will not deliver sucrose (CS-). Level press responses to CS+ and CS- were measured as an index of 'wanting'.

Amphetamine microinjection selectively enhanced

lever pressing for sucrose by the CS+ auditory stimulus, but not by the CS- auditory stimulus, indicating that amphetamine selectively enhanced the motivational element, 'wanting'. Therefore amphetamine had no effect on, or perhaps decreased, 'liking', but enhanced 'wanting', providing experimental evidence for the dissociation of the two which forms the basis of the incentive.

Figure 6.3 depicts representative hedonic tongue protrusions (reaction to sweet tastes) and aversive gapes (reaction to bitter tastes) from adult rat, young primate, and infant human (after Berridge).

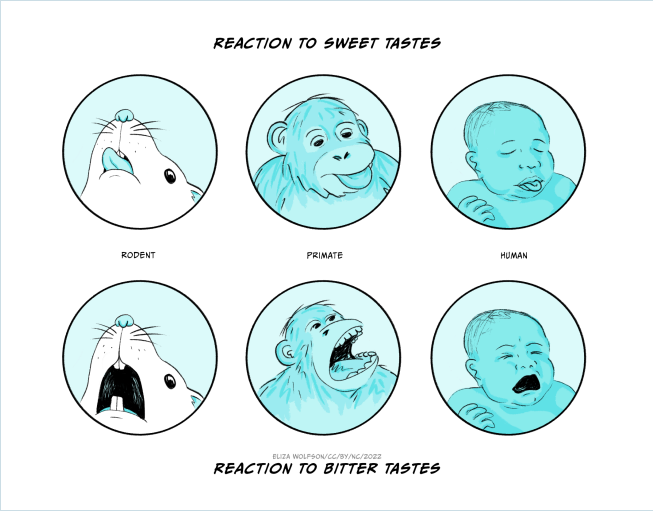

Fig. 6.3 Similar facial expressions to 'nice' and 'nasty' tastes in rodents and primates

Summary

(a) Increasing concentrations of amphetamine cause a small decrease in hedonic reaction, and an increase in aversive reaction, showing that amphetamine does not increase 'liking': indeed, it decreases it.

(b) Amphetamine causes an increase in responding to the auditory stimulus which has been paired with sucrose (CS+), but not to the auditory stimulus that was not paired with sucrose (CS-), showing that

amphetamine increases motivational drive, that is 'wanting'.

The mesolimbic dopamine pathway is primarily involved in control of incentive salience, while other parts of the basal ganglia circuitry, including those using opioids and acetylcholine, control the hedonic impact. This accounts for the role of dopamine is in the incentive salience aspect, but not the hedonic impact aspect: drugs which enhance dopamine function increase the incentive salience ('wanting'), increasing the animals motivation to pursue the goal, without increasing the hedonic impact ('liking'). While 'liking' of the drug may decrease with repeated exposure (probably due to tolerance), the motivation to take the drug increases through sensitisation of the incentive salience (incentive sensitisation). Thus the model explains the dissociation between the pleasure received from taking the drug, and the motivational drive to take it. The sensitisation of the incentive salience is similar to habit learning (aberrant learning model), but is distinguished from it by the fact that it is only one component of the response which is sensitised. Related to this, work from Everitt, Robbins and co-workers suggest that the switch to compulsive drug taking which characterises dependence may be mediated by a shift in the specific dopamine pathway controlling the response, from the neurones terminating in nucleus accumbens, when

responses are primarily goal directed, to neurones terminating more dorsally in the dorsal striatum when responses become compulsive (e.g. Everitt et al, 2001).

Addictive behaviours

There are a number of behaviours which share a lot of features in common with drug addiction. Behaviours like gambling and exercise can become compulsive with individuals carrying out the behaviours to the detriment of normal daily function or family relationships. A common feature with these behaviours is a dopaminergic component to the motivation, which may include activation of endorphin (an endogenous opioid, related to morphine) systems which in turn activate the mesolimbic dopamine pathway. Therefore some of these so called 'addictive behaviours' share many of characteristics of drug addiction, and evidence suggests that they may share similar neural mechanisms. A major research focus is aimed at identifying whether they are indeed different manifestations of the same process or different processes.

Treatment

Treatment options for drug addiction are fairly limited at present. The best long-term therapy is abstinence, although, as has been discussed earlier, people, specific cues and environments associated with drug taking can produce very

strong cravings, often leading to relapse, even after extended periods of abstinence – you will recall that, in rats, stimuli associated with reinforcement (e,g, drug administration) evoke dopamine release in nucleus accumbens long after the withdrawal of the reinforcer. In all therapeutic strategies for treating addiction, a vital consideration is that the individual must recognise that they have an addiction and they must be motivated to overcome it. Treatments can be physically and emotionally demanding, and without the motivation to stop, treatment is rarely successful.

Psychological therapies have proven fairly successful in sustaining abstinence. Cognitive behaviour therapy (CBT) helps recognize unhealthy behavioural patterns, identify triggers which may potentially lead to relapse, and develop coping strategies to overcome them. This may also include contingency management, which reinforces the positive aspects of avoiding drugs through specific rewards. Stepped management schemes are a form of group therapy which identifies negative consequences of addiction and through support networks develops strategies to overcome them. In the longer term, psychological therapies also look at aspects in the person's life beyond their addiction, and particularly any other pathological conditions they may experience. A key aim is to improve stress management, since we have seen that stress is a major precipitatory factor in relapse.

For psychological therapies to be effective, the individual must first stop taking the drugs, a process often called

detoxification: given the compulsive nature of drug addiction, this in itself can be a major challenge. The four main approaches used are drug elimination, agonist therapy, antagonist therapy or aversion therapy. Pharmacological treatments which reduce the impact of withdrawal symptoms and cravings can also help during the detoxification processes.

Drug elimination is where the person simply does not take the drug any more. Sometimes the drug is simply withdrawn, in a single step (e.g. very often when smokers give up smoking), but more often, particularly for more serious addictions, the daily intake of drug is slowly reduced under clinically controlled conditions, until the addict is no longer dependent on the drug. One of the main problems with this approach is that the person normally experiences withdrawal symptoms, which can be extremely unpleasant in some cases, and are a major motivator to relapse into a drug-taking habit.

Antagonist therapy is where an antagonist for the addictive drug is given to block the action of the drug. This form of therapy is rarely used as it induced very severe withdrawal effects, so much so that when antagonist therapy is used, the individual is normally anaesthetised or heavily sedated. Agonist therapy is probably the most widely used treatment for coming off drugs. In this case an agonist for the addicted drug, or in some cases the drug itself, is given but in a very controlled way, reducing the amount given over a period of time: normally also the drugs and/or route of delivery is less harmful. Finally, aversion therapy can be effective in some

cases, but is not widely used. This is where drug taking is paired with an aversive stimulus, such that a conditioned association is made between the drug and the aversive stimulus. For example an emetic drug is given alongside the addicted drug to induce sickness, making the addicted drug-taking aversive – you will remember the important role of conditioning in the development of addiction: well, it can also be used to treat it.

Key Takeaways

- Drug addiction is the compulsive use of drugs, to the detriment of daily functioning and relationships.
- Drugs which can become addictive have a wide variety of primary pharmacology, but all share the property that they provoke increased dopamine release in the mesolimbic pathway projecting from VTA in the midbrain to the nucleus accumbens in the forebrain.
- Many behavioural procedures in experimental animals have shown that this pathway is

important in motivation and that animals show a strong motivation to work (e.g. press a lever) in order to receive injections of drugs with addictive potential, providing a link between natural motivation networks and addiction.

- The incentive sensitisation model accounts for the phenomena of addiction by proposing a dissociation between motivation and hedonia, which can be demonstrated experimentally, and which accounts for the observation that drug addicts often report a heightened drive to take drugs, yet the enjoyment from taking them is diminished.

References and further reading

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, *28*(3), 309-369. https://doi.org/10.1016/S0165-0173(98)00019-8

Caine, S. B., Negus, S. S., Mello, N. K., Patel, S., Bristow, L., Kulagowski, J., Vallone, D., Saiardi, A., & Borrelli, E. (2002). Role of dopamine D2-like receptors in cocaine self-

administration: studies with D2 receptor mutant mice and novel D2 receptor antagonists. *The Journal of Neuroscience, 22*(7), 2977-2988. https://doi.org/10.1523/JNEUROSCI.22-07-02977.2002

Di Chiara, G. (1999). Drug addiction as dopamine-dependent associative learning disorder. *European Journal of Pharmacology*, *375*(1–3), 13-30. https://doi.org/10.1016/S0014-2999(99)00372-6

Di Chiara, G., & Imperato, A. (1988). Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proceedings of the National Academy of Sciences of the United States of America [PNAS] 85*(14), 5274-5278. https://doi.org/10.1073/pnas.85.14.5274

Everitt, B. J., Dickinson, A., & Robbins, T. W. (2001). The neuropsychological basis of addictive behaviour. *Brain Research Reviews*, *36*(2–3), 129-138. https://doi.org/10.1016/S0165-0173(01)00088-1

Everitt, B. J., & Robbins, T. W. (2016). Drug addiction: Updating actions to habits to compulsions ten years on. *Annual Review of Psychology*, *67*(1), 23-50. https://doi.org/10.1146/annurev-psych-122414-033457

Franken, I. H. A. (2003). Drug craving and addiction: integrating psychological and neuropsychopharmacological approaches. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *27*(4), 563-579. https://doi.org/10.1016/S0278-5846(03)00081-2

Goodman, A. (2008). Neurobiology of addiction: An integrative review. *Biochemical Pharmacology*, *75*(1), 266-322. https://doi.org/10.1016/j.bcp.2007.07.030

Heilig, M., MacKillop, J., Martinez, D., Rehm, J., Leggio, L., & Vanderschuren, L. J. M. J. (2021). Addiction as a brain disease revised: Why it still matters, and the need for consilience. *Neuropsychopharmacology*, *46*(10), 1715-1723. https://doi.org/10.1038/s41386-020-00950-y

Kalivas, P. W., & Weber, B. (1988). Amphetamine injection into the ventral mesencephalon sensitizes rats to peripheral amphetamine and cocaine. *Journal of Pharmacology and Experimental Therapeutics*, *245*(3), 1095-1102.

Koob, G.F. (2005). The neurocircuitry of addiction: Implications for treatment. *Clinical Neuroscience Research*, *5*(2-4), 89-101. https://doi.org/10.1016/j.cnr.2005.08.005

Koob, G. F., & Volkow, N. D. (2009). Neurocircuitry of addiction. *Neuropsychopharmacology*, *35*, 217-238. https://doi.org/10.1038/npp.2009.110

Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: A neurocircuitry analysis. *The Lancet Psychiatry*, *3*(8), 760-773. https://doi.org/10.1016/S2215-0366(16)00104-8

McKendrick, G., & Graziane, N. M. (2020). Drug-induced conditioned place preference and its practical use in substance use disorder research. *Frontiers in Behavioral Neuroscience*, *14*, 582147. https://doi.org/10.3389/fnbeh.2020.582147

Negus, S. S., & Miller, L. L. (2014). Intracranial self-stimulation to evaluate abuse potential of drugs. *Pharmacological Reviews*, *66*(3), 869-917. https://doi.org/10.1124%2Fpr.112.007419

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, *47*(6), 419. https://doi.org/10.1037/h0058775

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247-291. https://doi.org/10.1016/0165-0173(93)90013-P

Robinson, T. E., & Berridge, K. C. (2001). Incentive-sensitization and addiction. *Addiction*, *96*(1), 103-114. https://doi.org/10.1046/j.1360-0443.2001.9611038.x

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473-500. https://doi.org/10.1146/annurev.neuro.23.1.473

Vezina, P. (1993). Amphetamine injected into the ventral tegmental area sensitizes the nucleus accumbens dopaminergic response to systemic amphetamine: An in vivo microdialysis study in the rat. *Brain Research*, *605*(2), 332-337. https://doi.org/10.1016/0006-8993(93)91761-G

Wise, R. A. (1998). Drug-activation of brain reward pathways. *Drug and Alcohol Dependence*, *51*(1–2), 13-22. https://doi.org/10.1016/S0376-8716(98)00063-5

Wyvell, C. L., & Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: Enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *Journal of Neuroscience*, *20*(21), 8122-8130. https://doi.org/10.1523/JNEUROSCI.20-21-08122.2000

About the author

Dr Andrew Young
UNIVERSITY OF LEICESTER

Dr Andrew Young obtained a BSc degree in Zoology from the University of Nottingham, and his Ph.D in Pharmacology from the University of Birmingham. He then spent four years as a post doctoral researcher at Imperial College, London, studying glutamate release in the context of mechanisms of epilepsy, before moving to the Institute of Psychiatry (King's College, London) for nine years to study dopamine signalling in models of schizophrenia and addiction. In 1997 he was appointed as Senior Research Fellow in the School of Psychology at University of Leicester and is now Associate Professor in that department. His research interests focus mainly on neurochemical function, particularly dopamine, in attention and motivation, and in models of schizophrenia and addiction. He teaches topics in biological psychology and the biological basis of mental disease to both undergraduate and

postgraduate students in the School of Psychology and Biology.

16.

AFFECTIVE DISORDERS

Dr Andrew Young

Learning Objectives

- Know the main symptom clusters associated with affective disorders, including bipolar disorder and major depression
- Be aware of the diagnostic criteria used
- Know the fundamentals of the monoamine theory of depression
- Understand the theoretical underpinning of current approaches to pharmacological therapy for depression, based on the monoamine theory, and appreciate the shortcomings of current approaches

- Understand the theoretical basis linking depression to abnormalities in stress responses in the brain and appreciate how this theoretical framework informs novel antidepressant drug development.

Overview of affective disorders

Affective disorders, or mood disorders, are a group of psychological disturbances characterised by abnormal emotional state, and generally manifest as depressive disorders. When considering depressive disorders, the two most prevalent conditions are unipolar (major) depression and bipolar disorder, characterised by alternating depression and mania, although other conditions including dysthymia, cyclothymic disorder, seasonal affective disorder and pre and post-natal depression are also important (Figure 6.4). They occur across the lifespan, although incidence in pre-adolescents is low, and their characteristics are essentially the same across all ages and across cultures.

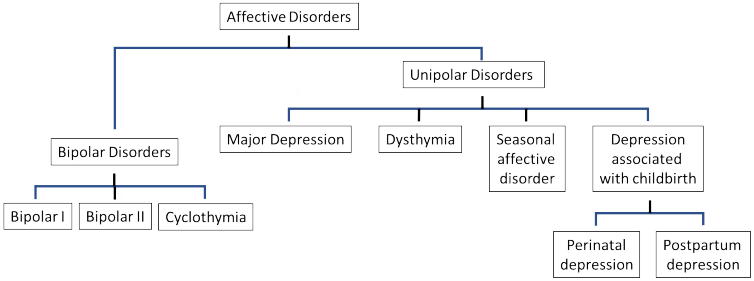

Figure 6.4. Types of affective disorder

Depression is characterised by persistent feelings of sadness, loss of interest, feelings of worthlessness and low self-esteem. Major depression and dysthymia share similar symptoms, with prolonged bouts of depressed mood: the main difference is the severity of the symptoms, with dysthymia showing less severe and less enduring symptoms. Bipolar disorder, on the other hand is characterised by similar periods of depression, but interspersed with periods of extreme euphoria, high activity and excitement and inflated self-esteem, termed mania. As with the depressive illnesses, the main difference between bipolar I, bipolar II and cyclothymia is the severity of the symptoms and the degree of interference with daily life.

Diagnosis of affective disorders uses diagnostic criteria laid down in the American Psychiatric Association International Classification of Diseases 11th Revision (DSM-5) or the World Health Organisation's International Classification of Diseases 11th Revision (ICD-11), which focus on the key features of

the condition and give guidance to clinicians for diagnosing the conditions.

Bipolar disorder

Bipolar disorder, formerly called manic depression, is characterised by cycles of extreme mood changes, from periods of a severely depressed state, resembling major depression (see below), to periods of extreme euphoria, high activity and excitement (termed mania). During a manic period people may experience inflated self-esteem and poor judgement, which may lead to them undertaking risky and often destructive behaviours; and a reduced need for sleep and a general restlessness, accompanied by physical agitation and a reduced ability to concentrate. They often deny that there is anything wrong, and become irritable, particularly when challenged about dubious decision making. It is not clear what causes mania. Genetic factors are implicated, since bipolar disorder tends to run in families, although no specific genes have yet been identified that link to it. However, genetic factors only account for around half of the vulnerability, so clearly environmental and social factors are also important.

There are three levels of severity of bipolar disorder. Bipolar I is the most severe form, and is characterised by manic episodes which last at least a week, while depressive episodes last for at least two weeks. The symptoms of both can be very severe and often require hospitalisation. Bipolar II is similar,

but less severe: in particular, the manic episodes are less intense, and less disruptive (often termed hypomania) and do not last as long. People with bipolar II are normally able to manage their symptoms themselves, and rarely require hospitalisation. There is a risk of progressing to bipolar I disorder, without correct treatment, but this can be kept to around 10% with the correct management. The least severe category is cyclothymia disorder, where people experience repeated and unpredictable mood swings, but only to mild or moderate degrees.

Diagnosis

Diagnosis of bipolar disorders require presence of both depressive symptoms (as below) and three or more of the listed features of mania (extreme euphoria, high activity, inflated self-esteem, poor judgement: see 'Diagnostic criteria for bipolar I disorder' box below). The main difference between diagnostic criteria for bipolar I, bipolar II and cyclothymia are the degree of severity and the time course of the expression of symptoms.

Diagnostic criteria (DSM-5) for bipolar I disorder

For a diagnosis of bipolar I disorder, it is necessary to meet the following criteria for a manic episode. The manic episode may have been preceded by and may be followed by hypomanic or major depressive episodes.

Manic episode

A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day.

During the period of mood disturbance and increased energy or activity, 3 (or more) of the following symptoms (4 if the mood is only irritable) are present to a significant degree and represent a noticeable change from usual behaviour:

- Inflated self-esteem or grandiosity
- Decreased need for sleep
- More talkative than usual or pressure to keep talking
- Flight of ideas or subjective experience that thoughts are racing
- Distractibility
- Increase in goal-directed activity or psychomotor agitation
- Excessive involvement in activities that have a high potential for painful consequences
- The mood disturbance is sufficiently severe to cause marked impairment in social or occupational functioning, or to necessitate hospitalisation to prevent harm to self or others, or there are psychotic features.
- The episode is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication, or other treatment) or to another medical condition.

Source: *The Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; DSM–5; American Psychiatric Association, 2013)

Incidence of bipolar disorder

Bipolar disorder is present in around 2% of the population, with bipolar I more common than bipolar II: lifetime prevalence is 1% and 0.4% respectively. Unlike major depression (see below) bipolar disorders are equally prevalent in males and females. Bipolar disorder can occur at any stage in the lifespan, although it is rare in pre-adolescents. Peak age of onset is between 15 and 25 years, although diagnosis may be considerably later, with the average age of onset of bipolar I disorder (18 years) a little earlier than for bipolar II disorder (22 years). It is a major cause of cognitive and functional impairment and suicide in young people.

Pathology of bipolar disorder

A number of brain abnormalities have been described in bipolar disorder, some of which overlap with those seen in unipolar depression (see below), but others appear to be specific to bipolar, and may represent changes responsible for the episodes of mania. Although the underlying neuronal abnormality causing mania is not well understood, changes in a number of chemical markers related to the regulation of pathways modulating neurotransmitter function and neurotrophic pathways have been described in cortex, amygdala, hippocampus and basal ganglia, suggesting compromised intracellular chemical signalling. Notably, there

is evidence for dysregulation of intracellular signalling pathways which regulate the function of a number of neurotransmitters, most notable of which are dopamine, serotonin, glutamate and GABA. This in turn may lead to the dysregulation of these transmitters which has been reported in mania. The decreased brain tissue volume reported in bipolar disorder, reflecting reduced number, density and size of neurones, may link to the compromised neurotrophic pathways leading to mild neuro-inflammatory responses and neurodegeneration reported in localised brain regions in mania. Therefore, although the pathology of mania seen in bipolar disorder is not well understood, it appears most likely that it derives from abnormalities in intracellular signalling cascades, perhaps related to localised neurodegeneration through decreased neurotrophic factors.

Treatment

First line treatment for bipolar disorder is antipsychotic medication: haloperidol, olanzapine, quetiapine or risperidone. These drugs target dopamine and serotonin signalling in the brain, and are likely to be downstream of the primary abnormalities associated with mania. If antipsychotic treatment is ineffective, then the mood stabilisers, including lithium, valproate or lamotrigine may be prescribed, either alone or in combination with antipsychotic drugs. Lithium has been widely used in the treatment of mania since its

introduction in 1949, but the mechanisms through which is has its mood-stabilising effects are still poorly understood. However, recent evidence has linked it to modulation of intracellular signalling pathways, particularly involving adenyl cyclase, inositol phosphate and protein kinase C: by competing with other metal ions which normally regulate these reactions (e.g, sodium, calcium, magnesium), but which may have become dysregulated, it is able to reverse instabilities in these reactions. Interestingly, other drugs, which also have mood-stabilising effects, including valproate and lamotrigine, also modulate these same intracellular signalling cascades. Therefore, the actions of lithium and other mood-stabilising drugs on these pathways provide supporting evidence for abnormalities in these intracellular signalling mechanisms in mania, perhaps opening novel routes for pharmacological therapy, but also provide plausible mechanism through which the drugs exerts their therapeutic action.

In addition to pharmacological treatment, psychotherapy has an important role to play in treatments of bipolar disorder. This may include cognitive behaviour therapy, which helps the individual to manage stress, and replace unhealthy negative beliefs with healthy positive beliefs; and well-being therapy which aims to help the individual manage stress, replace negative beliefs with positive beliefs and improve quality of life generally, rather than focusing on the symptoms. Psychotherapy is particularly important in managing

cyclothymia, to minimise the risk that it will develop into bipolar I or II disorder.

Major depression

Major depression is characterised by persistent feelings of sadness, which manifests as enduring and pervasive, 'blocking out' all other emotions. Associated with this is a loss of interest in aspects of life (termed anhedonia) which may start as general lethargy, but in its extreme it is a complete loss of interest in all aspects of daily life, including health and well-being. In addition to these emotional symptoms there is also a spectrum of physiological and behavioural symptoms, including sleep disturbances, psychomotor retardation or agitation, catatonia, fatigue or loss of energy. There are also cognitive symptoms including poor concentration and attention, indecisiveness, worthlessness, guilt, poor self-esteem, hopelessness, suicidal thoughts and delusions with depressing themes. Dysthymia, also termed persistent depressive disorder (DSM-5) essentially relates to similar symptoms, but less severe and with a more chronic time course. An individual can suffer from both major depression and dysthymia, which is termed double depression.

Diagnosis

The diagnostic criteria for major depression according to DSM-5 require the occurrence of feelings of sadness or low

mood and loss of interest in the individual's usual activities, occurring most of the day for at least two weeks (Table 2). Importantly, the symptoms must cause the individual clinically significant distress or impairment in social, occupational, or other important areas of functioning, and must not be a result of substance abuse or another medical condition. Diagnosis of dysthymic disorder is similar to that for major depression, but less severe: symptoms in all domains are at the mild to moderate level.

Summary of DSM-5 criteria for Major Depressive Episode

Five (or more) of the following have been present during the same two-week period and represent a change from previous functioning; at least one of the symptoms is either (a) depressed mood or (b) loss of interest or pleasure:

- Depressed most of the day, nearly every day
- Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly

every day
- Significant weight loss when not dieting or weight gain or decrease or increase in appetite nearly every day
- Insomnia or hypersomnia nearly every day
- Psychomotor agitation or retardation nearly every day
- Fatigue or loss of energy nearly every day
- Feelings of worthlessness or excessive or inappropriate guilt nearly every day
- Diminished ability to think or concentrate, or indecisiveness, nearly every day
- Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide
- The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning
- The episode is not attributable to the physiological effects of a substance nor to another medical condition
- The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophrenia,

schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders

- There has never been a manic episode or a hypomanic episode.

Source: *The Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; DSM–5; American Psychiatric Association, 2013)

Incidence

The overall prevalence of depression worldwide is estimated at around 5% of the population. Although there are some regional variations, prevalence rates world-wide are fairly similar with women around twice as likely (5 to 6%) as men (2 to 4%). As with bipolar disorder, incidence in pre-adolescents is very low, but the condition begins to emerge in adolescence, peaks in late middle age and then declines in old age. World-wide, depression is the leading cause loss of functionality at the population level, including absence from work and treatment costs, and major depression is the most prevalent mental disorder associated with the risk of suicide.

Causes of depression

Like many mental illnesses, the underlying cause is not yet known. It is likely that genetic, environmental and social factors contribute, and the exact origin may be different in different people. The main risk factors for an individual developing depression are a family history of depression, particularly if they experience severe or recurrent episodes, a history of childhood trauma, and major stressful life changes. In addition, some physical illnesses and medications can bring on a depressive episode.

Evidence suggests that offspring of people who suffer major depression are 2 to 3 times more likely to suffer from major depression themselves compared to the rate in the populations as a whole. This figure rises to 4 or 5 times greater risk if we consider only offspring of parents with recurrent depression or depression which developed early in life. Studies on identical twins suggest that major depression is around 50% hereditable, although this may be higher in the case of severe depression. Although there is clearly a genetic link, there is no one gene which is responsible for this vulnerability. Rather a vulnerability for depression is promoted by combinations of genetic changes. In adoption studies, a higher risk of an adopted child developing depression has been found if an adoptive (unrelated) parent has depression than if they are unaffected. This gives a clear indication that, as well as genetic influences, parents also clearly have a social influence.

Stress seems to be the most important environmental factor involved in the incidence of depression. The stress-diathesis model puts forward the notion that it is the interaction between stress and the individual's genetic background which determine the expression of depression. Studies on childhood trauma show that children who have experienced emotional abuse, neglect and sexual abuse have an increased likelihood of developing depression in the future of around three-fold, and around 80% of depressive episodes in adults are preceded by major stressful life events. Therefore it is likely that stressful life events, be they in the distant past or more recent, are both a vulnerability factor and a precipitatory factor in the origin of depression.

Beck's cognitive triad provides a mechanism through which stressful life events may impact on altered cognition leading to a tendency to interpret every-day events negatively leading to the development of depression. Essentially he proposed that the combination of early life experiences and acute stress led to negative views of oneself, the world and the future (the cognitive triad), which in turn created negative schema with a cognitive bias towards negative aspects of a situation, an overemphasis on negative inferences and an overgeneralisation of negative connotations to all aspects of a situation. While these factors may in themselves be sufficient to invoke a depressive episode, it becomes more likely in those with a genetic predisposition (Figure 6.5).

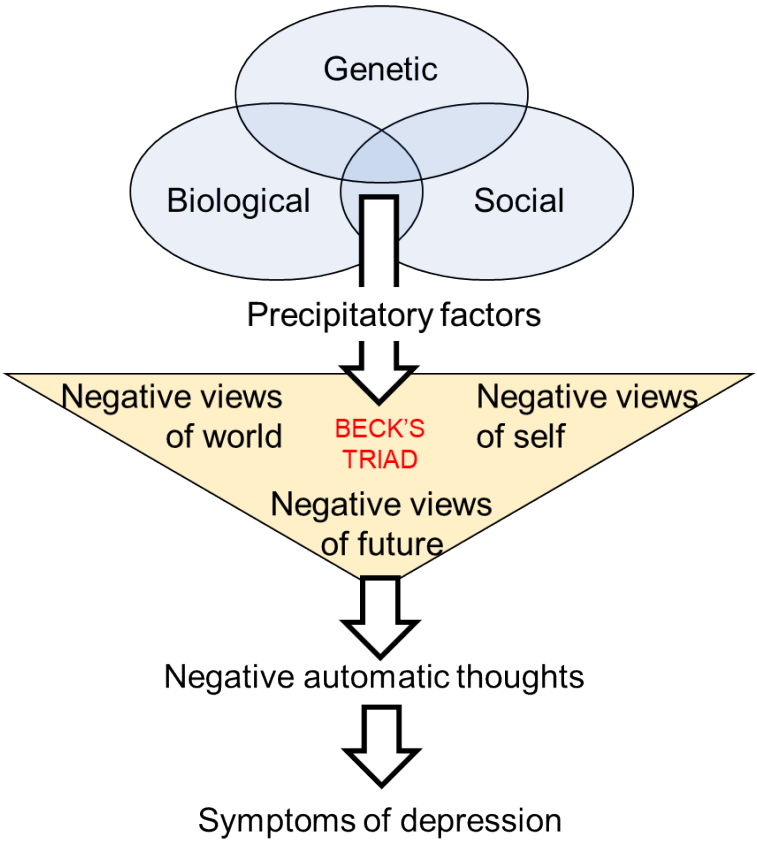

Fig 6.5. Origins of depression

Vulnerability is brought about by a combination of a genetic predisposition, biological factors (e.g. brain structure, hormones) and social factors (e.g. upbringing, childhood trauma). Precipitatory factors, such as stressful life events, trigger negative views on one's self, the world and the future

leading to negative automatic thoughts, including negative attribution, and the symptoms of depression.

Brain structural abnormalities in depression

In the search for a biological origin for depression no consistent abnormalities in brain structure or connectivity have been identified linked to depression. There have been reports of decreased tissue volumes in prefrontal, anterior cingulate cortices and hippocampal regions associated with major depression, but results across studies are inconsistent, although individuals with reduced hippocampal volume seem to be more prone to relapse. Neuroimaging studies have found a number of functional abnormalities in localised brain regions associated with depression, with prefrontal and anterior cingulate cortices emerging as the most likely areas of dysfunction, but again there is a lack of consistency between studies. A decrease in metabolism has been reported in dorsal prefrontal cortex, which is reversed after antidepressant treatment. Decreased metabolism has also been reported in the ventral portion of the anterior cingulate cortex, a region that has extensive connection with brain regions involved in mood regulation, including amygdala, orbitofrontal cortex and medial prefrontal cortex. On the other hand, insular volume and insular activation in response to negative stimuli has been reported to be increased in major depression, suggesting a

heightened sensitivity to adverse stimuli and situations. It has been suggested that depression is not caused by an abnormality in a single brain region, but rather imbalances of connectivity distributed across a number of brain regions. Connectivity studies have shown that the decrease in metabolism in regions of the frontal cortex correlates with increased metabolic function in striatal areas, suggesting a dysregulation of corticostriatal networks in depression.

The monoamine theory of depression

The drug reserpine, extracted from the Indian snake root plant, was historically used as a tranquilliser and for treating hypertension. It was found to cause severe, often suicidal tendencies in patients treated with it. In the 1960s, it was found that the pharmacological action of reserpine is to deplete releasable monoamine neurotransmitters (dopamine, noradrenaline, serotonin) in the terminals, by preventing their storage in vesicles. This therefore supports the view that depression may be accompanied by a reduction in monoamine neurotransmitters, a view that was brought together by Joseph Schildkraut in 1965 in the monoamine theory of depression, in which he concluded that depression is caused by reduced monoamine neurotransmitter function. Subsequently, the primary importance of serotonin and noradrenaline were realised, with dopamine playing a lesser role. This is consistent

with the observation that many of the drugs used to treat depression are more effective in influencing serotonin and noradrenaline systems than dopamine systems.

Serotonin is involved in many functions which are disrupted in depression, including pain sensitivity, emotionality and responses to negative consequences. Some studies have shown that the serotonin metabolite, 5HIAA, is reduced in the cerebrospinal fluid (CSF) of depressed patients, although the data are inconsistent. Low 5-HIAA seems to be particularly associated with aggressive hostile and impulsive behaviour and has been reported in violent suicide attempts. Decreased serotonin is found in post mortem brains of some depressed patients, but again the data are inconsistent. Thus, overall depression does seem to be associated with decreases in brain serotonin function.

There is little evidence of decreased noradrenaline or noradrenaline function in post mortem brains of depressed people. Although some studies found changes in the noradrenaline metabolite (MHPG) in CSF or blood of depressed patients, no consistent decrease has been found in depressed patients, as would be predicted if decreased noradrenaline levels were causing depressive symptoms. However, increased MHPG has been observed after successful treatment with antidepressant, which would be consistent with the antidepressants increasing noradrenaline function. Therefore, though these findings suggest some involvement

of noradrenaline in depression, the precise relationship is not clear.

Treatment

Prior to the 1950s, there were no suitable drug treatments available to treat the symptoms of major depression. Earlier drug treatments were non-specific and simply aimed at suppressing troublesome symptoms, often by extreme sedation. Relief of symptoms was poor and many patients developed dependence and/or toxic reactions. Shock therapy, first introduced for the treatment of schizophrenia in the 1930s, involves inducing seizures, which were seen to be beneficial in treating mental disorders. Early shock treatment used insulin shock and chemical shock (Cardiazol) but an alternative to these, electroconvulsive shock (ECT), was introduced in the 1940s: this involved passing an electrical current through the brain, inducing seizures. It was seen as a safer way of inducing seizures than either insulin or Cardiazol, and became widely used in the treatment of mental disorders, including depression. The induction of severe seizures, although therapeutically beneficial, often caused fractures, broken teeth and torn muscles and ligaments from the violent convulsions, and in many cases left residual long-term amnesia and personality changes. Although ECT is still used in extreme cases, where patients do not respond to drug treatment, it now performed in a controlled environment using much lower

currents, and is carried out using muscle relaxants and under general anaesthetic to prevent injury and distress during the procedure. Although the precise mechanism of ECT is not fully understood, it appears to cause changes in brain chemistry which rapidly alleviate symptoms of a number of mental conditions including depression. ECT is one of the most effective treatments for severe depression, particularly in patients who do not respond to drug treatment.

Monoamine therapy

In the early 1950s, a drug called iproniazid was being used as an antibiotic in the treatment of tuberculosis, and clinicians reported that it also seemed to elevate the mood of the patients. It was therefore tested on depressed patients and found to alleviate the symptoms of depression. Empirical studies ensued and in 1956 the first formal report of an antidepressant effect of iproniazid was published by Kline in 1956, and it was subsequently marketed as an antidepressant. Pharmacologically, iproniazid is a monoamine oxidase inhibitor (MOAI). It blocks the action of the enzyme, monoamine oxidase, which breaks down the monoamine neurotransmitters serotonin, noradrenaline and dopamine, thus increasing their concentrations in the synaptic cleft. The therapeutic benefits seen with iproniazid suggest that these monoaminergic transmitters may be depleted in depression. In particular, iproniazid blocks MAO-A, which breaks down

serotonin and noradrenaline preferentially, whereas dopamine is broken down by MOA-B. These effects of iproniazid form part of the evidence that changes in serotonin and noradrenaline are more important in depression than dopamine changes.

MAOIs are associated with a number of side effects mediated in the brain, including insomnia, confusion, drowsiness and nausea. However, the most problematic, called tyramine-induced hypertension crisis, derives from the fact that as well as blocking the breakdown of monoamine neurotransmitters, MAOIs also block the breakdown of the amino acid, tyramine, in the liver. This leads to an increase of tyramine in the blood causing dangerous increases in blood pressure and intracranial bleeding, which can be fatal. Therefore care with diet is necessary, since many foods contain high levels of tyramine. One such food is cheese, hence it is commonly called the 'cheese effect', but also include wine and chocolate.

Newer reversible inhibitors of MAOIs (RIMA: e.g. moclobemide) reduce this danger. Unlike standard MOAIs, which once bound to the enzyme remain bound, binding of the RIMAs is reversible. Therefore, when tyramine concentrations increase it competes with the drug for binding at the enzyme. In this way tyramine levels never reach the dangerous levels required to trigger tyramine-induced hypertension crisis.

Another class of drugs which largely superseded MAOIs are

the tricyclic antidepressants, so called because of their chemical structure containing three benzene rings. The first of these was imipramine, which was approved for use in 1959: structurally it is similar to the antipsychotic drug, chlorpromazine, and actually derived from drug development to try to isolate a drug with similar antipsychotic properties as chlorpromazine, but without the motor side effects. Imipramine showed very little antipsychotic effect, but did have antidepressant properties. Since then a number of other tricyclic antidepressants have been developed, including amitriptyline, clomipramine, desipramine and nortriptyline. Pharmacologically, they inhibit the reuptake of serotonin and noradrenaline back into the terminal after release. As with MAOIs, this prolongs the time the transmitter molecules are in the synaptic cleft and therefore increases their trans-synaptic signalling. Interestingly, the tricyclics with the best antidepressant profile are those which primarily block serotonin and noradrenaline reuptake: those which block dopamine reuptake are less effective, supporting the view that dopamine is less involved in the origin of depression than serotonin or noradrenaline. The main problem with the tricyclic antidepressants is that they are not very specific: therefore as well as blocking monoamine reuptake they are also antagonists at acetylcholine, noradrenaline and histamine receptors. Therefore, not surprisingly, as well as treating depressive symptoms they also have wide ranging side-effects, most notably hypotension, cardiac arrhythmia/arrest, sedation and memory disturbances,

some of which can be fatal, particularly in overdose. Although they are effective and relatively cheap, the side effects limit compliance and they are generally no longer used as a first line treatment.

The efficacy of tricyclic antidepressants paved the way for the development of more specific drugs, which have the required therapeutic action, but without the problematic side effects. These are serotonin reuptake inhibitors (SSRIs), noradrenaline reuptake inhibitors (NRIs), and serotonin and noradrenaline reuptake inhibitors (SNRI) which, as their names indicate, block only reuptake of serotonin, noradrenaline or serotonin and noradrenaline respectively, without having the non-specific effects on other neurotransmitter systems which cause the array of side effects seen with tricyclics. These drugs show much better clinical efficacy, with fewer side-effects (although they do still produce some side effects), thus improving compliance.

SSRIs are powerful inhibitors of serotonin reuptake, with minimal effects on noradrenaline reuptake, or on other neurotransmitter systems. Examples include fluoxetine, paroxetine, sertraline and citalopram. As such they induce fewer side effects, although those that they do induce are mediated through serotonin systems outside the primary therapeutic targets in the brain, and are therefore hard to dissociate from the required therapeutic action. These side effects include acute anxiety and panic attacks, akathisia

(constant restlessness and inability to remain still), sleep disturbances and nausea.

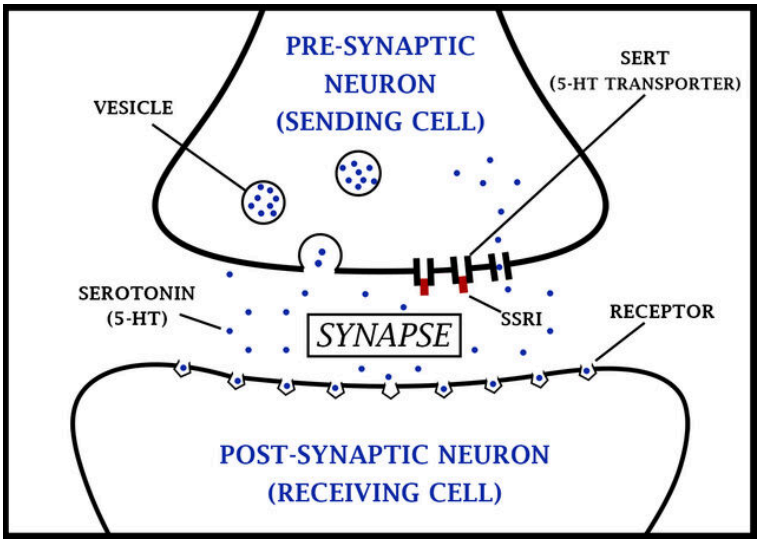

Fig 6.6. How SSRIs work

SNRIs inhibit the reuptake of both serotonin and noradrenaline, and as such show a similar reuptake blocking action to the tricyclic antidepressants, but without receptor mediated side-effects. Examples include duloxetine and venlafaxine. Side effects of SNRIs include nausea, insomnia, and loss of appetite. NRIs, for example atomoxetine and reboxetine, block only noradrenaline reuptake. Although they are generally less effective as antidepressants, they can be a preferable therapeutic route for severe depression and where depression is accompanied by significant anxiety.

SSRIs or SNRIs are currently the first line of treatment

for major depression. Although they are not fully effective in controlling the symptoms, around 50% of patients show good control of symptoms, with a further 25% showing some improvement but still exhibiting debilitating symptoms. Even in those patients where symptoms are well controlled, there is a 4 to 6 week delay before the antidepressant effect emerges. This implies that a more complex action than simply blocking of reuptake is involved, as this direct pharmacological effect will occur within 1 to 2 hours of the drug administration. The other major limitation of SSRI and SNRI treatment is the side effects associated with treatment. Although these are generally not severe or life threatening, they are nevertheless unpleasant and cause some disruption to daily life, leading many patients to discontinue drug treatment after recovery from a depressive episode, with the high risk of relapse. This therefore provides a challenge for future drug design, to develop drugs with a faster antidepressant action, which are effective in all patients and with fewer side effects.

Transcranial Magnetic Stimulation

The recent advance in transcranial magnetic stimulation (TMS) technology has enabled targeted brain stimulation to be employed. In the field of depression, there has been some success in TMS treatments particularly stimulating areas of the dorsolateral prefrontal cortex. These are the same areas which

have been shown to have reduced metabolism in depressed patients. Imaging studies have highlighted some functional abnormalities in brain circuits in depressed patients, and as our understanding of these aspects improves, there is the potential for this type of treatment to become more effective in a wider group of patients.

Novel approaches to antidepressant drugs

All current pharmacological approaches to treating depression derive from the monoamine theory of depression and aim to increase serotonergic and/or noradrenergic transmission. Novel drugs licensed as antidepressants over the last three decades have essentially relied on very similar pharmacology, the main improvements being in specificity, leading to reduced side-effects, and potency. While clearly these monoamines are involved, and any alternative theory as to the origin of depression must be able to account for the at least moderate efficacy of these drugs are alleviating symptoms, the anomalies highlighted above (onset delay, efficacy) suggest that they may not be targeting the core deficit. However, what holds us back in developing novel therapeutic strategies which do target the core deficit, is that the core deficit has not yet been conclusively identified.

To address this issue, two key lines of evidence have been important. First, the prominence of stress as a predisposing

and precipitating factor in psychological models of depression, and second, that the common process in many animal models used to study depression and antidepressants involves applying stressors to the animals (e.g. forced swim test, learned helplessness, chronic mild stress, maternal separation). These observations suggest that the body's stress response system may be involved and that function may be compromised in depression. These stress responses are controlled in the hypothalamus-pituitary-adrenal cortex system (HPA-axis).

A key element of this response is the release of glucocorticoids from the adrenal cortex, a secretory organ located on the dorsal the kidneys. In humans, the main glucocorticoid is cortisol, and this is responsible for many of the reactions to stress in the body – blood pressure, heart rate, reduced gut motility, arousal, etc. Under normal conditions, high cortisol levels in the blood also act as a negative feedback mechanism in the brain to switch off the HPA-axis, by reducing activity in the corticotrophin releasing hormone (CRH) producing cells in the hypothalamus (note that CRH was previously called corticotrophin releasing factor, CRF, before its hormone characteristics were identified). Thus the stress response is self-limiting and returns to normal levels when the stressor is removed.

The Hypothalamus-Pituitary-Adrenal cortex system (HPA-axis)

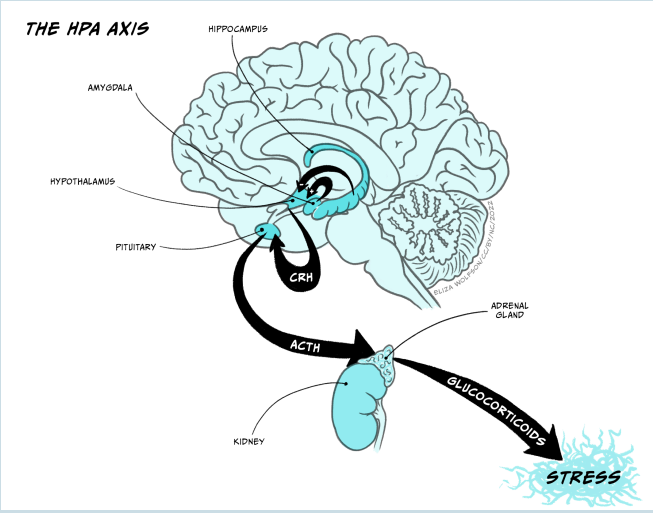

Fig 6.7. The HPA-axis. Diagram of main components of the HPA axis (black), under inhibitory control from the hippocampus and excitatory control from the amygdala

The hypothalamus-pituitary-adrenal axis (HPA-axis) is crucial in controlling the body's response to stressors and aversive situations. Neurosecretory

cells project from the hypothalamus to the pineal stalk, where they release corticotrophin releasing hormone into the pituitary postal blood capillaries. Increased CRH activates secretory cells in the posterior lobe of the pituitary (median eminence) to release adrenocorticotrophin (ACTH) into the blood system, which triggers the body's response to the stressor – essentially the 'fight or flight' response. Two key inputs regulate the activity of the CRH producing cells in the hypothalamus, and through that regulate the activity in the HPA-axis. These inputs arise from hippocampus, which inhibits CRH cells, and the amygdala, which activates CRH cells. It is well documented that the amygdala is important in responses to aversive stimulation and it is likely that it is a key structure in activating HPA-axis responses to a stressor.

Several lines of evidence suggest that HPA-axis function is abnormal in depression. Many depressed patients have been found to have elevated levels of ACTH and cortisol, enlarged pituitary and adrenal glands, raised levels of CRH in cerebrospinal fluid and abnormal circadian rhythm of cortisol,

all indicative of dysregulation of the HPA-axis in depression. Importantly, although the mechanism is unclear, treatment with current antidepressant drugs reduces CRH levels in depressed patients, indicating effective treatment of depression normalises, at least to some extent, the HPA dysfunction. Furthermore, Cushing's disease, which involves excessive secretion of glucocorticoids, is commonly followed by depression, and corticosteroids used in treatment of arthritis often cause depression. Therefore there is a substantial body of clinical evidence suggesting that hyperfunction of HPA may be linked to depression.

There is also support for this view from animal studies. CRH administration in rodents causes an increase in cortisol and also increases behaviours resembling symptoms of depression (e.g. insomnia, loss of appetite), effects which are reversed by antidepressant drugs, again providing a link between current treatment and HPA-axis function. Also in rodents, maternal separation, modelling childhood trauma, causes an elevation of stress-induced CRH, ACTH and cortisol release in adulthood, and increased CRH gene expression, effects which again are reversed by antipsychotic drugs. The effects of antidepressants on augmented gene expression are of particular interest and may account for the delay in therapeutic effectiveness of these drugs, since changes in gene expression are likely to take weeks, rather than hours, to manifest.

Novel treatments strategies derived from HPA research

CRH-related therapy

In animal studies, CRH antagonists have shown an antidepressant-like profile: the CRH antagonist LWH234 decreases time spent immobile in the forced swim test. In clinical trials the CRH antagonist, R121919, shows early promise, providing a significant improvement in depression, with minimal side effects: notably, the depression worsens again after the end of drug treatment.

Neurokinin-related therapy

Animal models have shown antidepressant-like effects of NK-1 receptor antagonists and clinical trials have shown that the NK-1 antagonist, MK869, reduces depressive symptoms in patients with major depression, to a similar extent to SSRIs. Notably, however, the time course is also similar to SSRIs, so these drugs, although potentially beneficial in the

future, may share the same delayed response as current medicines.

The CRH-secreting cells of the hypothalamus are under direct inhibitory control of the hippocampus. Circulating cortisol activates the hippocampus, thus increasing inhibition of the hypothalamus, providing a plausible mechanism for the self-regulating negative feedback curtailing HPA activation. Damage to the hippocampus would lead to a prolongation of HPA activation, with an increase in CRH, ACTH and cortisol. Quantitative magnetic resonance imaging (MRI) studies have shown decreased hippocampal volume in severely depressed patients, providing evidence that such damage may have occurred. In addition, cortisol has a regulatory influence, both positive and negative, on many genes expressed in the brain, which can account for behavioural changes, including depressed mood, after continued overexposure.

Animal studies have shown that high levels of circulating cortisol have neurotoxic effects on hippocampal neurons. These include decreased dendritic branching, loss of dendritic spines (the location of synapses) and reduced hippocampal neurogenesis. This effect may be mediated through reduced levels of neurotrophic factors, in particular brain derived neurotrophic factor (BDNF), which are essential for maintaining healthy neurons: reduced BDNF compromises

neuronal function, which can lead to severe functional deterioration, or even death, of the neurons. Low BDNF may be responsible for the reduction in dendrites seen with prolonged high levels of circulating cortisol. BDNF levels are decreased in the brains of people committing suicide, and chronic stress reduces hippocampal BDNF in rats and suggests promoting BDNF activity as a possible therapeutic target. BDNF is involved in a number of intracellular second messenger cascades where therapeutic drugs could act. Monoamine antidepressant drugs may protect vulnerable cells by preventing the decrease in BDNF, and importantly this may be dependent on chronic treatment, consistent with the delay in onset of therapeutic effect, although the mechanism is unclear. For example, chronic, but not acute, antidepressant treatment increases BDNF in both experimental animals and humans, and prevents stress-induced reductions in BDNF, probably through up-regulation of intracellular second messenger pathways responsible for BDNF production.

Antidepressant effects of ketamine

Ketamine, and particularly its active enantiomer, S-

ketamine (esketamine) is well known as an anaesthetic, but recently has received attention in management of a number of clinical conditions, including depression, where it has been found to provide a rapid onset (within four hours) antidepressant effect in treatment-resistant depression. Although it is now licenced for use in treatment-resistant depression in the United States, uncertainties about functional outcomes, side-effects and cost effectiveness have delayed its adoption as a main-line treatment elsewhere.
The mechanism through which ketamine exerts its antidepressant action is unclear. The main documented pharmacological action of ketamine is as a non-competitive antagonist at NMDA-type glutamate receptors. However, it also has pharmacological effects at other neurotransmitter systems, including monoamine, opioid and cholinergic mechanisms, all of which may contribute to its antidepressant effect. However, ketamine also exerts a powerful regulation of intracellular signalling cascades that increase neuronal and glial trophic factors (BDNF, GDNF) and inhibit microglia associated with inflammation. These factors, in turn, lead to a decrease in neurodegeneration and an

increase in neuronal proliferation and synaptogenesis. This potential mechanism of action of ketamine is particularly pertinent in view of the evidence of hippocampal degeneration and lowered BDNF in depression, and indeed, chronic ketamine treatment reverses the reduced BDNF seen in hippocampus in depression. These observations have given renewed impetus to the search for mechanisms of intracellular regulation which may be responsible for depression, and which could be novel targets for antidepressant action.

It is well documented that the amygdala is important in processing aversive and emotional stimuli. Amongst many other outputs, mediating various responses to such stimuli, the amygdala exerts an excitatory influence on CRH releasing cells in the hypothalamus. It is the balance between the excitatory input from the amygdala, present during the stressor, and the inhibitory input from the hippocampus, including from the negative feedback mechanisms, which controls the activity in the HPA-axis. Thus when the stressor is still present, the amygdala retains its excitatory drive, overcoming the inhibitory influence of the hippocampus, but when the stressor recedes the excitatory drive diminishes and the

inhibitory influence from the hippocampus predominates and reduces activity HPA activity. However, if hippocampal function is compromised, the negative feedback will be ineffective, and the HPA-axis will remain active, maintaining a high concentration of circulating cortisol, potentially causing further damage to the hippocampus, creating a self-perpetuating cycle of damage. Similarly, during chronic stress, the amygdala remains active, driving the HPA-system and maintaining a potentially toxic level of cortisol in the circulation for an extended period, leading to compromise of hippocampal function. Finally, during neurodevelopment, hippocampal cells are particularly susceptible to high cortisol levels, causing potentially irreversible changes, explaining why childhood trauma may be particularly damaging.

This opens another route for potential novel therapeutic strategies, by reducing the excitatory drive from the amygdala. Neurokinin peptide neurotransmitters, including substance P, are involved in signalling about aversive stimuli, particularly in the amygdala. Some studies have shown increases in substance P in the CSF of depressed patients, suggesting abnormalities in substance P signalling, which could potentially be normalised with neurokinin (NK) receptor antagonists.

Other depressive illnesses

Although bipolar disorder and major depression are the two main types of affective illness, there are others which should

not be overlooked. Seasonal affective disorder (SAD: termed Major Depressive Disorder with Seasonal Pattern in DSM-5) is a type of depression, with similar features to major depression, which is brought about by seasonal change. It affects around 2% of the population, occurring most often in young adults, and women are more affected than men. It is believed that decreased sunlight during winter months and increased sunlight in summer months affects the natural diurnal rhythms that control hormones, sleep and moods, and it is particularly prevalent in people living in extreme northerly and southerly regions of the earth, where the daylight differences between summer and winter are most extreme. Notably the majority of SAD is associated with decreased sunlight in the winter months, with only around 10% of cases associated with increased sunlight in the summer months. For winter SAD, light therapy is reasonably effective, where people are exposed to light of certain wavelengths from a specialised light box for around 30 minutes each day. Alternatively standard antidepressants and cognitive behavioural therapy are also effective.

Depression associated with pregnancy and childbirth are relatively common. Around 15% of women experience symptoms of depression during pregnancy, which, although often mild, can be severe in some cases. Similarly, after the baby is born, 60 to 80% of mothers experience mild depression, often termed ‘baby blues’, which is normally transient, with each bout lasting no more than an hour. Normally it does not

occur beyond 2 to 3 weeks after birth, and does not generally require treatment more than providing practical and emotional support for the mother. More serious, though, is postpartum depression, occurring in around 10% of new mothers. Where symptoms are more severe, bouts last for longer and it continues for months or even years after birth. The cause of these depressive conditions is thought to relate to the rapid and extreme hormonal changes that occur during pregnancy and childbirth, although, like other forms of depression, social stress and trauma may trigger or exacerbate the depression. Interestingly, postnatal depression has also been found in 5 to 10% of new fathers, indicating that there cannot be an entirely hormonal basis: the stress of the changed lifestyle may also be an important trigger. Psychological therapy, including cognitive behaviour therapy and life-style advice (exercise, diet) has been found to be effective in most cases, although antidepressants are appropriate for more severe symptoms and where the symptoms do not respond to psychological treatments. Importantly, if left untreated, postpartum depression can develop into persistent major depression.

Summary

The two main affective disorders are bipolar disorder, characterised by fluctuations between severe depression and mania, and major depression characterised by severe

depression alone. There are essentially three levels of bipolar disorder, distinguished by the severity of the symptoms: the most severe is bipolar I, followed by bipolar II, then cyclothymia. Similarly, dysthymia is distinguished from major depression by the less severe symptoms. The monoamine theory of depression places a major emphasis on decreased functionality of serotonin and noradrenaline in the origin of depression, and current antidepressant treatments derive very much from that view. First line treatment normally uses SSRIs or SNRIs which block the reuptake of serotonin and serotonin and noradrenaline respectively, although other classes of antidepressant drugs, tricyclic antidepressants and MAOIs, also aim to increase function of these neurotransmitters. However, none of these drugs are effective in all patients, they are all associated with side effects, some more severe than others, and there is always a delay of around 4 to 6 weeks before they begin to show antidepressant effects. Therefore an alternative therapeutic approach has been sought.

A common theme in psychological and animal models of depression is the role of stress, leading to attention being paid to the HPA-axis, the main stress control system, in relation to depression. Studies in both animal models and in depressed patients have indicated that the HPA-axis may be compromised in depression and dysregulation of the inhibitory and excitatory inputs from hippocampus and amygdala (respectively) have been implicated. In particular decreased function in hippocampus leads to reduced

inhibition of the HPA-axis, and so an enhanced, or extended, stress response, and this may provide a target for future drug treatments. Importantly, also, monoamine antidepressant drugs have been shown to modulate HPA-function, particularly when given chronically, providing a plausible route for their therapeutic action, and an explanation for the delay in onset of their antidepressant action.

Key Takeaways

- The two main affective disorders are:
    - Bipolar disorder, characterised by fluctuations between severely depressed mood and extreme euphoria, high activity and excitement (mania)
    - Major depression (unipolar depression), characterised by a severely depressed state alone
- Mechanisms underlying bipolar disorder, and its treatment with mood-stabilising drugs are poorly understood, but are likely to involve

regulation of intracellular signalling pathways controlling neuronal activity and neurotransmitter function

- The monoamine theory of depression posits that major depression is caused by an underactivityof monoamine neurotransmitters, particularly serotonin and noradrenaline
- Current treatments for major depression are based on the monoamine theory, and aim to increase the function of serotonin and noradrenaline. These treatments include monoamine oxidase inhibitors, which prevent the enzymatic breakdown of the transmitters, tricyclic antidepressants, which block the reuptake of both serotonin and noradrenaline; and specific reuptake inhibitors for serotonin (SSRIs) or both serotonin and noradrenaline (SNRIs)
- These drugs are not effective in all patients, they are relatively slow in onset, and they are accompanied by problematic side effects, meaning that for many people there is inadequate control of symptoms
- Compromised stress responses, mediated

through the hypothalamus-pituitary-adrenal cortex (HPA) axis has been proposed as an underlying cause of depression
- Treatments targeting stages in the HPA axis signalling are showing promise as novel antidepressant drugs.

References and further reading

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th edn. American Psychiatric Publishing.

Bissette, G., Klimek, V., Pan, J., Stockmeier, C., & Ordway, G. (2003). Elevated concentrations of CRF in the locus coeruleus of depressed subjects. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *28*(7), 1328–1335. https://doi.org/10.1038/sj.npp.1300191

Deussing, J. M. (2006). Animal models of depression. *Drug discovery today: disease models*, *3*(4), 375-383. https://doi.org/10.1016/j.ddmod.2006.11.003

Elhwuegi, A. S. (2004). Central monoamines and their role in

major depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *28*(3), 435-451. https://doi.org/10.1016/j.pnpbp.2003.11.018

Gainetdinov, R. R., & Caron, M. G. (2003). Monoamine transporters: from genes to behavior. *Annual Review of Pharmacology and Toxicology*, *43*, 261-284. https://doi.org/10.1146/annurev.pharmtox.43.050802.112309

Holsboer, F. (2000). The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology*, *23*(5), 477-501. https://doi.org/10.1016/S0893-133X(00)00159-7

Kato, T. (2019). Current understanding of bipolar disorder: Toward integration of biological basis and treatment strategies. *Psychiatry and clinical neurosciences*, *73*(9), 526-540. https://doi.org/10.1111/pcn.12852

Kramer, M. S., Cutler, N., Feighner, J., Shrivastava, R., Carman, J., Sramek, J. J., Scott, A.R., Guanghan, L., Snavely, D., Wyatt-Knowles, E., Hale, J.J., Mills, S.G., MacCoss, M., Swain, C.J., Harrison, T., Hill, R.G., Hefti, F., Scolnick, E.M., Cascieri, M.A., ...Rupniak, N. M. (1998). Distinct mechanism for antidepressant activity by blockade of central substance P receptors. *Science*, *281*(5383), 1640-1645. https://doi.org/10.1126/science.281.5383.1640

Matrisciano, F., Bonaccorso, S., Ricciardi, A., Scaccianoce, S., Panaccione, I., Wang, L., Ruberto, A., Tatarelli, R., Nicoletti, F., Girardi, P., & Shelton, R. C. (2009). Changes in BDNF serum levels in patients with major depression

disorder (MDD) after 6 months treatment with sertraline, escitalopram, or venlafaxine. *Journal of Psychiatric Research*, *43*(3), 247-254. https://doi.org/10.1016/j.jpsychires.2008.03.014

McIntyre, R. S., Rosenblat, J. D., Nemeroff, C. B., Sanacora, G., Murrough, J. W., Berk, M., Brietzke, E., Dodd, S., Gorwood, P., Ho, R., Iosifescu, D.V., Jaramillo, C.L., Kasper, S., Kratiuk, K., Lee, J.G., Lee, Y., Lui, L.M.W., Mansur, R.B., Papakostas, G., ...Stahl, S. (2021). Synthesizing the evidence for ketamine and esketamine in treatment-resistant depression: an international expert opinion on the available evidence and implementation. *American Journal of Psychiatry*, *178*(5), 383-399. https://doi.org/10.1176/appi.ajp.2020.20081251

Nemeroff, C. B. (1998). The neurobiology of depression. *Scientific American*, *278*(6), 42-49.

Nestler, E. J., Gould, E., & Manji, H. (2002). Preclinical models: status of basic research in depression. *Biological Psychiatry*, *52*(6), 503-528. https://doi.org/10.1016/S0006-3223(02)01405-1

Nestler, E. J., & Hyman, S. E. (2010). Animal models of neuropsychiatric disorders. *Nature Neuroscience*, *13(*10), 1161-1169. https://doi.org/10.1038/nn.2647

Ng, J., Rosenblat, J. D., Lui, L. M., Teopiz, K. M., Lee, Y., Lipsitz, O., Mansur, R.B., Rodrigues, N.B., Nasri, F., Gill, H., Cha, D.S., Subramaniapillai, M., Ho, R.C., Cao, B., & McIntyre, R. S. (2021). Efficacy of ketamine and

esketamine on functional outcomes in treatment-resistant depression: a systematic review. *Journal of Affective Disorders*, *293*, 285-294. https://doi.org/10.1016/j.jad.2021.06.032

Pizzagalli, D. A., & Roberts, A. C. (2022). Prefrontal cortex and depression. *Neuropsychopharmacology*, *47*(1), 225–246. https://doi.org/10.1038/s41386-021-01101-7

Rupniak, N. M. (2002). New insights into the antidepressant actions of substance P (NK1 receptor) antagonists. *Canadian Journal of Physiology and Pharmacology*, *80*(5), 489-494. https://doi.org/10.1139/y02-048

Scheepens, D. S., van Waarde, J. A., Lok, A., de Vries, G., Denys, D., & van Wingen, G. A. (2020). The Link Between Structural and Functional Brain Abnormalities in Depression: A Systematic Review of Multimodal Neuroimaging Studies. *Frontiers in Psychiatry*, *11*, 485. https://doi.org/10.3389/fpsyt.2020.00485

Schildkraut, J. J. (1965). The catecholamine hypothesis of affective disorders: a review of supporting evidence. *American journal of Psychiatry*, *122*(5), 509-522. https://doi.org/10.1176/ajp.122.5.509

Takamiya, A., Kishimoto, T., & Mimura, M. (2021). What Can We Tell About the Effect of Electroconvulsive Therapy on the Human Hippocampus?. *Clinical EEG and Neuroscience*. https://doi.org/10.1177/15500594211044066

Waters, R. P., Rivalan, M., Bangasser, D. A., Deussing, J. M.,

Ising, M., Wood, S. K., Holsboer, F., & Summers, C. H. (2015). Evidence for the role of corticotropin-releasing factor in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, *58*, 63-78. https://doi.org/10.1016/j.neubiorev.2015.07.011

Zhang, F. F., Peng, W., Sweeney, J. A., Jia, Z. Y., & Gong, Q. Y. (2018). Brain structure alterations in depression: Psychoradiological evidence. *CNS Neuroscience & Therapeutics*, *24*(11), 994-1003. https://doi.org/10.1111%2Fcns.12835

Zobel, A. W., Nickel, T., Künzel, H. E., Ackl, N., Sonntag, A., Ising, M., & Holsboer, F. (2000). Effects of the high-affinity corticotropin-releasing hormone receptor 1 antagonist R121919 in major depression: the first 20 patients treated. *Journal of Psychiatric Research*, *34*(3), 171-181. https://doi.org/10.1016/S0022-3956(00)00016-9

About the author

Dr Andrew Young
UNIVERSITY OF LEICESTER

Dr Andrew Young obtained a BSc degree in Zoology from the University of Nottingham, and his Ph.D in Pharmacology from the University of Birmingham. He then spent four years as a post doctoral researcher at Imperial College, London, studying glutamate release in the context of mechanisms of

epilepsy, before moving to the Institute of Psychiatry (King's College, London) for nine years to study dopamine signalling in models of schizophrenia and addiction. In 1997 he was appointed as Senior Research Fellow in the School of Psychology at University of Leicester and is now Associate Professor in that department. His research interests focus mainly on neurochemical function, particularly dopamine, in attention and motivation, and in models of schizophrenia and addiction. He teaches topics in biological psychology and the biological basis of mental disease to both undergraduate and postgraduate students in the School of Psychology and Biology.

17.

SCHIZOPHRENIA

Dr Andrew Young

Learning Objectives

- Know the main symptom clusters associated with schizophrenia, and be aware of the diagnostic criteria used
- Know the fundamentals of the three main biochemical theories of schizophrenia (dopamine, glutamate, serotonin)
- Appreciate how an understanding of the mechanism of action of drugs affecting behaviour, and those used in treatment of schizophrenia help us understand the biochemical changes occurring in the disease

- Be aware of the main classes of drugs used to treat schizophrenia, and their advantages and limitations.

Overview of schizophrenia

Schizophrenia is a severe and persistent mental disorder which causes profound changes in social, emotional and cognitive processes, with a major impact on daily lives. It has a prevalence of 0.3 to 0.7% worldwide, and is characterised by disturbances in thought processes, perception, behaviour and cognition. These symptoms normally emerge in late adolescence and early adulthood (17 to 30 years), with males generally showing earlier onset than females.

Although the term schizophrenia was only coined in the nineteenth century, there are descriptions of symptoms resembling schizophrenia from as early as ancient Egypt, Greece and China, and examination of the manifestations of people who were termed "possessed" or "mad" suggest that they probably suffered from schizophrenia. In the Middle Ages, mental illnesses were separated into four main categories: idiocy, dementia, melancholia and mania. At the time, mania was a general term describing a condition of insanity where the individual exhibited hallucinations, delusions and severe

behavioural disturbances, rather than the more specific diagnostic term used today (see Affective Disorders).

The German psychiatrist Emil Kraepelin (1856 – 1926) found these categories unhelpful in understanding the presentation, progression and outcome of mental diseases. In the 1890s, he put forward the idea of grouping together symptoms associated with similar outcomes, which he believed provided different manifestations of a single progressive disease, which he called *dementia praecox*, or early dementia, characterised by *dementia paranoides*, hebephrenia and catatonia. However, at the time, Kraepelin's views were not widely accepted, and indeed were ridiculed by many clinical professionals.

The widespread adoption of these ideas across the psychiatric community can be put down to the work of the Swiss psychiatrist Eugen Bleuler (1857 – 1939), published in 1911. He developed Kraepelin's diagnostic ideas, but he conceptualised the condition as a more psychological disorder, rather than the neuropathological disorder conceived by Kraepelin. He regarded hallucinations and delusions, the key features described by Kraepelin, as accessory symptoms, suggesting that the core (cardinal) symptoms related more to anhedonia and social withdrawal aspects of the condition (negative symptoms: see below), which were present in all cases. He also coined the term schizophrenia, as he believed the term *dementia praecox* was misleading. The name refers to a splitting of the mind, derived from the Ancient Greek,

schizo – split and *phren* – the mind. This refers to dissociated thinking and an inability to distinguish externally generated stimuli from internally generated thoughts: notably, he was not referring to dual personality, which is a completely different psychological phenomenon.

In the 1950s, another German psychiatrist, Kurt Schneider (1887 to 1967) asserted that hallucinations, thought disturbances and delusions (positive symptoms), which he termed 'first rank' symptoms, were the most relevant. The diagnostic principles laid down by Kraepelin, Bleuler and Schnieder form the basis of the systems of diagnosis used today, the Diagnostic and Statistical Manual of Mental Disorders (DSM: American Psychiatric Association) and the International Classification of Diseases (ICD: World Health Organisation).

Aetiology of schizophrenia

It is now clear that no one cause underlies schizophrenia, but that it is determined by the interaction between genetic, biological and social factors.

Genetic factors

Studies from the early twentieth century showed that relatives of people with schizophrenia were more likely to develop schizophrenia than the population as a whole, suggesting some

familial, genetic influence. More recently studies on fraternal (dizygotic) and identical (monozygotic) twins showed a substantially higher incidence of schizophrenia in twins whose co-twin suffered schizophrenia.

Twin studies

Concordance (percentage) in schizophrenia compares the incidence rate between two individuals. In dizygotic twins concordance is 15 to 25% – that is, if one twin has schizophrenia, there is a 15 to 25% chance that the other twin will also have it. In monozygotic twins, the concordance rate is 40 to 65%.

This compares to a concordance rate in the population as a whole of approximately 0.3 to 0.7%, indicating a clear hereditable, genetic component to vulnerability. However, given that dizygotic twins share 50% of their genes and monozygotic twins share 100%, the concordance rates are considerably lower that would be expected if schizophrenia were entirely genetically determined (50% and 100%

respectively). Adoption studies showed that the concordance rates in twins raised separately were similar to this, even when they were unaware that they were twins. Moreover, twins of healthy biological parents who were adopted by foster parents, of which one of the latter went on to develop schizophrenia, did not have a higher risk of themselves developing schizophrenia. Therefore, it is not social factors of upbringing that are influencing the concordance rates in twins, but rather the genetic influence.

Importantly, these data do not suggest that being a twin is a risk factor for developing schizophrenia: the concordance rate amongst twins is the same as in the population as a whole. Rather, the data show that if one twin has schizophrenia, the other twin has a higher than normal probability of also having it.

It is thought that genetic factors contribute approximately 50% of the vulnerability for schizophrenia, and molecular genetic approaches over the last three decades aimed to identify specific genes or groups of genes involved in this susceptibility. A number of candidate genes have been identified, although their precise involvement in the development of schizophrenia

is still uncertain. However, it is unlikely that any one single gene is responsible for the vulnerability, but rather a combination of genes across the genome. This may account, to some extent, for the variation on presentation of the disease across different sufferers, if the combination of 'vulnerability' genes is different in different individuals.

Biological factors

A number of biological risk factors have been suggested, including pregnancy and birth complications, maternal infection during pregnancy, and possibly infections and/or exposure to toxins during development. However, there is still considerable uncertainly as to the relative contribution of any of these, or how exactly they influence the progression of the disease.

Looking first at pregnancy complications, there is evidence that maternal infection, particularly during the second trimester of pregnancy, is correlated with a raised chance of developing schizophrenia. Children born in the spring (March and April in the northern hemisphere and September and October in the southern hemisphere), where the second trimester coincides with the winter months during which viral infections are at their peak, have a higher incidence than children born outside these months. In addition, there is a high incidence of the disease in children born shortly after major influenza epidemics: it remains to be seen what impact

the COVID-19 pandemic will have on the incidence of schizophrenia in the future. This effect may be mediated by pro-inflammatory cytokines, which have been shown to alter foetal neurodevelopment, particularly during the period of high proliferation and specialisation in the second trimester. Similarly, food shortage or malnutrition, particularly in early pregnancy, and maternal vitamin D deficiency during pregnancy are reported to increase the risk.

Birth trauma has also been identified as a risk factor. Premature labour and low birthweight are both associated with an increased risk, although both these may be a result of pregnancy complications rather than birth complications *per se*. However, asphyxiation during birth has also been identified as a risk factor, and there is a high incidence in babies born with forceps delivery: this could be due to the trauma of the use of forceps, or it could be the outcome of the delivery complication which necessitated the use of forceps.

Social factors

There has been considerable focus on whether childhood trauma, such as dysfunction of the family unit, neglect or sexual, physical or emotional abuse increases the probability of developing schizophrenia in the future. While such trauma undoubtedly increases the severity of a schizophrenic episode and the distress caused in individuals, and predicts a worse

long term outcome, it is debatable whether there is a causal connection between childhood trauma and schizophrenia.

The urban environment has also been suggested as a risk factor: an increased incidence of schizophrenia has been reported in people who grew up in urban surroundings suggesting that social conditions such as social crowding, social adversity, social isolation and poor housing may have an influence on the incidence of the condition. However, urban surroundings are often associated with poverty and poor diet, which may provide a more biological and less social account for the increased incidence. Similarly, people are more likely to have higher exposure to toxins (e.g. lead) in an urban environment than in a more rural environment. So, although social factors cannot be ruled out further studies, particularly using longitudinal designs, are required to identify specific relationships.

Precipitatory factors

People with genetic, biological and social vulnerabilities, even those in the high-risk group, do not necessarily go on to experience a schizophrenic episode. Rather, precipitatory factors are triggers which evoke schizophrenia in people who are at risk (Figure 10). The main trigger that has been identified is stress, often brought about by traumatic life events, such as bereavement, accident, break up of a relationship, unemployment, homelessness or abuse. Importantly these are

not sufficient to trigger a schizophrenic episode in themselves, but can trigger them in people who already have a vulnerability. It is possible also that premorbid changes occurring before the first psychotic episode (see below) may alter the individual's perception of traumatic events or their ability to deal with them, and so pre-diagnosis schizophrenia may exacerbate the impact of life-changing events, rather than the other way around.

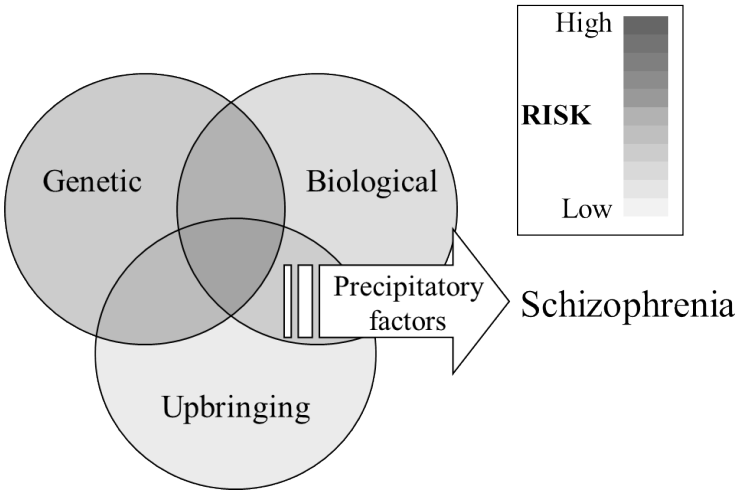

Fig 6.8. Interaction between risk factors and precipitatory factors in the origin of schizophrenia. Vulnerability to schizophrenia is determined by genetic, biological and social factors, with varying degrees of risk (depicted by the depth of shading): where multiple vulnerabilities are present the risk is increased. In vulnerable individuals, precipitatory factors, most commonly associated with stress from adverse life events, trigger a schizophrenic episode.

Neurodegeneration/ neurodevelopment

We have seen that the main vulnerabilities for schizophrenia are laid down very early in development during pregnancy, at birth and during early childhood, yet the outcome – a schizophrenic episode – normally occurs in early adulthood, some 15 to 20 years later. This suggests a neurodevelopmental aetiology, with abnormal development 'unmasked' by changes occurring in adulthood. The neurodevelopmental hypothesis is also supported by structural evidence from the brains of people with schizophrenia, where cortical volume is seen to be less than in control participants. Importantly this apparent loss of cortical tissue is not accompanied by significant increases in glial cells (which occurs with neurodegeneration), indicating that the reduction is not due to degeneration, supporting a neurodevelopmental explanation.

Brain development is a rapid and extremely complex process and is susceptible to damage from many sources. The main period of vulnerability starts in the second trimester of pregnancy, when neurogenesis and neuronal migration are at their peak, and continues through later pregnancy, birth and early childhood, when synaptogenesis occurs. Stress during this period, including inflammation, malnutrition or drugs has a major impact of foetal brain development, which can lead to developmental abnormalities, leading to a vulnerability to schizophrenia. However, we do not yet have a good

understanding of what these neurodevelopmental changes are, how they are triggered, how genetic, biological and social factors influence them, nor how they progress to leave the individual vulnerable to schizophrenia.

Although the wealth of evidence suggests a neurodevelopmental basis for the brain abnormalities in schizophrenia, there is also some evidence for neurodegeneration. Psychotic episodes appear to increase in severity over time, and the response to antipsychotic medication reduces over time, suggesting a progressive, neurodegeneration mechanism. It has been proposed that having a psychotic episode may be damaging to the brain, accounting for the increased likelihood and severity of subsequent episodes. If this is the case then it emphasises the importance of early intervention to prevent psychotic episodes developing. In this context, it is interesting that recent evidence suggests that both negative and cognitive symptoms may predate positive symptoms, and may act as a premorbid marker for people who are at risk. This then opens the possibility for psychological interventions before the onset of a psychotic episode.

The disconnection hypothesis of schizophrenia (Friston & Frith, 1995)

Many studies have shown both **structural** and **functional abnormalities** in the brains of schizophrenic patients.

Classical theories of schizophrenia suggest that impaired function is explained by pathological changes in specific localised brain areas, and that the type of dysfunction exhibited (i.e. the symptoms) depends on the particular areas damaged.

The **disconnection theory**, on the other hand, proposes a **dysregulation of connectivity** between regions in neural networks. Although areas may appear both structurally and functionally normal, their interactions within the neural networks controlling behaviour are abnormal, through a failure to establish a proper pattern of synaptic connection. This idea is also consistent with the neurogenerative processes occurring during the second trimester, where vulnerability to damage seems to be highest,

since this is the main time of neuronal migration, and marks the start of synaptogenesis.

Disconnection, therefore, causes a failure of appropriate functional integration between regions, rather than specific dysfunctions of the regions themselves. Thus the abnormality is expressed as an output from certain regions of the brain which are dependent on activity in other areas, and abnormal responses would only be seen when the specific activity involved interactions with other parts of the brain: so, for example, behaviours controlled in the frontal cortex are modulated by incorrect information coming from the temporal cortex. The important distinction is summed up as the distinction between 'the pathological interaction of two cortical areas and the otherwise normal interaction of two pathological areas' (Friston, 1998, p. 116).

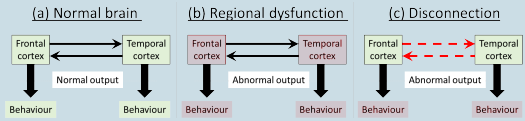

Figure 6.9. The Disconnection hypothesis. Comparing mechanisms controlling behavioural output in (a) the normal brain; (b) where behavioural dysfunction is accounted for by abnormalities in local brain regions; (c) where behavioural dysfunction is

accounted for by abnormal connectivity between normally functioning brain regions. Green indicates normal function; red indicated abnormal function.

Symptoms

The presentation of schizophrenia is very diverse, meaning that two individuals, both with a diagnosis of schizophrenia, may exhibit a very different spectrum of symptoms. Even within an individual, the symptoms may change over time, so that they may present differently from one month to the next.

Symptoms were originally divided into two groups: positive symptoms (type 1), characterised by exacerbation of normal

behaviour; and negative symptoms (type 2), characterised by a suppression of normal behaviour. However, more recently it became clear that there is actually a third cluster, cognitive symptoms, characterised by changes in cognitive executive function. Notably, negative and cognitive symptoms likely predate the onset of psychotic (positive) symptoms and are stable across the duration of the illness in most patients: they generally do not respond well to treatment and often persist after recovery from an acute psychotic episode.

Positive symptoms

Positive symptoms are symptoms that manifest as an enhancement or exaggeration of normal behaviour, where a patient loses touch with reality (psychosis). The most common symptoms in this domain are hallucinations, delusions, and abnormal and disorganised thoughts. These are essentially the symptoms referred to by Schneider as 'first order symptoms', and are often the first noticeable sign of illness.

Hallucinations (sensing things that are not there) are the most common symptom, and are experienced by around 75% of people living with schizophrenia. They are most commonly experienced in the auditory modality, where people 'hear voices', which may be people commenting on what they are doing and/or giving commands. Voices frequently appear angry or insulting and often demanding, although at times they can be neutral. People can also experience hallucinations

in visual, somatosensory or olfactory modalities, where they see, feel or smell (respectively) objects or people that are not actually present. Although not necessarily debilitating in terms of everyday life, experiencing hallucinations can be frightening and distressing for the individual.

Delusions are false beliefs that do not go away, even with evidence that they are not true. They are often associated with delusional perceptions (delusions of reference), where a normal perception takes on a specific, erroneous meaning. This may in turn reflect disturbances in salience attribution, that is, assigning inappropriate importance to unimportant stimuli. The most frequent type of delusions are **persecutory delusions** where the individual believes that people, including close friends or family, are working against them, commonly associated with large organisations such as government, MI5, or CIA. This leads to extreme mistrust of people, often the people who are trying to help them. However, patients also experience **grandiose delusions**, where they believe they are exceptionally talented or famous; **somatic delusions**, where they believe they are ill or deformed; and **delusions of control**, where they believe that their thoughts are being controlled by an outside force (thought insertion, thought removal, thought broadcasting). The specific form of delusion is often influenced by the person's own lifestyle, life events and social surroundings.

Disorganised thought, or formal thought disorder, is normally manifest as disorganised speech, where discourse is

fragmented and lacks logical progression. This often includes non-existent words (neologisms), and disorganised behaviour, where the individual exhibits unusual and unpredictable behaviour and may show inappropriate emotional responses. At its worst, narrative becomes a completely incoherent jumble of words and neologisms, sometimes referred to as 'word salad', reflecting severely disorganised thought processes and poverty of thought content.

There is strong evidence that positive symptomatology is related to temporal lobe dysfunction, and with abnormalities in dopamine function in the basal ganglia, particularly in the mesolimbic pathway, perhaps reflecting a dysregulation of glutamate-dopamine actions in the output from temporal cortex. Positive symptoms respond reasonably well to antipsychotic medication, which reduces dopamine function, emphasising the importance of dopamine systems in their generation.

Negative symptoms

Negative symptoms manifest as general social withdrawal, reduced affective responsiveness (emotional blunting), a lack of interest (apathy), desire (avolition) and motivation (abulia) and reduced pleasure (anhedonia). In its extreme this can lead to mutism (not speaking) and catatonia (immobility), often for extended periods of time. Negative symptoms may be present in the premorbid phase, before the first psychotic

episode, but may also emerge during or after a psychotic episode. The brain mechanisms underlying negative symptoms are currently not well understood, although frontal cortex abnormalities are implicated, and, as they do not respond well to treatment, they are a particularly debilitating component of the condition.

Cognitive symptoms

Cognitive symptoms encompass a number of difficulties with learning, memory, attention, planning and problem solving. Cognitive impairments occur in the majority of people with schizophrenia, if not all, and can be extremely severe and persistent. Indeed the degree of cognitive impairment contributes substantially to long term debilitation and is good predictor of outcome. Cognitive changes normally occur before the first psychotic episode, and may contribute to the individual's abnormal perceptions and attribution which subsequently manifests as positive symptoms. As with negative symptoms, cognitive symptoms probably originate from frontal cortex dysfunction, and do not generally respond well to antipsychotic treatment, emphasising a clinical need for better treatment strategies.

Diagnosis of schizophrenia

The diagnosis of schizophrenia primarily uses one of two diagnostic tools:

1. Diagnostic and Statistical Manual of the (DSM: American Psychiatric Association)
2. International Classification of Diseases (ICD: World Health Organization)

Both have had several iterations.

In the UK, ICD is the most widely used. Early versions of ICD and DSM exhibited a number of important differences in the conceptualisation of schizophrenia and diagnostic criteria, leading to different diagnoses between countries using ICD (e.g. UK) and DSM (e.g. USA). However, the most recent versions of each, ICD-11 (2019) and DSM-5 (2013) have very similar diagnostic criteria, improving consistency and applicability to clinical practice.

Diagnosis of schizophrenia (ICD 11)

At least two of the following symptoms must be present (by the individual's report or through observation by the clinician or other informants) most of the time for a period of 1 month or more. At least one of the qualifying symptoms should be from item a) through d) below:

a. Persistent delusions (e.g., grandiose, reference, persecutory).

b. Persistent hallucinations (most commonly auditory, but may be any sensory modality).

c. Disorganized thinking (formal thought disorder). When severe, the person's speech may be so incoherent as to be incomprehensible ('word salad').

d. Experiences of influence, passivity or control (i.e., the experience that one's feelings, impulses, actions or thoughts are not generated by oneself).

e. Negative symptoms (e.g. affective flattening, alogia, avolition, asociality, anhedonia).

f. Grossly disorganized behaviour that impedes goal-directed activity (e.g. bizarre, purposeless or unpredictable behaviour; inappropriate

emotional) Psychomotor disturbances such as catatonic restlessness, posturing, negativism, mutism, or stupor.

g. Symptoms are not a manifestation of another medical condition and are not due to effects of a substance or medication on the central nervous system, including withdrawal effects.

Source: World Health Organization (2019). International Statistical Classification of Diseases and Related Health Problems (11th ed.). https://icd.who.int/

Changes in brain structure in schizophrenia

Although many changes in brain structure have been reported in schizophrenia, no one signature abnormality has been identified. The most consistent changes that have been observed is a reduction in overall brain volume, particularly in grey matter, cortical thinning and increased ventricle size. However, these are also seen in normal aging and in other brain diseases, so are not unique to schizophrenia.

In particular, cortical thinning has been seen in the prefrontal cortex, an area important for logical thinking,

inference, problem solving and working memory, perhaps explaining the prevalence of disorganised thoughts and disrupted executive function and working memory in people with schizophrenia. This idea is supported by observations from functional imaging studies, showing reduced prefrontal cortex activity when people with schizophrenia perform cognitive tasks.

Medial temporal lobe structures, including entorhinal cortex and hippocampus, are also important in attention and working memory function, and imaging studies have also shown reduced tissue volume in these areas. In addition, both the temporal cortex and prefrontal cortex have substantial connectivity to the basal ganglia, including the nucleus accumbens (ventral striatum), an area critically involved in salience detection and response selection.

Thus, it is feasible that disruption of dopamine transmission in nucleus accumbens, perhaps under abnormal direction from temporal and frontal cortex inputs, may underlie salience attribution deficits (inappropriate attribution of importance to irrelevant stimuli) seen in schizophrenia. Notably, dopamine is a critical neurotransmitter in the nucleus accumbens, which may account for the efficacy of dopamine antagonists in treating these symptoms.

Therefore, although we do not yet know precisely what structural changes there are in the brains of people with schizophrenia, there is substantial evidence from recent

imaging studies, which is beginning to give us clues as to how subtle changes in brain connectivity may translate into the sorts of dysfunctions seen in schizophrenia.

Biochemical theories of schizophrenia

There is strong evidence that certain illicit drugs, including cocaine, amphetamine, phencyclidine and LSD, can cause acute and transient changes which resemble aspects of schizophrenia. In addition, these drugs exacerbate symptoms in people with schizophrenia and can trigger a relapse in people recovering from a previous episode. These drug effects give pointers to neurochemical changes which may underlie schizophrenia, and have formed the basis of biochemical theories. Indeed, they have also provided animal models for studying the mechanisms underlying schizophrenia and development of better drugs.

The dopamine theory posits that schizophrenia is caused by an increase in sub-cortical dopamine function, particularly in the mesolimbic dopamine pathway, projecting from the ventral tegmental are in the midbrain to limbic forebrain areas, primarily nucleus accumbens, but also hippocampus and amygdala. The dopamine theory is based on three main observations:

- first, that drugs which increase dopamine function, including amphetamine, cocaine and L-DOPA (used in

the treatment of Parkinson's disease), cause schizophrenia-like symptoms;

- second, that the first generation of drugs used in treating schizophrenia (typical antipsychotics) are dopamine receptor antagonists;
- and third, there is some evidence for perturbation of dopamine signalling in post-mortem schizophrenic brains, although these may derive from the body's adaptive changes in response to long-term treatment with dopaminergic drugs.

Whilst there is a great deal of evidential support for the dopamine theory, there are fundamental limitations which indicate that, although dopaminergic systems are involved, the dopamine theory cannot provide a complete explanation of schizophrenia.

Looking at this in more detail, dopaminomimetic drugs like amphetamine cause behavioural changes in normal individuals which resemble some aspects of schizophrenia. However, the changes are limited to behaviours resembling positive symptoms only, including hallucinations, delusions and thought disorder, but do not evoke changes resembling negative or cognitive symptoms. Therefore, although increasing dopamine function does evoke behavioural changes resembling schizophrenia, it does not cause the full spectrum of symptomatology, but only those resembling positive symptoms. Similarly, typical antipsychotic drugs which target

dopamine receptors alone are moderately effective at treating positive symptoms, but are very poor at treating negative or cognitive symptoms: indeed there is some evidence that these typical antipsychotic drugs may exacerbate negative and cognitive symptoms, possibly through actions in frontal cortex where dopamine signalling has been reported to be reduced in schizophrenia.

Dopamine signalling as the final common pathway

In its original iteration, the dopamine theory of schizophrenia posited a general over-activity of dopamine in schizophrenia. In a later refinement, only subcortical dopamine was thought to be overactive, with a dopamine underactivity in the prefrontal cortex, accounting for the observation that dopamine receptor antagonists (typical antipsychotic medication) exacerbate negative and cognitive symptoms. However, even this conceptualisation has shortcomings, and does not

explain how risk factors translate into the symptoms and time course of schizophrenia.

In their reconceptualisation of the dopamine hypothesis, which they name 'version III: the final common pathway', Oliver Howes and Shitij Kapur (2009) bring together recent data from genetics, molecular biology and imaging studies to provide a framework to account for these anomalies.

Molecular imaging studies show increases in activity in the dopamine neurones in schizophrenia, implying that the abnormality lies in the input to dopaminergic neurones rather than the output from them. Notably, dysfunction in both frontal and temporal cortex has been shown to increase mesolimbic dopamine release, suggesting that core abnormalities in these areas can modify critical dopamine function.

Thus, abnormal function in multiple inputs leads to dopamine dysregulation as the final common pathway: the different behavioural manifestations seen in schizophrenia may be due to the actual combination of dysfunctional inputs to the dopaminergic system in each individual. Moreover, within this framework, the underlying damage could

be in the brain areas sending projections to the dopamine neurones, or in the connections themselves (see **Disconnection hypothesis**).

Notably, mesolimbic dopamine systems are implicated in salience attribution, therefore dysregulation of activity in this pathway would result in abnormal salience attribution which may underlie positive symptomatology.

Therefore an important goal for future drug development is to target the mechanisms converging on the dopamine systems, which are abnormal in schizophrenia, rather than on dopamine systems themselves, which are the target of current antipsychotics. This in turn relies on a fuller understanding of what systems are involved.

Glutamate is the most prevalent excitatory neurotransmitter in the mammalian brain, which acts at several different receptor types. Non-competitive antagonists at one of these receptor types, the NMDA receptor (e.g. phencyclidine, ketamine and dizocilpine (MK-801)), cause behavioural changes in normal people which resemble schizophrenia. In addition, when given to schizophrenia sufferers, they exacerbate the symptoms, providing evidence that the drug action mimics the disease state. Therefore this implies that a glutamate underactivity,

particularly at NMDA receptors may underlie schizophrenia. Importantly, unlike dopaminergic drugs which provoke behaviours resembling positive symptoms only, NMDA-receptor antagonists generate behavioural changes which resemble symptoms in all three domains – positive, negative and cognitive, implying that glutamate dysregulation is the core deficit in schizophrenia, and that dopamine abnormalities are downstream of this core deficit. There is also a body of evidence showing changes glutamate function in brains of people with schizophrenia, including reduced levels of glutamate and increased cortical glutamate binding in post mortem brains, and increased glutamate receptor density in living brains.

Other transmitters which have been implicated in schizophrenia are serotonin and gamma-aminobutyric acid (GABA). Lysergic acid diethylamide (LSD), an agonist at serotonin receptors, is an illicit drug taken recreationally, and causes reality distortions and hallucinations resembling positive symptoms of schizophrenia, implicating serotonin over activity in schizophrenia. This is consistent with the pharmacological action of atypical (second generation) antipsychotic drugs, many of which are 5HT-2 receptor antagonists. However there is little or no evidence for abnormalities in serotonin function in the brains of people with schizophrenia. Cortical GABA signalling has also been shown to be dysfunctional in the brains of people with

schizophrenia, but it is not clear how this impacts on cortical function leading to schizophrenia symptoms.

Treatment

Social and clinical outcome

Without pharmacological intervention, around 20% of people with schizophrenia recover well, although it is likely that they never actually show full recovery, hence the term 'near full recovery' is often used.

With pharmacological intervention, this figure rises to around 50% showing near full recovery and able to live independently or with family. A further 25% show moderate recovery, but still require substantial support: these generally live in supervised housing, nursing homes or hospitals. The remainder show little or no improvement.

In particular, negative and cognitive symptoms do

not respond well to treatment, and often form the most debilitating long-term dysfunctions.

(Data from Torrey, 2001)

Until relatively recently, there were no effective treatments for schizophrenia. In the nineteenth and early twentieth centuries, sufferers were usually installed in asylums, with little or no form of treatment offered, and little or no communication with the outside world. Where treatments were offered, these included **shock treatment** (insulin shock, pentylenetetrazol [*Metrazol*] shock, and electroconvulsive shock [ECT]) and even **frontal lobotomy** (a severing of the neurones connecting the frontal lobes to the remainder of the brain), both of which were severely debilitating, and had limited efficacy in treating the disease. In this situation, patients rarely showed any sort of recovery: indeed their condition often worsened during confinement.

Typical antipsychotic drugs

The drug **chlorpromazine** is a powerful tranquilliser, used in managing recovery after surgical anaesthetic. People who took it reported a feeling of well-being and calm. On this basis, during the 1950s, it was tried on people with schizophrenia, who often exhibited extreme agitation. It was found to

alleviate some of the symptoms of schizophrenia, notably the hallucinations, delusions and disorganised thought – all symptoms within the positive symptom domain – even at a much lower dose than that required for tranquilliser action.

Pharmacologically, chlorpromazine is a dopamine receptor antagonist, with some selectivity for D2-like receptors (D2, D3, D4) compared to D1-like (D1, D5), although at the time nobody knew what the pharmacology of the drug was: indeed it was not for another decade that dopamine was realised to be a neurotransmitter. It was also not effective in all patients, or against all symptoms, and was associated with some debilitating side effects. Nevertheless, at the time (mid 1950s), it formed a major breakthrough as the first pharmacological treatment for schizophrenia.

Following the discovery of the antipsychotic effect of chlorpromazine, many other dopamine D2-like receptor antagonists were tested as potential antipsychotic drugs. This led to the development of a whole class of antipsychotic drugs: the typical or first-generation antipsychotics. Of these, haloperidol is now the typical antipsychotic of choice, although there are several other typical antipsychotic drugs also licenced for use in UK (e.g. flupentixol, pimozide, sulpiride), which became the mainstay of pharmacological treatment for schizophrenia during the 1970s and 1980s. Originally these drugs were called neuroleptics, as they induced neurolepsis (immobility associated with their major

tranquilliser action). Now, they are called antipsychotics, reflecting their reduction of psychotic symptoms at doses much lower than those used to induce neurolepsis. Their antipsychotic efficacy is a direct result of their antagonist action at dopamine D2 receptors.

However, treatment with typical antipsychotic drugs has a number of drawbacks. Firstly they are not very effective: around 25% of patients fail to respond to treatment at all, and others (around 25%) show some improvement, but still show substantial symptomatology. In particular, typical antipsychotic drugs show little or no efficacy at treating negative or cognitive symptoms; they are mainly effective only on positive symptoms. Therefore, while treatment may alleviate positive symptoms, sufferers are left with residual and potentially severely debilitating negative and cognitive symptoms.

Another main drawback of typical antipsychotic drugs is that they produce sedative and motor side effects in the majority of patients. The most debilitating of these are the motor side effects, including resting tremor and akathisia (similar to those seen in Parkinson's disease), and tardive dyskinesia: each occurs in around 25% of people taking typical antipsychotic medication. These are caused by D2 receptor antagonism in the dorsal striatum (caudate nucleus and putamen) resembling the dopamine depletion seen in these areas in Parkinson's disease. Notably, the parkinsonian side effects recover on withdrawal of the drugs, but tardive

dyskinesia does not and motor function will progressively deteriorate irreversibly if the medication is continued. Finally, the antipsychotic effect of these drugs is not immediate, but takes several weeks to establish, creating a substantial delay between initiation of treatment and control of symptoms.

Atypical antipsychotic drugs

In the search for antidepressant drugs similar to the tricyclic antidepressant, imipramine, several drugs were discovered which had antipsychotic properties: one of these was clozapine. In sharp contrast to other antipsychotics used at the time, clozapine had good antipsychotic potency, but with minimal motor side effects and for this reason it was called an atypical antipsychotic (also known as second generation antipsychotic). Subsequently it was found that, as well as positive symptoms, it is at least somewhat effective at treating negative and cognitive symptoms, and it is effective in some people who do not respond to other antipsychotic drugs. Pharmacologically, too, it is rather different from typical antipsychotics, which are D2 receptor antagonists: clozapine has a wide ranging pharmacology with effects at dopamine, serotonin, acetylcholine, noradrenaline and histamine receptors. Clozapine was introduced as an antipsychotic medication in the early 1970s, but was withdrawn a few years later after a Finnish study reported a high incidence of severe, and potentially fatal blood disorders, agranulocytosis and

leucopoenia. However, after extensive studies, it was concluded that the occurrence of agranulocytosis (1%) and neutropenia (3%) in patients taking clozapine is relatively low, particularly beyond 18 weeks after the start of treatment, and it was reintroduced into the market in the 1990s, with strict monitoring controls in place. Thus, in the UK, patients need to have blood tests every week for the first 18 weeks of treatment, then fortnightly up to the end of the first year of treatment and every four weeks thereafter. If there is any sign of agranulocytosis or leukopenia, the drug has to be withdrawn permanently. This monitoring adds substantially both to the patient inconvenience and financial cost, and therefore, although clozapine is still the most effective antipsychotic available, it is only used in cases where other medications have not worked.

The discovery of the effectiveness of clozapine initiated a new approach to developing novel antipsychotic drugs. Rather than focussing on D2 receptor antagonists, drugs with much wider pharmacology were tested. Several more atypical antipsychotics derived from this approach, including olanzapine, currently the first line treatment, quetiapine, risperidone and lusaridone. Although they mostly have a range of pharmacological effects, the common action of these drugs and clozapine, is potent antagonist effects at both D2 and 5HT2 receptors: this dual action is believed to underlie the antipsychotic actions.

Although these drugs are not much more effective at

treating positive symptoms – even clozapine is only effective in around 85% of patients – they do have some limited efficacy at treating negative and cognitive symptoms, and they cause little or no motor side effects. However, there use is still limited by other side effects, including substantial weight gain and excessive salivation. In addition, their effects can be quite variable, and are normally significantly slower in onset that typical antipsychotics.

Third and fourth generation antipsychotics

Third generation antipsychotics, for example aripiprazole, brexpiprazole and cariprazine, are D2 receptor partial agonists, rather than full antagonists, which means that where endogenous dopamine levels are high, the drugs reduce its effect, but when they are low the drugs enhance its effect. They also have actions on second messenger pathways to modulate the actions on D2 receptors. Therefore they have a dopamine 'stabilising' effect. Some of them also have 5HT partial agonist actions. They are generally as effective as other antipsychotics, but with reduced side effects and are better tolerated. However, they are still not very effective at treating negative and cognitive symptoms. Adequate control of negative and cognitive symptoms, which are arguably the most pervasive and disruptive symptoms of schizophrenia is, at present, an unmet clinical need, and several alternative therapeutic

approaches are at the experimental stage, either in preclinical testing or in clinical trials, aiming to target actions beyond dopamine and serotonin receptors. Among these are drugs which modulate glutamate function, drugs acting on acetylcholine systems and drugs targeting a group of regulatory compounds called trace amines.

Psychological therapy

There are a number of psychological therapies available for treating schizophrenia, of which the most important are cognitive behavioural therapy (CBT) and family therapy. Although these therapies are not effective in all people or situations, they are showing great promise for future refinement. CBT primarily focusses on helping the individual understand their abnormal perceptions and work to overcome them, while family therapy involves working with the patient and their family to achieve a less stressful and more supportive environment.

Psychological therapy is often not effective during an acute psychotic episode, as presence of psychotic symptomatology make meaningful communication difficult and also make patients suspicious of caregivers. The most success has been achieved with people who have been stabilised pharmacologically first, where psychological therapy has been successful to maintain stability allowing reduction or even cessation of drugs. Interestingly, also, there has been some

success in using psychological therapies in individuals who have shown a vulnerability, or have shown evidence of existing negative or cognitive symptoms, but have not yet experienced a full psychotic episode. In this case therapy looks at adverse life events and the individuals reactions to them: this approach has shown some success in preventing the development of a psychotic episode. Given the evidence that psychotic episodes may in themselves cause damage, this is valuable in managing vulnerable patients, and emphasises the importance of being able to identify vulnerable individuals in the premorbid stage.

Current treatments

The current first line treatment is generally an atypical antipsychotic drug, normally olanzapine, alongside individual CBT and family therapy, although acutely symptomatic patients rarely respond well to psychological therapy: patients require stabilisation pharmacologically before psychological therapy becomes effective. If the first drug is not effective at controlling symptoms, or has unacceptable side effects, a second drug would be tried, normally another atypical antipsychotic drug, but for some patients a typical antipsychotic is more appropriate. Clozapine is only considered after two other antipsychotics have been tried, one of which must be an atypical drug.

In the post-acute period, following a schizophrenic episode, both pharmacological and psychological therapies are

generally continued in order to prevent relapse, although it is sometimes possible to slowly reduce drugs, with careful monitoring to guard against relapse, particularly with effective psychological therapy. However, in the post-acute phase, many patients choose not to take the drugs, believing that they are cured, or even choosing the risk of relapse rather than the side effects of the drugs. This alone is estimated to account for a relapse rate of around 20% of patients. In some cases, where adherence to oral preparations is unreliable, it is beneficial to give patients slow-release 'depot' preparation, known as long acting injectable drugs, or LAIs. Mostly these are typical antipsychotics, haloperidol, flupentixol or fluphenazine, but LAI preparations of atypical antipsychotics, including olanzapine, risperidone and aripiprazole are now available for clinical use.

Key Takeaways

- Schizophrenia occurs in approximately 0.5% of the population, with peak onset in early adulthood. It is characterised by a variety of symptoms, which cluster into three types:

positive (psychotic), negative and cognitive. Although positive symptoms are the most noticeable, and indeed it is usually the emergence of positive symptoms that alerts people to the problem, negative and cognitive symptoms may occur before a psychotic episode, and often endure long after recovery from a psychotic episode, causing substantial long-term debilitation. Vulnerability to schizophrenia depends on genetic, biological and social factors, which influence neurodevelopment, although little is known about the precise mechanisms. A psychotic episode is triggered in a vulnerable individual by precipitatory factors, the most prominent of which seems to be stress, particularly from adverse life events.

- Biochemical theories posit critical roles for glutamate and dopamine in the pathology of schizophrenia, although other transmitters, notably serotonin and GABA have also been implicated. It is thought that the primary deficit may lie in abnormal cortical glutamate function, supported by physiological and imaging studies showing decrease cortical

volume, and changes in markers of cortical glutamate function in schizophrenic brains. Negative and cognitive symptoms may be a result of abnormalities in frontal and/or temporal cortices, or in the communication between them, while dysregulated glutamate-dopamine signalling, particularly in the basal ganglia, may account for positive symptomatology.

- Current treatments rely heavily on drugs which act as antagonists at dopamine and serotonin receptors, the typical and atypical antipsychotics. They are reasonably effective at treating positive symptoms, perhaps reflecting the critical dopaminergic element in the expression of positive symptoms, but have little or no effect on negative or cognitive symptoms: they are also not effective in around 25% of sufferers, and cause unpleasant and debilitating side effects. Therefore there is a real clinical need for drugs which offer better control of symptoms in all three domains, with fewer side effects.

References and further reading

Bowie, C. R., & Harvey, P. D. (2006). Cognitive deficits and functional outcome in schizophrenia. *Neuropsychiatric disease and treatment*, *2*(4), 531–536. https://doi.org/10.2147/nedt.2006.2.4.531

Chen, Z., Fan, L., Wang, H., Yu, J., Lu, D., Qi, J., ... & Wang, S. (2022). Structure-based design of a novel third-generation antipsychotic drug lead with potential antidepressant properties. *Nature Neuroscience*, 25(1), 39-49. https://doi.org/10.1038/s41593-021-00971-w

Egerton, A., Modinos, G., Ferrera, D., & McGuire, P. (2017). Neuroimaging studies of GABA in schizophrenia: A systematic review with meta-analysis. *Translational psychiatry*, *7*(6), e1147. https://doi.org/10.1038/tp.2017.124

Ellenbroek, B. A. (2012). Psychopharmacological treatment of schizophrenia: What do we have, and what could we get? *Neuropharmacology*, *62*(3), 1371-1380. https://doi.org/10.1016/j.neuropharm.2011.03.013

Friston, K. J. (1998). The disconnection hypothesis. *Schizophrenia Research*, *30*(2), 115-125. https://doi.org/10.1016/S0920-9964(97)00140-0

Friston, K. J., & Frith, C. D. (1995). Schizophrenia: A disconnection syndrome? *Clinical Neuroscience, 3*(2), 89-97

Henriksen, M. G., Nordgaard, J., & Jansson, L. B. (2017).

Genetics of schizophrenia: Overview of methods, findings and limitations. *Frontiers in Human Neuroscience*, *11*, 322. https://doi.org/10.3389%2Ffnhum.2017.00322

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III–the final common pathway. *Schizophrenia Bulletin*, *35*(3), 549-562. https://doi.org/10.1093/schbul/sbp006

Jauhar, S., Johnstone, M., & McKenna, P. J. (2022). Schizophrenia. *The Lancet*, *399*(10323), 473–486. https://doi.org/10.1016/S0140-6736(21)01730-X

McCutcheon, R. A., Abi-Dargham, A., & Howes, O. D. (2019). Schizophrenia, dopamine and the striatum: From biology to symptoms. *Trends in Neurosciences*, *42*(3), 205-220. https://doi.org/10.1016/j.tins.2018.12.004

McKenna, P. J. (2013). *Schizophrenia and related syndromes*. Routledge.

Morgan, C., & Fisher, H. (2007). Environment and schizophrenia: Environmental factors in schizophrenia: Childhood trauma – a critical review. *Schizophrenia bulletin*, *33*(1), 3–10. https://doi.org/10.1093/schbul/sbl053

Orsolini, L., De Berardis, D., & Volpe, U.(2020) Up-to-date expert opinion on the safety of recently developed antipsychotics, *Expert Opinion on Drug Safety*, *19*(8), 981-998, https://doi.org/10.1080/14740338.2020.1795126

Seeman, P. (2013). Schizophrenia and dopamine receptors.

European Neuropsychopharmacology, *23*(9), 999-1009. https://doi.org/10.1016/j.euroneuro.2013.06.005

Tandon, R., Nasrallah, H. A., & Keshavan, M. S. (2009). Schizophrenia, "just the facts" 4. Clinical features and conceptualization. *Schizophrenia Research*, 110(1), 1-23. https://doi.org/10.1016/j.schres.2009.03.005

Torrey, E.F. (2001) Surviving Schizophrenia: A Manual for Families, Consumers, and Providers (4th Edition); HarperCollins.

World Health Organization (2019). *International Statistical Classification of Diseases and Related Health Problems* (11th ed.). https://icd.who.int/

About the author

Dr Andrew Young
UNIVERSITY OF LEICESTER

Dr Andrew Young obtained a BSc degree in Zoology from the University of Nottingham, and his Ph.D in Pharmacology from the University of Birmingham. He then spent four years as a post doctoral researcher at Imperial College, London, studying glutamate release in the context of mechanisms of epilepsy, before moving to the Institute of Psychiatry (King's College, London) for nine years to study dopamine signalling in models of schizophrenia and addiction. In 1997 he was appointed as Senior Research Fellow in the School of

Psychology at University of Leicester and is now Associate Professor in that department. His research interests focus mainly on neurochemical function, particularly dopamine, in attention and motivation, and in models of schizophrenia and addiction. He teaches topics in biological psychology and the biological basis of mental disease to both undergraduate and postgraduate students in the School of Psychology and Biology.

18.

AGEING: A BIOLOGICAL AND PSYCHOLOGICAL PERSPECTIVE

Professor Claire Gibson and Professor Harriet Allen

Learning Objectives

- To gain knowledge and understanding of the biological and cognitive changes that occur with ageing
- To understand methodological approaches to studying effects of ageing and strategies which may exist to promote healthy (cognitive ageing)

- To understand the subtle sensory changes which may occur with ageing.

Ageing can be defined as a gradual and continuous process of changes which are natural, inevitable and begin in early adulthood. Globally, the population is ageing, resulting in increasing numbers and proportion of people aged over 60 years. This is largely due to increased life expectancy and has ramifications in terms of health, social and political issues. The ageing process results in both physical and mental changes and although there may be some individual variability in the exact timing of such changes they are expected and unavoidable. Whilst ageing is primarily influenced by a genetic process it can also be impacted by various external factors including diet, exercise, stress, and smoking. As humans age their risk of developing certain disorders, such as dementia, increases, but these are not an inevitable consequence of ageing.

Healthy ageing is used to describe the avoidance or reduction of the undesired effects of ageing. Thus, its goals are to promote physical and mental health. For humans we can define our age by chronological age – how many years old a person is – and biological age, which refers to how old a person seems in terms of physiological function/presence of

disease. All of the systems described elsewhere in this textbook will undergo changes with ageing and here we have focused on the main cognitive and sensory ones. Gerontology is the study of the processes of ageing and individuals across the life span. It encompasses study of the social, cultural, psychological, cognitive and biological aspects of ageing using distinct study designs, as described in Table 1, and specific methodological considerations (see insert box).

Table 6.1. Advantages and disadvantages of different methodological approaches to study ageing

Study type	Description	Advantages	Disadvantages
Longitudinal	Data is collected from the same participants repeatedly at different points over time.	Easier to control for cohort effects as only one group involved Requires (relatively) fewer participants	Participant dropout rates increase over time. Resource intensive to conduct studies over long periods of time. Practice effects
Cross-sectional	Data is collected at a single time point for more than cohort. Cohorts are separated into age groups.	Efficient – all data collection completed within a relatively short time frame, studies can be easily replicated	Difficult to match age groups. Differences due to cohort/ historical differences in environment, economy etc
Sequential longitudinal/ Sequential cross sectional	Two or more longitudinal, or cross sectional designs, separated by time	Repeating or replicating helps separate cohort effects from age effects.	Complex to plan. Can be expensive

Accelerated longitudinal	A wide age range is recruited, split into groups and each group is followed for a few years.	Longitudinal data is collected from the same participants over time. Cross-sectional data collection occurs within shorter time frame.	Does not completely avoid cohort effects.

Credit: Claire Gibson

Methodological considerations for ageing studies

There are a number of issues that are important to consider when studying ageing. Many of these occur because it is difficult, or impossible, to separate out the effects of ageing from the effects of living longer or being born at a different time;

- Older people will have experienced more life events. This will be true even if the chances of

experiencing something are the same throughout life and not related to age. This means that older people are more likely to have experienced accidents, recovered from disease, or have an undiagnosed condition.

- People of different ages have lived at different times, and thus experienced different social, economic and public health factors. For example, rationing drastically changed the health of people who grew up in the middle of the twentieth century. It is likely that the Coronavirus pandemic will also have both direct and indirect effects in the longer term.
- The older people get, the more variable their paths through life become. It is often found that there is more variability in data from older people, however we also know that environmental factors have a strong effect on behavioural data.
- Non-psychological effects can directly impact on psychology. A good example of this is attention. Often as we age our range of movement becomes limited or slower. Since attention is often shifted by moving the head or eyes, reduction in the range of speed of

motion will directly impact shifts of attention. Of course, not all attention shifts involve overt head or eye movements.

- Generalised effects such as slowing need to be controlled for or considered before proposing more complex or subtle effects. One way to do this is to use analysis techniques such as z-scores, ratios or Brindley plots to compare age groups and to ensure there is always a within age group baseline or control condition.

Biological basis of ageing

Ageing *per se* involves numerous physical, biochemical, vascular, and psychological changes which can be clearly identified in the brain. As we age our brains shrink in volume (Figure 6.9), with the frontal cortex area being the most affected, followed by the striatum, and including, to a lesser extent, areas such as the temporal lobe, cerebellar hemispheres, and hippocampus. Such shrinkage can affect both grey and white matter tissue, with some studies suggesting there may be differences between the sexes in terms of which brain areas show the highest percentage of shrinking with ageing.

Magnetic Resonance Imaging (MRI) studies allow the study of specific areas of brain atrophy with age and show that decreases in both grey and white matter occur, albeit at different stages of the lifespan (for further information see Farokhian et al., 2018). Shrinkage in grey and white matter results in expansion of the brain's ventricles in which cerebrospinal fluid circulates (Figure 6.10). The cerebral cortex also thins as we age and follows a similar pattern to that of brain volume loss in that it is more pronounced in the frontal lobes and parts of the temporal lobes.

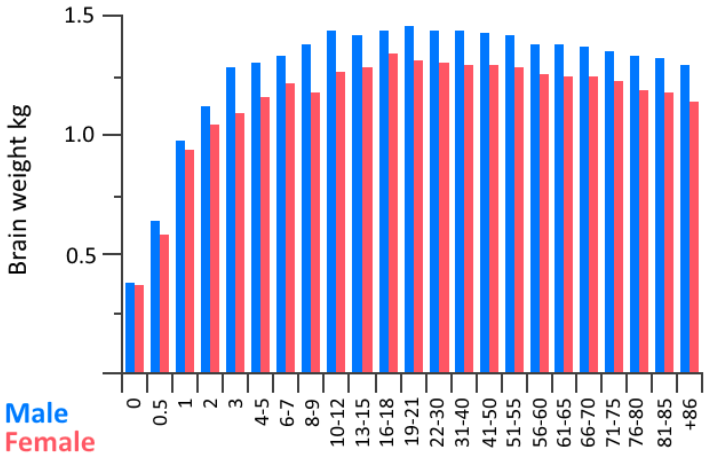

Fig 6.10. Graph showing (male and female) brain weight changes with age

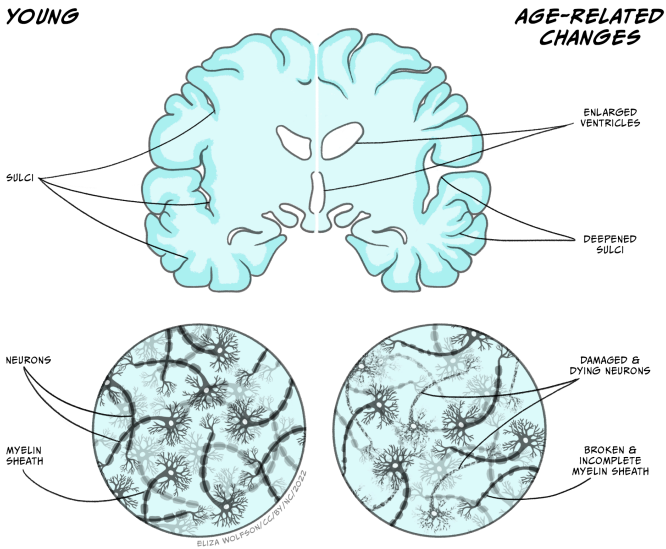

Fig 6.11. Neuroanatomical changes with ageing

Brain plasticity and changes to neural circuits

During normal ageing humans, and other animals, experience cognitive decline even in the absence of disease, as explained below. Some of this cognitive decline may be attributable to decreased, or at least disrupted, neuroplasticity. Neuroplasticity refers to the brain's ability to adapt and modify its structure and functions in response to stimuli – it is an important process during development and thought to underlie learning and memory. In the young, the potential for brain plasticity is high as they undergo rapid learning and

mapping of their environment. As we age, this capacity for learning, and therefore plasticity, declines, although it may be argued that we can retain some capacity for learning and plasticity through practice and training-based techniques (see insert box on 'Strategies to promote healthy cognitive ageing'). Synaptic changes, thought to be a major contributor to age-related cognitive decline, involve dendritic alterations in that dendrites shrink, their branching becomes less complex, and they lose dendritic spines. All in all, this reduces the surface area of dendrites available to make synaptic connections with other neurons and therefore reduces the effectiveness, and plasticity, of neural circuitry and associated cognitive behaviours (for review see Petralia et al., 2014).

Cellular and physiological changes

Certain pathological features, in particular the occurrence of beta amyloid (Ab) plaques (sometimes referred to as senile plaques) and neurofibrillary tangles, are typically associated with dementia-causing diseases such as Alzheimer's. Such features are described in detail in the **Dementias** chapter. However, it is important to note that they also occur with ageing, albeit in smaller amounts, and are more diffusely located compared to disease-pathology, but may also contribute to cell death and disruption in neuronal function seen in ageing.

Other physiological changes which occur with ageing, all of

which have been suggested to result in cognitive impairment, include **oxidative stress, inflammatory reactions** and **changes in the cerebral microvasculature**.

Oxidative stress is the damage caused to cells by free radicals that are released during normal metabolic processes. However, compared to other tissues in the body, the brain is particularly sensitive to oxidative stress, which causes DNA damage and inhibits DNA repair processes. Such damage accumulates over the lifespan resulting in cellular dysfunction and death. Ageing is associated with a persistent level of systemic inflammation – this is characterised by increased concentration in the blood of pro-inflammatory cytokines and other chemokines which play a role in producing an inflammatory state, along with increased activation of microglia and macrophages. **Microglia** are the brain's resident immune cells and are typically quiescent until activated by a foreign antigen. Upon activation, they produce pro-inflammatory cytokines to combat the infection, followed by anti-inflammatory cytokines to restore homeostasis. During ageing, there appears to be a chronic activation of microglia, inducing a constant state of neuroinflammation, which has been shown to be detrimental to cognitive function (Bettio et al., 2017; Di Benedetto et al., 2017). Finally, the ageing brain is commonly associated with decreased microvascular density, vessel thickening, increased vessel stiffness and increased vessel tortuosity (or distortion, twisting) which all result in compromised cerebral blood flow.

Any such disruption to cerebral blood flow is likely to result in changes in cognitive function (e.g. see Ogoh, 2017).

Neurotransmitter changes

During ageing the brain also experiences changes in the levels of neurotransmitters and their receptors in different regions of the brain largely, but not exclusively, involving the dopamine, serotonin and glutamate systems. Neurochemical changes associated with ageing are important to understand as they may be relevant when considering therapeutic targets aimed at stabilising or enhancing those brain functions which typically deteriorate with age.

- **Dopamine** is a monoamine neurotransmitter which plays a neuromodulatory role in many CNS functions including executive function, motor control, motivation, arousal, reinforcement and reward. During ageing, dopamine levels have been reported to decline by about 10% per decade from early adulthood onwards and have been associated with declining motor and cognitive performance (Berry et al., 2016; Karrer et al., 2017). It may be that reduced levels of dopamine are caused by reductions in dopamine production, dopamine producing neurons and/or dopamine responsive synapses.
- **Serotonin**, also known as 5-hydroxytryptamine (5-HT)

functions as both an inhibitory neurotransmitter and a hormone. It helps regulate mood, behaviour, sleep and memory all of which decline with age. Decreasing levels of different serotonin receptors and the serotonin transporter (5-HTT) have also been reported to occur with age. Areas particularly affected by loss of serotonin neurones include the frontal cortex, thalamus, midbrain, putamen and hippocampus (Wong et al., 1984).

- **Glutamate** is the primary excitatory neurotransmitter in the CNS synthesised by both neuronal and glial cells and high levels of glutamate, causing neurotoxicity, are implicated in a number of neurodegenerative disorders including multiple sclerosis, amyotrophic lateral sclerosis, Alzheimer's disease and schizophrenia. Neurotransmission at glutamatergic synapses underlies a number of functions, such as motor behaviour, memory and emotion, which are affected during ageing. Levels of glutamate are reported to decline with ageing – older age participants have lower glutamate concentrations in the motor cortex along with the parietal grey matter, basal ganglia and the frontal white matter (Sailasuta et al., 2008).

Ageing and genetics

The genetics of human ageing are complex, multifaceted and are based on the assumption that duration of lifespan is, at least

in part, genetically determined. This is supported by evidence that close family members of a centenarian tend to live longer and genetically identical twins have more similar lifespans than non-identical twins. The genetic theory of ageing is based on telomeres which are repeated segments of DNA (deoxyribonucleic acid) which are present at the ends of our chromosomes. The number of repeats in a telomere determines the maximum life span of a cell, since each time a cell divides, multiple repeats are lost. Once telomeres have been reduced to a certain size, the cell reaches a crisis point and is prevented from dividing further; thus the cell dies and cannot be replaced. Many believe this is an oversimplified explanation of the genetics of ageing and that actually a number of genetic factors, in combination with environmental factors, contribute to the ageing process (see reviews for further information – Melzer et al., 2020; Rodriguez-Rodoro et al., 2011).

Changes in cognitive systems

The changes in the biological systems and processes above translate into changes across all psychological processes. Some changes are found across all systems, but, as the sections below describe, changes with age are not uniform and there are specific changes in cognitive systems such as memory and attentional control and inhibition.

One of the strongest findings in ageing research is that there

is a general slowing of processes (Salthouse, 1996). In virtually every task, older people respond more slowly, on average, than younger people, reflecting many of the biological changes described above. This is such a ubiquitous finding that it is important to control or account for slowing before considering any further theories of cognitive change. Furthermore, slowing can have more subtle effects than simple changes in reaction times. If a cognitive process requires more than one step, a delay in processing the first step can mean that the entire processing stream cannot proceed, or information cannot synchronise between different sub-processes (see 'Methodological considerations' text box).

Memory

Ageing appears to have greater effects in some types of memory more than others. Memory for personally experienced events (i.e. **episodic memory**) undergoes the clearest decline with age (Ronnlund et al., 2005). Within this, the decline after age 60 seems to be greater for recall tasks that require the participant to freely recall items, compared to tasks where they are asked to recognise whether the items were seen before (La Voie & Light, 1994). In a meta-analysis of studies where participants were asked whether they remembered the context or detail of a remembered event, Koen and Yonelinas (2014) found that there was a significant difference in participants' recollection

of context, but much less reduction in the ability to judge the familiarity of a prior occurrence.

Changes in other types of memory are less clear. **Semantic memory**, that is the memory for facts and information, shows less decline (Nyberg, L. et al., 2003), as does **procedural memory** and **short-term memory** (Nilsson, 2003). **Working memory** is distinct from short-term memory in that items in memory typically have to be processed or manipulated and this also declines with age (Park et al., 2002).

When considering these studies on memory it should always be remembered that there is considerable variability in memory performance, even in the domains where studies find consistent evidence of decline. Some of these differences are likely to be due to differences in the rate of loss of brain structure (sometimes termed 'brain reserve'). Other variability is likely to be due to differences in how well people can cope or find alternative strategies to perform memory tasks. For example, it has been argued that some older people interpret the sense of familiarity or recognition differently to younger people and are more likely to infer they recalled the event. Higher levels of education earlier in life, as well as higher levels of physical or mental activity later in life, are associated with better memory. This variability illustrates the importance of considering participant sampling and cohort differences in ageing research.

Strategies to promote healthy cognitive ageing

Recently, application of behavioural interventions and non-pharmacological approaches has been demonstrated to improve some aspects of cognitive performance and promote healthy cognitive ageing. This is of particular relevance in older age when cognitive performance in particular domains declines. Such approaches, including cognitive training, neuromodulation and physical exercise are thought to improve cognitive health by rescuing brain networks that are particularly sensitive to ageing and/or augmenting the function of those networks which are relatively resilient to ageing. However, although such approaches may be supported by relatively extensive psychological studies, in terms of improvement or stabilisation of cognitive ability, evidence of underlying changes in biological structure/function, which is needed to support long term changes in cognitive behaviours, is more limited.

- **Cognitive or 'brain' training** – a program of regular mental activities believed to maintain or improve cognitive abilities. Based on the assumptions that practice improves performance, similar cognitive mechanisms underlie a variety of tasks and practicing one task will improve performance in closely related skills/tasks. Such training encompasses both cognitive stimulation and strategy-based interventions, are typically administered via a computer or other electronic medium and aim to restore or augment specific cognitive functions via challenging cognitive tasks that ideally adapt to an individual's performance and become progressively more difficult. Recent meta-analyses of randomized controlled trials of cognitive training in healthy older adults and patients with mild cognitive impairment (MCI) report positive results on the cognitive functions targeted (Basak et al., 2020; Chiu et al., 2017).
- **Neuromodulation** – non-invasive brain stimulation techniques, such as transcranial direct current stimulation (tDCS) and repetitive transcranial magnetic stimulation

(rTMS) have been shown to moderately improve cognitive functioning in older people (Huo et al., 2021) and improve cognitive performance in patients with MCI (Jiang et al., 2021). tDCS is approved as a safe, neuromodulatory technique which delivers a weak, electrical current via electrodes placed on the scalp to directly stimulate cortical targets. rTMS uses an electromagnetic coil to deliver a magnetic pulse that can be targeted at specific cortical regions to modulate neuronal activity and promote plasticity. Although some studies have combined cognitive training and neuromodulation approaches to enhance cognitive performance there is limited evidence of enhanced performance beyond that reported for either approach used in isolation.

- **Physical activity** – structured physical activity, in the form of moderate to vigorous aerobic exercise, has been reported to preserve and enhance cognitive functions in older adults. In particular, it moderately improves global cognitive function in older adults and improves attention, memory and

executive function in patients with MCI (Erickson et al., 2019; Song et al., 2018). However, the mechanisms by which exercise has these effects are not fully understood and it is likely that the mechanisms of exercise may vary depending on individual factors such as age, affective mood and underlying health status.

Attentional control

One example of slowing is that the response to a cue becomes slower with age. The extent of the difference in response times between age groups differs depending on the type of cueing. A cue can be used to create an alerting response, which increases vigilance and task readiness. Older people are slower to show this alerting response (Festa-Martino et al, 2004). A cue can also symbolically direct attention to a specific location. One commonly used example of a symbolic cue is an arrow pointing to a particular location. Several studies have shown that older people are comparatively slower at responding to this type of cue (see Erel & Levy 2016 for an extensive and useful review). On the other hand, cues can also capture attention automatically, for example a loud noise or bright

light. In contrast to the alerting and symbolic cues, numerous studies show that older people maintain an automatic orienting response (e.g. Folk & Hoyer, 1992). The attention effects above are likely to be due, in part, to other effects of ageing, such as sensory change (see section on Sensory change with age). For example, in the study by Folk and Hoyer (1992) older people were slower to respond to arrow cues only when they were small, not larger. This illustrates that the effects of changes in perception in ageing need to be considered when interpreting ageing effects.

Attentional processes are also often measured by visual search tasks. In these tasks participants search for a particular target amongst distracters. The target might be specified in advance ('search for the red H') or be defined by its relation to the distracters ('find the odd one out'). The distracter can be similar to the target, typically leading to slower search (Duncan & Humphreys, 1989), or be sufficiently different that it 'pops out' (Treisman & Gelade, 1980). The distracters might be completely different to the target, allowing people to search based on a feature (e.g. a red H among blue As), or share some features with the target, which is usually termed conjunction search (e.g. a red H among red As and blue As). Visual search performance can be used to test multiple attentional mechanisms such as attentional shifting, attention to different features, response times and attentional strategies.

Older people show slower performance than younger people for conjunction search, compared to relatively

preserved performance in feature-based search (Erel & Levy, 2016). This can be considered analogous to the differences in cueing above. The quick performance in feature search is often considered to be automatic and involuntary (Treisman & Gelade, 1980) and the slower performance in conjunction search to require more complex and voluntary processes. Older people have slower performance when they are required to make more attentional shifts to find the target (Trick & Enns, 1998). However, the role of declines in other processes such as discrimination of the target and distractors, inhibition and disengagement from each location and general slowing are also likely to play a part.

Inhibition

In contrast to directing attention towards a target, we also need to ignore, avoid or suppress irrelevant actions. In psychology this is often referred to as 'inhibition' and it can refer to the ability to ignore distracting items or colours on screen, or to resist the urge to make a specific repeated or strongly cued action. In ageing research, it is often measured by the Stroop task (where participants must name a word but ignore its ink colour) a go/no-go task (where participants must press a button in response to a target on most trials, and avoid that press on a few trials with a different stimulus) or flanker or distracter tasks (where participants' performance with and without a distraction is measured).

Hasher and Zacks (1988) proposed that an age-related decline in inhibition underlies many differences in performance between older and younger people. Deficits can be found in many tasks which appear to be based on inhibition. Kramer et al. (2000) asked people to search for a target item among distracters. For example, older people's reaction times were more affected (slowed) when there was a particularly salient (visible) distracter also on the screen. Older people also do not show as strong 'negative priming' as younger people. Negative priming is found when a distracter on a previous trial becomes the target on the current trial. This extended effect on performance is attributed to the distracter being inhibited so if inhibition is reduced then the negative priming effect is reduced.

On the other hand, some of the findings of inhibitory deficits can be attributed to other factors. For example, differences in Stroop task performance can be partly due to differences in the speed of processing of colours and words with age (Ben-David & Schneider, 2009), and inhibition of responses rather than the sensory profile of the distracter itself (Hirst et al., 2019). A meta-analysis by Rey-Mermet & Gade (2018) suggested that older people's inhibitory deficit is likely to be limited to the inhibition of dominant responses.

Sensory changes with age

Most of us will have noticed someone using reading glasses, or

turning up the TV to better hear the dialogue in a favourite film or drama. Although these are commonly assumed to be the main effects of ageing, the sensory changes with age are varied and some are quite subtle.

Vision

With age there are changes in the eye and in the visual pathways in the brain. A reduction in the amount that the eye focuses at near distances means that almost everyone will need reading glasses at some point. The lens also becomes thicker and yellows, reducing the amount of light entering the optic nerve and the brain. This affects how well we can see both colour and shape. For colour vision, the yellowing of the eye has a greater effect on the shorter, i.e blues and greens wavelengths of light (Ruddock, 1965, Said, 1959). So, for example, when matching red and green to appear the same brightness, more green has to be added for older, compared to younger, participants (Fiorentini et al., 1996).

For shape perception, the reduction in the amount of light passing through the eye reduces people's ability to resolve fine detail (Weale, 1975; Kulikowski, 1971). Furthermore, neural loss and decay with age also contribute to declining visual ability, especially for determining the shape of objects. In general, for coarser patterns (lower spatial frequencies), the differences in performance between older and younger people are likely to be due to cortical changes. For finer detailed

patterns (high spatial frequencies) the loss is more likely to be due to optical factors. Note that glasses correct for acuity, which is mostly driven by the ability to detect and discriminate between small and fine detailed patterns, but visual losses are far more wide-ranging and subtle than this.

Older people also have reduced ability to see movement and things that are moving. Older individuals tend to misjudge the speed of moving items (Snowden & Kavanagh, 2006) and the minimum speed required to discriminate direction of motion is higher for older, compared to younger, people (Wood & Bullmore, 1995). On the other hand, other studies have found that age-related deficits in motion processing are absent or specific to particular stimuli (Atchley & Andersen, 1998) and there are also reports of improvements in performance with age. For instance, older adults are quicker than younger adults to be able to discriminate the direction of large moving patterns (Betts et al., 2005; Hutchinson et al., 2011). This illustrates an important point about ageing and vision: many of the age-related deficits in motion processing are not due to deficits at the level of motion processing *per se*, but due to sensitivity deficits earlier in the processing stream. When presenting stimuli to older adults, it is worth noting that slight changes in details of the stimuli (for instance, their speed or contrast) might make dramatic changes in visibility for older, compared to younger, adults (Allen et al., 2010).

Hearing

As with vision, the effects of age on hearing include declines in the ear as well as in the brain. Also similar to vision there is a reduction in sensitivity to high frequencies. In hearing, high frequencies are perceived as high notes. The ear loses sensitivity with age and this loss starts with the detection of high tones and then goes on to affect detection of low tones (Peelle & Wingfield, 2016).

Although a loss of the ability to hear pure tones is very common in older people, one of the most commonly reported issues with hearing is a loss of ability to discriminate speech when it is in background noise (Moore et al., 2014). This causes trouble with hearing conversations in crowds, as well as dialogue in films and TV. The deficit is found both in subjective and objective measures of hearing in older people. Interestingly, there is an association between ability to hear speech in noise and cognitive decline (Dryden et al., 2017). One suggestion is that this reflects a general loss across all systems of the brain, but another interesting suggestion is that the effort and load of coping with declining sensory systems causes people to do worse on cognitive tasks.

Touch

Touch perception is perhaps the least well understood sense when it comes to ageing. We use our hands to sense texture and

shape, either pressing or stroking a surface. Beyond this, our entire body is sensitive to touch to some degree, for example touch and pressure on the feet affect balance, and touch on our body tells us if we are comfortable. We know that ageing affects the condition of the skin as well as the ability to control our movements. Skin hydration, elasticity and compliance are all reduced with increased age (Zhang & Duan, 2018). Changing the skin will change how well it is able to sense differences in texture and shape. There are also changes in the areas of the brain that process touch, and in the pathways that connect the skin and the brain (McIntyre et al., 2021), both affecting basic tactile sensitivity (Bowden & McNulty, 2013; Goble et al.,1996).

Taste and smell

Taste and smell are critically important for quality of life and health and also show decline with ageing. Loss of appetite is a common issue for the old, and loss of smell and taste contribute to this. Loss of taste or smell is unpleasant at any age, making food unpalatable, but also making it difficult to identify when food is 'off' or when dirt is present. The change in smell and taste is gradual over the life span, but by the age of 65 there are measurable differences in the ability to detect flavours or smells (Stevens, 1998).

Key Takeaways

- Ageing causes natural and inevitable changes in both brain structure and function – however, we don't yet fully understand the rate of change and the processes involved.
- Changes to the brain which may affect cognitive and sensory functions occur at molecular, synaptic and cellular levels – some of which, but not all, is driven by genetic factors.
- Understanding the mechanisms of ageing is important as this may identify approaches to try and alleviate age-related decline in cognition and sensory functions, along with identifying psychological and lifestyle factors which may help promote healthy cognitive ageing.

References and further reading

Allen, H. A., Hutchinson, C. V., Ledgeway, T., & Gayle, P. (2010). The role of contrast sensitivity in global motion processing deficits in the elderly. *Journal of Vision, 10* (15). 1-10 https://doi.org/10.1167/10.10.15

Atchley, P., & Andersen, G. J. (1998). The effect of age, retinal eccentricity, and speed on the detection of optic flow components. *Psychology and Aging, 13*(2), 297-308. https://doi.org/10.1037/0882-7974.13.2.297

Basak, C., Qin, S., & O'Connell, M.A. (2020). Differential effects of cognitive training modules in healthy aging and mild cognitive impairment: A comprehensive meta-analysis of randomized controlled trials. *Psychology and Aging*, *35*(2), 220–49. https://doi.org/10.1037/pag0000442

Ben-David, B. M. & Schneider, B. A. (2009). A sensory origin for color-word effects in aging: A meta-analysis. *Aging Neuropsychology and Cognition, 16*(5), 505-534. https://doi.org/10.1080/13825580902855862

Bettio, L.E.B., Rajendran, L., & Gil-Mohapel, J. (2017). The effects of aging in the hippocampus and cognitive decline. *Neuroscience & Biobehavioral Reviews*, *79*, 66-86. https://doi.org/10.1016/j.neubiorev.2017.04.030

Betts, L. R., Taylor, C. P., Sekuler, A. B. & Bennett, P. J. (2005). Aging reduces center-surround antagonism in visual motion processing. *Neuron, 45*(3), 361-366. https://doi.org/10.1016/j.neuron.2004.12.041

Berry, A.S., Shah, V.D., Baker, S.L., Vogel, J.W., O'Neil, J.P., Janabi, M., Schwimmer, H.D., Shawn, M.M., & Jagust, W.J. (2016). Aging affects dopaminergic neural mechanisms of cognitive flexibility. *The Journal of Neuroscience*, *36*(50), 12559-12569. https://doi.org/10.1523%2FJNEUROSCI.0626-16.2016

Bowden, J.L. & McNulty, P. A. (2013). Age-related changes in cutaneous sensation in the healthy human hand. *Age*, *35*(4), 1077-1089. https://doi.org/10.1007%2Fs11357-012-9429-3

Chiu, H.-L., Chu, H., Tsai, J.-C., Liu, D., Chen, Y.-R., Yang, H.-L., & Chou, K.-R, (2017). The effect of cognitive-based training for the healthy older people: A meta-analysis of randomized controlled trials. *PLoS ONE*, *12*(5), e0176742. https://doi.org/10.1371/journal.pone.0176742

Di Benedetto, S., Muller, L., Wenger, E., Duzel, S., & Pawelec, G. (2017). Contribution of neuroinflammation and immunity to brain aging and the mitigating effects of physical and cognitive interventions. *Neuroscience and Biobehavioral Reviews*, *75*, 114-128. https://doi.org/10.1016/j.neubiorev.2017.01.044

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The association between cognitive performance and speech-in-noise perception for adult listeners: A systematic literature review and meta-analysis. *Trends in Hearing, 21*. https://doi.org/10.1177/2331216517744675

Duncan, J., & Humphreys, G. W. (1989). Visual search and

stimulus similarity. *Psychological Review, 96*(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Erel, H. & Levy, D. A. (2016) Orienting of visual attention in aging. *Neuroscience and Biobehavioral Reviews*, *69*, 357-380. https://doi.org/10.1016/j.neubiorev.2016.08.010

Erickson, K.I., Hillman, C., Stillman, C.M., Ballard, R.M., Bloodgood, B., Conroy, D.E., Macko, R., Marquez, D., Petruzzello, S.J., & Powell, K.E. (2019). Physical activity, cognition, and brain outcomes: A review of the 2018 physical activity guidelines. *Medicine & Science in Sports & Exercise*, *51*(6), 1242-1251. https://doi.org/10.1249/mss.0000000000001936

Farokhian, F., Yang, C., Beheshti, I., Matsuda, H., & Wu, S. (2018) Age-related gray and white matter changes in normal adult brains. *Aging and disease*, *8*(6), 899-909. https://doi.org/10.14336%2FAD.2017.0502

Festa-Martino, E., Ott, B.R., & Heindel, W.C. (2004). Interactions between phasic alerting and spatial orienting: effects of normal aging and Alzheimer's disease. *Neuropsychology*, *18*(2), 258-68. https://doi.org/10.1037/0894-4105.18.2.258.

Fiorentini, A., Porciatti, V., Morrone, M. C. & Burr, D. C. (1996). Visual ageing: Unspecific decline of the responses to luminance and colour. *Vision Research, 36*(21), 3557-3566. https://doi.org/10.1016/0042-6989(96)00032-6

Folk, C.L., & Hoyer, W.J. (1992). Aging and shifts of visual spatial attention. *Psychology and Aging, 7*(3), 453–465. https://doi.org/10.1037//0882-7974.7.3.453

Goble, A.K., A.A. Collins, & R.W. Cholewiak, (1996). Vibrotactile threshold in young and old observers: The effects of spatial summation and the presence of a rigid surround. *Journal of the Acoustical Society of America*, *99*(4), 2256-2269. https://doi.org/10.1121/1.415413

Hasher, L. & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology-General*, *108*(3), 356-388. https://doi.org/10.1037/0096-3445.108.3.356

Hirst, R. J., Kicks, E. C., Allen, H. A. & Cragg, L. (2019). Cross-modal interference-control is reduced in childhood but maintained in aging: A cohort study of stimulus- and response-interference in cross-modal and unimodal stroop tasks. *Journal of Experimental Psychology-Human Perception and Performance*, *45*(5), 553-572. https://doi.org/10.1037%2Fxhp0000608

Huo, L., Zhu, X., Zheng, Z., Ma, J., Ma, Z., Gui, W., & Li, J. (2021). Effects of transcranial direct current stimulation on episodic memory in older adults: a meta-analysis. *Journal of Gerontology: Series B*, *76*(4), 692–702. https://doi.org/10.1093/geronb/gbz130

Hutchinson, C., Allen, H. & Ledgeway, T. (2011) When older is better: Superior global motion perception in the elderly. *Perception, 40*, 115-115.

Jiang, L., Cui, H., Zhang, C., Cao, X., Gu, N., Zhu, Y., Wang, J., Yang, Z., & Li, C. (2021). Repetitive transcranial magnetic stimulation for improving cognitive function in patients with mild cognitive impairment: A systematic review. *Frontiers in Aging Neuroscience*, *12*. https://doi.org/10.3389/fnagi.2020.593000

Karrer, T.M., Josef, A.K., Mata, R., Morris, E.D., & Samanez-Larkin, G.R. (2017). Reduced dopamine receptors and transporters but not synthesis capacity in normal aging adults: A meta-analysis. *Neurobiology of Aging*, *57*, 36-46. https://doi.org/10.1016/j.neurobiolaging.2017.05.006

Koen, J.D., & Yonelinas, A.P. (2014). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection and familiarity: a meta-analytic review. *Neuropsychology Review 24*(3), 332-54. https://doi.org/10.1007/s11065-014-9266-5.

Kramer, A. F., Hahn, S., Irwin, D. E. & Theeuwes, J. (2000). Age differences in the control of looking behavior: Do you know where your eyes have been? *Psychological Science*, *11*(3), 210-217. https://doi.org/10.1111/1467-9280.00243

Kulikowski, J. J. (1971). Some stimulus parameters affecting spatial and temporal resolution of the human eye. *Vision Research, 11*(1), 83-93. https://doi.org/10.1016/0042-6989(71)90206-9

La Voie, D., & Light, L. L. (1994). Adult age differences in

repetition priming: A meta-analysis. *Psychology and Aging*, *9*(4), 539–553. https://doi.org/10.1037/0882-7974.9.4.539

Liu, X. Z. & Yan, D. (2007). Ageing and hearing loss. *Journal of Pathology, 211*(2), 188-197. https://doi.org/10.1002/path.2102

McIntyre, S., Nagi, S.S., McGlone, F., & Olausson, H. (2021). The effects of ageing on tactile function in humans. *Neuroscience*, *464*, 53-58. https://doi.org/10.1016/j.neuroscience.2021.02.015

Melzer, D., Pilling, L.C., & Ferrucci, L. (2020) The genetics of human ageing. *Nature Reviews Genetics*, *21*, 88-101. https://doi.org/10.1038/s41576-019-0183-6

Moore, D. R., Edmundson-Jones, M., Dawes, P., Fortnum, H., McDormack, A., Pierzycki, R.H., & Munro, K.J. (2014). Relation between speech-in-noise threshold, hearing loss and cognition from 40-69 years of age. *PLoS One, 9*(9), e107720. https://doi.org/10.1371/journal.pone.0107720

Nilsson L-G. (2003) Memory function in normal aging. *Acta Neurologica Scandinavica, 107*, 7–13. https://doi.org/10.1034/j.1600-0404.107.s179.5.x

Nyberg, L., Maitland, S.B., Ronnlund, M., Backman, L., Dixon, R.A., Wahlin, A., Nilsson, L.G. (2003). Selective adult age differences in an age- invariant multifactor model of declarative memory. *Psychology and Aging, 18*(1), 149–160. https://doi.org/10.1037/0882-7974.18.1.149

Ogoh, S. (2017). Relationship between cognitive function and regulation of cerebral blood flow. *Journal of Physiological Sciences*, *67*, 345-351. https://doi.org/10.1007/s12576-017-0525-0

Park, D. C., Lautenschlager, G. Hedden, T., & Davidson, N. S. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299-320. https://doi.org/10.1037/0882-7974.17.2.299

Peelle, J. E. & Wingfield, A. (2016). The neural consequences of age-related hearing loss. *Trends in Neurosciences*, *39*(7), 486-497. https://doi.org/10.1016/j.tins.2016.05.001

Petralia, R.S., Mattson, M.P., & Yao, P.J. (2014). Communication breakdown: the impact of ageing on synapse structure. *Ageing Research Reviews*, *14*, 31-42. https://doi.org/10.1016/j.arr.2014.01.003

Rey-Mermet, A. & Gade, M. (2018). Inhibition in aging: What is preserved? What declines? A meta-analysis. *Psychonomic Bulletin & Review*, *25*, 1695-1716. https://doi.org/10.3758/s13423-017-1384-7

Rodriguez-Rodero, S., Fernandez-Morera, J.L., Menendez-Torre, E., Calvanese, V., Fernandez, A.F., & Fraga, M.F. (2011). Aging genetics and aging. *Aging and Disease*, *2*, 186-195. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3295054/

Ronnlund, M., Nyberg, L., Backman, L., & Nilsson, L.G. (2005). Stability, growth and decline in adult life span development of declarative memory: cross-sectional and

longitudinal data from a population-based study. *Psychology and Aging, 20*(1), 3–18. https://doi.org/10.1037/0882-7974.20.1.3

Ruddock, K. H. (1965). The effect of age upon colour vision. II. Changes with age in light transmission of the ocular media. *Vision Research*, *5*(1-3), 47-58. https://doi.org/10.1016/0042-6989(65)90074-X

Said, F. S. W. R. A. (1959).The variation with age of the spectral transmissivity of the living human crystalline lens. *Gerontologia, 3*, 213-231. https://doi.org/10.1159/000210900

Sailasuta, N., Ernst, T., & Chang, L. (2008). Regional variations and the effects of age and gender on glutamate concentrations in the human brain. *Magnetic Resonance Imaging*, *26*(5), 667–75. https://doi.org/10.1016/j.mri.2007.06.007

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*(3), 403-428. https://doi.org/10.1037/0033-295X.103.3.403

Snowden, R. J. & Kavanagh, E. (2006). Motion perception in the ageing visual system: Minimum motion, motion coherence, and speed discrimination thresholds. *Perception, 35*(1), 9-24. https://doi.org/10.1068/p5399

Song, D., Yu, D.S.F., Li, P.W.C., & Lei, Y. (2018) The effectiveness of physical exercise on cognitive and psychological outcomes in individuals with mild cognitive impairment: A systematic review and meta-analysis.

International Journal of Nursing Studies, *79*, 155–164. https://doi.org/10.1016/j.ijnurstu.2018.01.002

Stevens, J. C., Cruz, L. A., Marks, L. E. & Lakatos, S. (1998). A multimodal assessment of sensory thresholds in aging. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*, *53*(4), 263-272. https://doi.org/10.1093/geronb/53b.4.p263

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136. https://doi.org/10.1016/0010-0285(80)90005-5

Trick, L.M., & Enns, J.T. (1998). Lifespan changes in attention: The visual search task. *Cognitive Development. 13*(3), 369–386. https://doi.org/10.1016/S0885-2014(98)90016-8

Weale, R. A. (1975) Senile changes in visual acuity. *Transaction of the Ophthalmological Society*, *95*(1), 36-38.

Wong, D. F., Wagner, H.N., Dannals, R.F., Links, J.M., Frost, J.J., Ravert, H.T., Wilson, A.A., Rosenbaum, A.E., Gjedde, A., Douglass, K.H., Petronis, J.D., Folstein, M.F., Toung, J.K.T., Bums, H.D., & Kuhar, M.J. (1984). Effects of age on dopamine and serotonin receptors measured by positron tomography in the living human brain. *Science, 226*(4681), 1393-1396. https://doi.org/10.1126/science.6334363

Wood, J. M. & Bullmore, M. A. (1995). Changes in the lower displacement limit for motion with age. *Ophthalmic and*

Physiological Optics, 15(1), 31-36. https://doi.org/10.1046/j.1475-1313.1995.9592789.x

Zhang, S.B. & Duan, E. K. (2018). Fighting against skin aging: The way from bench to bedside. *Cell Transplantation*, *27*(5), 729-738. https://doi.org/10.1177/0963689717725755

About the authors

Professor Claire Gibson

UNIVERSITY OF NOTTINGHAM

https://twitter.com/PreclinStroke

Professor Claire Gibson obtained a BSc degree in Neuroscience from the University of Sheffield and her PhD from the University of Newcastle. She then gained a number of years' experience researching the mechanisms of injury following CNS damage – initially focusing on spinal cord injury and moving on later to cerebral stroke. She is now a Professor of Psychology at the University of Nottingham whose research pursues the mechanisms of damage and investigates novel treatment approaches following CNS disorders, focusing primarily on stroke and neurodegeneration. She regularly teaches across the spectrum

of biological psychology to both undergraduate and postgraduate students.

Professor Harriet Allen
UNIVERSITY OF NOTTINGHAM

Professor Harriet Allen received her BSc and PhD in Psychology from the University of Nottingham. She then worked in Montreal, at McGill University, and the University of Birmingham, UK before returning to the University of Nottingham where she is now a Professor of Lifespan Psychology. She researches how sensory processes interact with attention over the lifespan and teaches research methods, cognitive psychology and perception to undergraduates and postgraduates.

19.

DEMENTIAS

Professor Claire Gibson and Dr Catherine Lawrence

Learning Objectives

- To gain an overview of the symptoms and main causes of dementia along with approaches used to diagnose dementia
- To understand the symptoms and pathological consequences of the main causes of dementia – focusing on Alzheimer's disease, vascular dementia and dementia with Lewy bodies
- To gain an understanding of the various pharmacological and psychological

approaches to treat Alzheimer's disease

Dementia is a **syndrome** associated with a progressive decline in brain functioning, most commonly affecting memory. Symptoms of dementia can be wide-ranging and have huge individual variability which may include, but are not limited to:

- loss of memory
- apathy
- difficulties in language
- difficulties in judgement
- difficulties in motor control
- speed of (cognitive) processing

A person with dementia may also experience paranoia, hallucinations, and find it challenging to make decisions and live independently. There are currently no cures for dementia. However, depending on the type of dementia, and the underlying cause, treatments may exist which can stabilise symptoms and slow the progression of the disease.

There are many different causes of dementia (see Table 6.2) with Alzheimer's disease (AD) being the most common (see

Figure 6.11). Whilst some symptoms of dementia, such as memory loss, might be expected with the normal ageing process, dementia is a syndrome in which the deterioration in cognitive function is beyond that which might be expected from the usual consequence of biological ageing and symptoms are usually severe enough to interfere with daily activities.

Table 6.2. Common causes of dementia

Type of dementia	Brief description of cause
Alzheimer's Disease	Progressive degeneration of brain tissue
Vascular Dementia	Block or reduction in blood flow to the brain
Mixed Dementia	Several types of dementia contribute to symptoms
Dementia with Lewy Bodies	Abnormal aggregates of protein that develop inside neurons
Frontotemporal Dementia	Progressive degeneration of the temporal and frontal lobes of the brain
Parkinson's Disease with Dementia	Development of dementia symptoms as disease progresses
Other	May include conditions such as Creutzfeld-Jacob Disease; Depression; Multiple Sclerosis, Down's syndrome

Credit: Claire Gibson

Risk factors for dementia

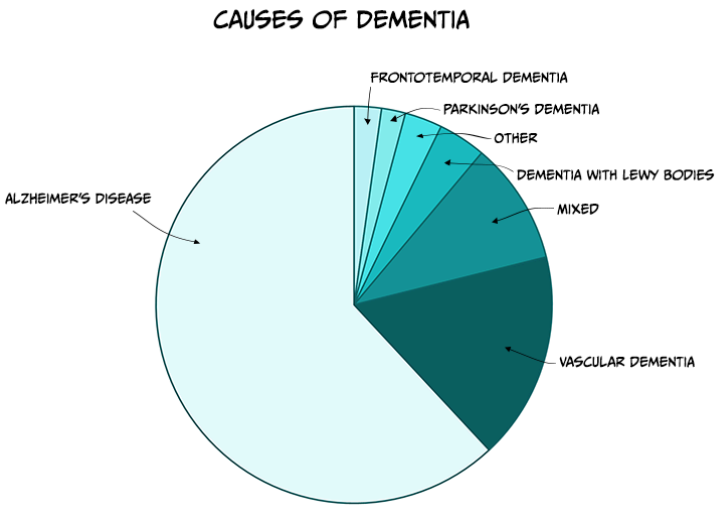

Fig 6.12. Pie chart showing different causes of dementia, as a percentage of all cases of dementia

Although age is the strongest risk factor known for dementia it does not occur as an inevitable consequence of biological ageing. Additionally, dementia does not exclusively affect older people – young onset dementia, typically defined as onset of symptoms before the age of 65, accounts for up to 9% of all cases of dementia. There are various risk factors which have been identified to increase the risk of developing dementia, including:

- smoking

- excessive alcohol use
- low levels of physical activity
- high cholesterol
- atherosclerosis
- social isolation
- obesity
- mild cognitive impairment (MCI)

There are also a number of known genetic risk factors for developing dementia, in particular Alzheimer's Disease (see below). It is likely that the development of dementia occurs due to a combination of various risk factors – some of which are modifiable (e.g. diet, physical activity) and some which are not (e.g. genetic).

Mild Cognitive Impairment

MCI is typically an early stage of memory loss, or other cognitive ability loss, such as language or visual/spatial perception. Individuals diagnosed with MCI are able to maintain the ability to live independently and perform most activities of daily living. Importantly, people with MCI exhibit a decline in memory and/or other cognitive areas beyond a level we would expect to see during normal ageing. MCI is not a type of dementia but it is associated with a higher risk of developing dementia, in particular AD (Boyle et al., 2006).

Alzheimer's Disease

AD is a neurodegenerative disorder that leads to cognitive decline and memory loss. AD is characterised pathologically by the accumulation of **extracellular beta amyloid (Aβ) plaques**, **neurofibrillary tangles**, and **neuroinflammation**.

First described by Alois Alzheimer in 1907, AD is the most common form of dementia, accounting for between 60-80% of total dementia cases. Currently it is estimated that 30 million people worldwide have AD, which is predicted to rise to up to 90 million by 2050. In the UK there are over 500,000 people living with AD and, if the prevalence remains the same, this is forecast to rise to over 1 million by 2025 and 2 million by 2050 (Prince et al., 2014). The risk of AD increases with age affecting 1 in 20 people under the age of 65, 1 in 14 over the age of 65 and 1 in 6 over the age of 80. In England and Wales, AD was one of the leading causes of death accounting for over 10% of deaths registered in 2021 (Office of National Statistics, 2022).

There are two types of AD: early onset (familial, EOAD) or late onset (sporadic, LOAD), which are diagnosed before or after the age of 65 respectively. EOAD is rare and accounts only for up to 5% of all AD cases. It is thought to be caused by mutations in one of three genes: amyloid precursor protein (APP), presenilin 1 (PSEN1) or presenilin 2 (PSEN2), that lead to increased production of Aβ plaques. No single gene mutation is thought to be the cause of the more common

LOAD, but it is suggested to be driven by a complex interplay between genetic and environmental factors. However, genetic mutations have been identified in LOAD that increase the risk of developing AD, the strongest being the apolipoprotein E4 gene (APOE4). More recently, genome-wide association studies implicate genes associated with the innate immune system and microglia (the resident immune cells of the brain), including the phagocytic receptors CD33 and TREM2 (triggering receptor expressed on myeloid cells 2) (Griciuc & Tanzi, 2021).

Symptoms

AD develops slowly over several years, sometimes decades, so the symptoms are not always obvious at first and also depend on the stage of the disease. The predominant symptoms of AD are progressive and irreversible impairment in memory and cognitive function. One of the first signs of cognitive decline in AD is the generalised disruption of declarative memory, including the inability to learn and remember new facts (**semantic memory deficit**) and recall past experiences (**episodic memory deficit**) and this is often characterised by an abnormally rapid rate of forgetfulness (Holger, 2013). Symptoms that occur early in the disease include forgetting the names of objects and places, misplacing items (losing your house keys), and repetition such as asking the same question several times. Deficits in episodic memory are one of the best

indicators of early AD compared to other forms of dementia and have been reported in the pre-clinical stage of the disease.

As AD progresses, other cognitive deficits manifest, including disruptions in language (**aphasia**), spatial orientation (e.g. judging distances), attention and executive functions. In contrast, procedural memory (habits and skills) remains relatively unaffected until the late stages of the disease when there are significant problems with both short- and long-term memory.

AD is also associated with various behavioural and psychological symptoms including depression, anxiety, apathy, irritability, aggression, disinhibition and reduced curiosity. These changes all form part of the behavioural and psychological symptoms of dementia (BPSD), which are commonly seen in people with AD in either the early or late stages of the disease but which can fluctuate throughout its progression. Research also indicates that BPSD might contribute to cognitive decline as the disease progresses (Gottesman & Stern, 2019).

Disturbances in sleep-wake patterns are also a common feature, with people with AD displaying increased sleepiness during the day and increased wakefulness at night. AD patients often exhibit a shift in their body clock, tending to wake up later in the day and going to sleep later than non-demented controls. Similarly, circadian shifts in eating patterns are observed in people with AD who show a tendency to have their biggest meal during breakfast and display increased

preference for sweet food. However, considerable weight loss is also a common symptom which can lead to frailty and weakness. Over time, the ability to perform everyday activities becomes increasingly impaired and eventually leads to permanent dependence on caregivers.

Neuropathology

Macroscopic features

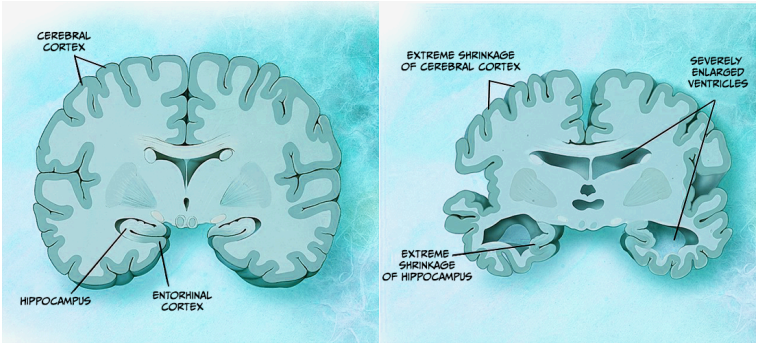

Fig. 6.13. Diagrammatic representation of macroscopic changes that occur with Alzheimer's disease

Several pathological features can be seen macroscopically in the brain of someone with AD (DeTure & Dickson, 2019). These features get worse with disease progression and can be visualised using imaging (e.g. magnetic resonance imagining, MRI) and post-mortem analysis. For example, **cortical**

atrophy (thinning), which is characterised by enlarged sulcal spaces and atrophy of the gyri, is seen prominently in the frontal and temporal cortices of people with AD. As a result of this atrophy there is a reduction in brain weight and ventricular enlargement (see Figure 6.13). The hippocampus, a crucial region for learning and memory, also shows atrophy thought to be due to neuronal loss. However, while these features suggest that someone has AD, they can sometimes be seen in other dementias and also in clinically normal people.

Microscopic features

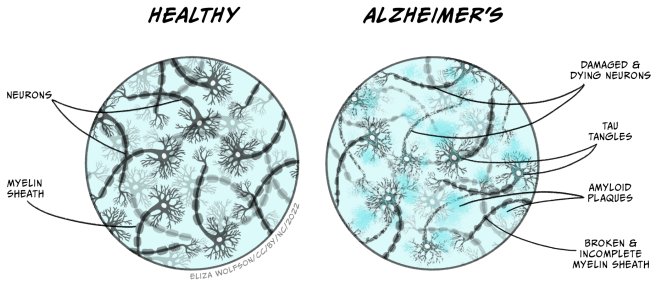

Fig 6.14. Microscopic changes with Alzheimer's

The key neuropathological hallmarks that define AD are the presence of Aβ plaques (or senile plaques) and neurofibrillary tangles (DeTure & Dickson, 2019). It is worth noting that the presence of such plaques and tangles are not unique to AD as

they are also seen during the ageing process, but the density and location are distinct in AD.

These features are initially located in temporal lobe structures e.g. hippocampus and entorhinal cortex, but can spread to other areas as the disease progresses. Aβ plaques consist of insoluble aggregates of Aβ that are found in the brain parenchyma. The genetic mutations seen in people with EOAD affect the processing of amyloid precursor protein (APP), which subsequently leads to a build-up of Aβ plaques. APP is processed by three enzyme complexes known as α, β and γ secretase. APP is normally cleaved by α-secretase then γ-secretase but, in people with EOAD, APP is processed by β- and γ-secretase which results in different species Aβ being produced that are far more prone to aggregation.

The other characteristic sign in AD is the presence of neurofibrillary tangles composed of hyperphosphorylated tau. Normally, tau's role is to regulate elements of the microtubule cytoskeleton such as stabilisation and facilitation of axonal transport. In AD, tau becomes hyperphosphorylated and forms neurofibrillary tangles inside neurons. This disturbs microtubule structure causing major problems in the neuronal function and ultimately leads to neuronal cell death.

Other features of AD pathology include synaptic loss that precedes neuronal loss and strongly correlates with cognitive decline (Terry et al., 1991). There is also an inflammatory response that is observed in the brains of people with AD,

including microglia and astrocyte activation around Aβ plaques that is thought to contribute to disease pathogenesis.

Diagnosis

Currently there is no simple and reliable test for diagnosing AD. If a person is suspected to have AD, their cognitive ability will be evaluated using tests that assess memory, concentration and attention, language and communication skills, orientation, and visual and spatial abilities.

The most commonly used test to measure cognitive impairment is the **Mini Mental State Exam** (MMSE), which was first introduced in 1975 by Marshal Folstein and colleagues. The MMSE is a 30-point assessment involving a series of questions and tests that each score points if answered correctly. Specifically, the MMSE test measures short-term memory (e.g. memorising an address and recalling it a few minutes later), attention and concentration (e.g. spelling a simple word backwards), language (e.g. identifying common objects by name), orientation to time and place (e.g. knowing where you are, and the day of the week) and comprehension and motor skills (drawing a slightly complicated shape such as copying a pair of intersecting pentagons). Scores of 24 or higher generally indicate normal cognition, while scores below this can indicate mild (19-23), moderate (10-18) or severe (9) cognitive impairment. The MMSE can therefore be used to indicate how severe a person's symptoms are, but if repeated

it can assess changes in cognitive ability and how quickly their AD is progressing. The onset of cognitive symptoms indicates severe neurodegeneration has already taken place and diagnosis at an earlier stage is therefore needed.

Furthermore, while cognitive impairment is a symptom of AD, several other conditions associated with dementia (e.g. vascular dementia) can lead to cognitive decline and a reduction in MMSE score, so in order to determine the likelihood of AD, patients might also undergo a brain scan. Neuroimaging techniques such as MRI have dramatically advanced the ability to diagnose people with AD at an earlier stage (Kim et al., 2022). Structural MRI is used to detect changes in brain structure such as cerebral atrophy and ventricular enlargement. Positron emission tomography (PET) imaging can detect other characteristic hallmarks of AD such as brain hypometabolism, characterised by decreased brain glucose consumption, and Aβ burden, but these types of scans are more commonly used in research rather than as a clinical diagnostic tool.

Treatments

Despite large scientific research efforts, there is no cure for AD and treatment options are limited with a few pharmacological treatments and non-pharmacological interventions available. Examples of non-pharmacological therapies to improve memory, problem-solving skills, mood and wellbeing include

cognitive stimulation therapy, cognitive rehabilitation and reminiscence/life story work (see insert box). In the UK there are four pharmacological treatments licenced for AD. These include three acetylcholinesterase (AChE) inhibitors donepezil (Aricept), rivastigmine (Exelon) and galantamine (Reminyl) and the N-methyl-D-aspartate (NMDA) receptor antagonist memantine (Namenda). However, these drugs only provide symptomatic effects by reducing the severity of common cognitive symptoms, and while they can increase the quality of life, they do not alter the course and progression of the disease. There are also some drugs that might be prescribed for the symptoms of BPSD including the antipsychotic medicines risperidone or haloperidol or antidepressants if depression is suspected as a cause of anxiety. While there are no disease-modifying treatments licenced in the UK, in 2021 the United States Food and Drug Administration (FDA) approved the use of aducanumab (Aduhelm), a monoclonal antibody designed to bind and eliminate aggregated Ab for treatment in AD, although there are still uncertainties around the benefits it may bring. Therefore, effective interventions to halt or reverse the neurodegeneration seen in AD are still needed.

Psychological approaches for

dementia

Psychological approaches are aimed at improving cognitive abilities (e.g. cognitive training/stimulation), enhancing emotional well-being (e.g. activity planning, reminiscence therapy), reducing behavioural symptoms (e.g. music therapy) and promoting everyday functioning (e.g. occupational therapy). Whilst such approaches do not prevent or delay the progression of the underlying cause of dementia, they can improve the quality of life for the patient and their caregivers (Logsdon et al., 2007; Woods et al., 2018). Some specific examples:

- Cognitive Training/Stimulation – adapted from rehabilitation programmes designed for individuals with neurological disorders (e.g. stroke, traumatic brain injury) its goal is to improve memory, attention and general cognitive function. It typically involves strategies such as memory training, general problem solving (including games and

puzzles), use of mnemonic devices and/or use of external memory aids such as notebooks, calendars.

- Reminiscence Therapy – involves discussing events and experiences from an individual's past aiming to stimulate memories, mental activity and improve well-being. It is usually supported by external aids such as photographs, music, objects and may involve direct discussion with the individual or involve a wider family or social group.
- Music/Art Therapy – aimed at improving the mood, alertness and engagement of individuals with dementia. Through engagement with music and/or art this can help trigger memories, stimulate communication and build confidence – all of which impact positively on the quality of life for individuals with dementia. This type of approach allows for self-expression and engagement of individuals which has been shown to reduce agitation and distressing behaviour.

Cholinesterase inhibitors

Acetylcholine (ACh) is a neurotransmitter produced in cholinergic neurones that has a role in memory, thinking, language and attention. In AD there is a loss of cholinergic neurones particularly in the hippocampus, cortex and amygdala, which leads to a reduction in ACh. The enzyme AChE breaks down ACh, and AChE inhibitors (donepezil, rivastigmine, and galantamine) are therefore believed to treat the cognitive symptoms of AD by increasing the levels of ACh in the brain. These drugs are used treat the symptoms of mild to moderate AD and can lead to an improvement in thinking, memory, communication or day-to-day activities. In some people a noticeable improvement is not seen, but their symptoms do not worsen as quickly as expected. Some common side-effects of AChE inhibitors are diarrhoea, feeling or being sick, trouble sleeping, muscle cramps and tiredness.

Memantine

Memantine is the most recent drug to be approved for the treatment of AD in the UK. Pharmacologically, it is a non-competitive antagonist of the NMDA receptor. In AD, NMDA receptor over-activity due to an excess of glutamate is thought to result in neuronal cell death as well as calcium-dependent neurotoxicity, therefore memantine is thought to prevent these toxic effects of glutamate and reduce the symptoms of AD. Memantine is recommended for people

with severe AD, or those with moderate AD who are unable to use AChE inhibitors, but is often used in combination with these drugs. Common side effects include drowsiness, dizziness, constipation, headaches and shortness of breath.

Vascular dementia

Vascular dementia is the second commonest cause of dementia after AD and it occurs as a consequence of reduced blood flow to the brain. All cells within the brain require a constant supply of oxygen and nutrients in order to function and these are delivered via the blood supply within the brain's vascular network. Any interruption or reduction in the blood flow within the brain, for example as a consequence of stroke, can result in impaired function of brain cells, cell death and disruption of cognitive and motor processes.

Symptoms occurring following vascular dementia tend to vary quite considerably between individuals as they depend on the location of the damage and symptoms may develop suddenly, for example following a stroke, or more gradually, such as with small vessel disease. Some symptoms of vascular dementia may be similar to those of other types of dementia – however, although memory loss is typical of the early stages of AD, it is not usually the main early symptom of vascular dementia. The most common cognitive symptoms in the early stages of vascular dementia include problems with planning/organising and decision making, slower speed of cognitive

processing, inattention and short periods of confusion. There are three main types of vascular dementia:

- Subcortical vascular dementia – reported to be the most common type of vascular dementia, occurs as a consequence of disease of the very small arteries that lie within subcortical regions of the brain and is termed small vessel disease (Tomimoto, 2011). Over time, the walls of these vessels thicken and therefore the vessel lumen narrows leading to reduced blood flow (Figure 6.14) and subsequent areas of brain damage termed 'infarcts' are produced. Subcortical structures of the brain are important for processing complex activities such as memory and emotions. It can be distinguished from AD because it is associated with more extensive white matter infarcts and less severe atrophy of the hippocampus.
- Multi-infarct dementia – occurs when an individual experiences a series of mini-strokes, which are sometimes referred to as a transient ischemic attacks. Such mini-strokes cause a temporary reduction in blood flow to the brain and whilst the patient may only experience temporary symptoms at the point of experiencing the mini-stroke they can result in generation of infarcts. Over time, if a number of infarcts develop then the cumulative damage may be sufficient for the individual to develop symptoms of dementia (McKay and Counts,

2017).

- Post-stroke dementia – about 20% of individuals who experience an ischaemic stroke will develop dementia within the following 6 months. An ischaemic stroke is caused by the presence of a clot in a blood vessel which reduces blood flow within that cerebral blood vessel, resulting in tissue loss and brain dysfunction (Mijajlović et al., 2017). Factors which increase the risk of cardiovascular disease and ischaemic stroke, such as hypertension and high cholesterol, also increase the risk of cognitive decline post-stroke.

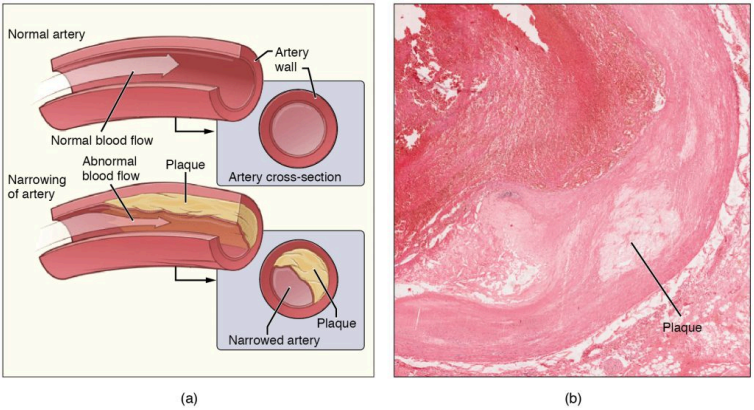

Fig 6.15. View of blood vessel artery and arteriosclerotic blood vessel

Dementia with Lewy Bodies

Dementia with Lewy bodies (DLB) is a progressive disease associated with abnormal deposits of a protein called alpha-synuclein in neuronal and non-neuronal cells within the brain (Outerio et al., 2019). These deposits, termed Lewy bodies, named after FH Lewy, the German doctor who first identified them, affect neurotransmitter functioning, in particular ACh and dopamine, which in turn disrupts cognitive functioning, movement, behaviour, and mood. DLB causes a range of symptoms, some of which are shared with AD and some with Parkinson's disease, resulting in DLB being commonly wrongly diagnosed (see insert box for methods used to diagnose dementia). However, symptoms more commonly associated with DLB rather than other causes of dementia include sleep disturbances, visual hallucinations and motor symptoms. Having a family member with DLB also may increase a person's risk, though LBD is not considered a genetic disease. Variants in three genes, APOE, synuclein alpha (SNCA) and glucocerebrosidase (GBA), have been associated with an increased risk, but for the majority of DLB cases, the cause is unknown.

Testing for dementia

There is no single test for dementia but medical doctors use information from a variety of approaches (listed below) to determine if an individual is experiencing dementia and, if so, what the underlying cause is. Understanding the underlying cause of a patient's dementia is important as it will inform treatment approaches and determine the likely progression of the disease and associated symptoms.

- Medical history to ascertain how any symptoms are affecting daily life along with ensuring any other existing medical conditions (e.g. hypertension) are being treated appropriately.
- Tests of cognitive ability – typically involve neuropsychological tests of memory, attention, problem solving and awareness of time and place.
- Blood tests to check for other conditions

which may be causing symptoms which mimic those seen in dementia. Such tests may typically check the function of the liver, kidneys and thyroid.

- Brain scans can detect signs of brain damage which may help identify the underlying cause of dementia. For example, MRI scans can provide more detailed information about blood vessel damage that might indicate vascular dementia or show atrophy in specific brain areas. For example, hippocampal atrophy is a strong indicator of AD whereas atrophy in the frontal and temporal lobes are more typical of frontotemporal dementia. Other types of scan, e.g. CT scan, may be used to rule out the presence of a brain tumour. Clinical research studies tend to use scans more, such as a PET scan, to identify markers of interest e.g. glucose, in specific relation to disease progression and/or evaluation of potential new therapeutics. It is likely the majority of individuals will not receive a brain scan if the various other tests and assessments show that dementia is a likely diagnosis.

Other dementias

Mixed dementia may be diagnosed when a person has more than one underlying cause of dementia – most commonly this would be co-occurrence of AD and vascular dementia, although other possible combinations are possible such as AD and DLB. Mixed dementia tends to be more common in older age groups (over 75 years of age), and is reported to account for 10% of all dementia diagnoses.

Frontotemporal dementia (FTD) is a rare form of dementia sometimes referred to as Pick's disease or frontal lobe dementia. It typically occurs at a younger age than other forms of dementia, with 60% of cases occurring in people aged 45 to 64 years old. FTD occurs as a consequence of selective degeneration within the frontal and temporal lobes. In the early stages of FTD individuals tend to display changes to their personality and behaviour and/or aphasia. Aphasia is when a person has difficulty with their language or speech and in FTD is usually caused by damage to the left temporal lobe. Compared to disorders such as AD patients with FTD tend to have good memory performance in the early stages although this does become progressively worse as the disease progresses.

Key Takeaways

- Dementia is a complex syndrome with different underlying causes and huge variability in symptom presentation
- Symptoms associated with dementia typically involve cognitive processes, in particular memory, but can also affect other behaviours, motor control, sleep and mood
- There are no cures for dementia which is progressive and symptoms worsen over time. However, some of the more common causes of dementia, i.e. AD, do have treatments which have been shown to be effective in slowing and stabilising the progression of symptoms
- Psychological approaches for dementia are also important as they have been demonstrated to improve wellbeing and quality of life not only for the patient but also for those involved in caring for dementia patients

- A huge amount of research is invested in further understanding the pathology underlying dementia and identifying novel treatment approaches.

References

Boyle, P.A., Wilson, R.S., Aggarwal, N.T., Tang, T., & Bennett, D. A. (2006). Mild cognitive impairment: Risk of Alzheimer disease and rate of cognitive decline. *Neurology*, *67*(3), 441-445. https://doi.org/10.1212/01.wnl.0000228244.10416.20

DeTure, M.A., & Dickson, D.W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration, 14(32), 1-18*. https://doi.org/10.1186/s13024-019-0333-5

Gottesman, R.T., & Stern, Y. (2019). Behavioral and psychiatric symptoms of dementia and rate of decline in Alzheimer's disease. *Frontiers in Pharmacology, 10*, 1062. https://doi.org/10.3389/fphar.2019.01062

Griciuc, A., & Tanzi, R.E. (2021). The role of innate immune genes in Alzheimer's disease. *Current Opinion in Neurology, 34*(2), 228-236. https://doi.org/10.1097/wco.0000000000000911

Holger, J. (2013). Memory loss in Alzheimer's disease. *Dialogues Clinical Neuroscience, 15(4)*, 445-454. https://doi.org/10.31887%2FDCNS.2013.15.4%2Fhjahn

Kim, J., Jeong, M., Stiles, W.R., & Choi, H.S. (2022). Neuroimaging modalities in Alzheimer's disease: Diagnosis and clinical features. *International Journal of Molecular Sciences, 23*(11), 6079. https://doi.org/10.3390/ijms23116079

Logsdon, R.G., McCurry, S.M., & Teri, L. (2007) Evidence-based interventions to improve quality of life for individuals with dementia. *Alzheimer's care today, 8*(4), 309-318. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2585781/

McKay, E., & Counts, S.E. (2017) Multi-infarct dementia: A historical perspective. *Dementia and Geriatric Cognitive Disorders Extra*, *7*, 160-171. https://doi.org/10.1159/000470836

Mijajlović, M.D., Pavlović, A., Brainin M., Heiss, W.-D., Quinn, T.J., & Ihle-Hansen, H.B. (2017) Post-stroke dementia – a comprehensive review. *BMC Medicine*, *15*(11), 1-12. https://doi.org/10.1186/s12916-017-0779-7

Office of National Statistics (2022, July 1). *Deaths Registered in England and Wales, 2021*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2021

Outeiro, T.F., Koss D.J., Erskine, D., Walker, L., Kurzawa-Akanbi, M., & Burn, D. (2019) Dementia with Lewy bodies: an update and outlook. *Molecular Neurodegeneration*, *14*(5), 1-18. https://doi.org/10.1186/s13024-019-0306-8

Prince, M., Knapp, M., Guerchet, M., McCrone, P., Prina, M., Comas-Herrera, M., Wittenberg, A., Adelaja, R., Hu, B., King, B., Rehill, D., & Salimkumar, D. (2014). *Dementia UK: Update*. Alzheimer's Society. http://www.alzheimers.org.uk/dementiauk

Terry, R.D., Masliah, E., Salmon, D.P., Butters, N., DeTeresa, R., Hill, R., Hansen, L.A., & Katzman, R. (1991). Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment. *Annals*

of Neurology, 30(4), 572–80. https://doi.org/10.1002/ana.410300410

Tomimoto, H. (2011). Subcortical vascular dementia. *Neuroscience Research, 71*(3), 193-199. https://doi.org/10.1016/j.neures.2011.07.1820

Woods, B., O'Philbin, L., Farrell, E.M., Spector, A.E., & Orrell, M. (2018) Reminiscence therapy for dementia. *Cochrane Database of Systematic Reviews*, *3*, CD001120. https://doi.org/10.1002%2F14651858.CD001120.pub3

About the authors

Professor Claire Gibson
UNIVERSITY OF NOTTINGHAM
https://twitter.com/PreclinStroke

Professor Claire Gibson obtained a BSc degree in Neuroscience from the University of Sheffield and her PhD from the University of Newcastle. She then gained a number of years' experience researching the mechanisms of injury following CNS damage – initially focusing on spinal cord injury and moving on later to cerebral stroke. She is now a Professor of Psychology at the University of Nottingham

whose research pursues the mechanisms of damage and investigates novel treatment approaches following CNS disorders, focusing primarily on stroke and neurodegeneration. She regularly teaches across the spectrum of biological psychology to both undergraduate and postgraduate students.

Dr Catherine Lawrence
UNIVERSITY OF MANCHESTER
https://braininflamelab.org/research
https://twitter.com/big_research

Dr Catherine Lawrence obtained a BSc degree in Pharmacology and her PhD from the University of Manchester. She then gained over two years' experience in the commercial sector working as a Clinical Research Associate in the pharmaceutical industry but, returned to academic research at the University of Manchester on a post-doctoral position (funded by AstraZeneca). In 2004 she secured a position as a Senior Research Scientist at AstraZeneca, but in 2005 was awarded an RCUK fellowship at the University of Manchester. In 2010 she became a lecturer and a senior lecturer in 2015. Her current research interests are Alzheimer's disease and stroke and in particular understanding how diet can influence these disorders and the involvement of inflammation.

20.

PLACEBOS: A PSYCHOLOGICAL AND BIOLOGICAL PERSPECTIVE

Professor Jose Prados and Professor Claire Gibson

Learning Objectives

To gain an understanding of the following:

- The definition of a placebo effect
- The biological and psychological mechanisms of the placebo effect
- The importance of placebos in clinical trial

design and their ethical considerations
- The contribution of placebos to our understanding of complex disorders i.e. pain, depression.

Definition of a placebo

The term **placebo**, derived from the Latin for *'I shall please',* is used in modern medicine to describe a dummy substance or other treatment that has no obvious or known direct physiological effect. The most common examples of a placebo include an inert tablet (e.g. sugar pill) or injection, via intramuscular or intravenous routes, of a control solution (typically saline), but can also include a surgical procedure. However, such inert treatments can have measurable effects and benefits in patient groups due to the context of their administration and expectation. Such effects are not limited to the individual's subjective evaluation of symptom relief but can include measurable physiological changes such as altered gastric secretion, blood vessel dilation and hormonal changes.

The placebo effect

A medical treatment/procedure is associated with a complex psychosocial context that might affect the outcome of the therapy (see Figure 6.15). To determine the effects of the psychosocial context on the patient it is necessary to eliminate the specific action of the treatment and to replicate the context of the treatment administration with administration of the active treatment itself. Thus, a placebo is given in which the patient believes they are receiving an effective therapy and therefore expects to experience its benefits such as symptom relief. The placebo effect, or response, is the outcome that follows this administration of a placebo. It is essential administration takes place within the design of a clinical trial (see insert box) to evaluate the potential effectiveness of new treatments and eliminate the influence of patient expectation on outcome, as drug effects may be influenced by the patient's history and beliefs/expectations about the drug/treatment being developed.

In part, the effect of a placebo may be explained as an outcome of classical conditioning. For example, in the case of pain relief (see insert box), if there is a history of an injection causing pain relief (e.g. morphine). Thus, by association, the syringe and the context of the injection can acquire some pain-relieving capacity i.e. an association between the procedure (conditional stimulus), the drug (unconditional stimulus) and the pain relief (unconditional response). However,

conditioning cannot fully explain the placebo effect in all scenarios – for example, if a person is told that pain-relief is to be expected there can be some tendency for it to be experienced. A wealth of neuroimaging and neurobiological studies report changes in brain activity and brain function following placebo administration, supporting the notion of a biological basis of the placebo effect.

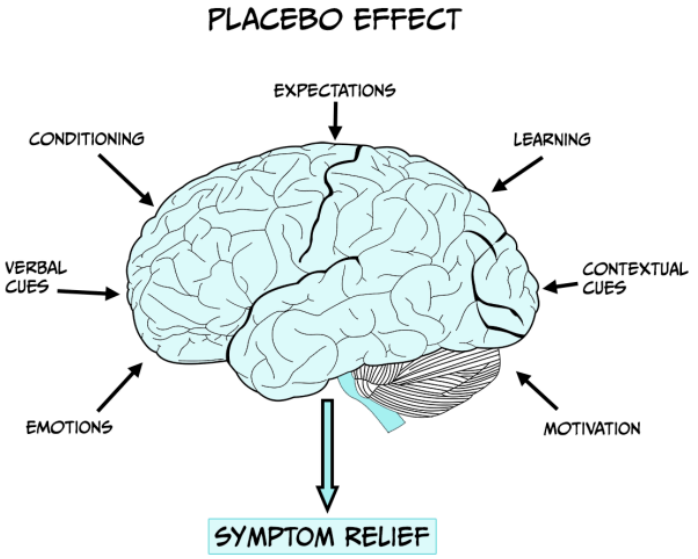

Figure 6.16. Psychological influences on the ability of a placebo to produce symptom relief

Pain

Pain is a highly complex and individual experience which results in behavioural, chemical, hormonal and neuronal responses. Humans may experience occasional pain which activates the autonomic, central, and peripheral nervous systems as well as chronic pain over a number of months and even years. Chronic pain substantially impacts the quality of life of affected individuals and there is a demand to develop new and effective therapies. Various treatment approaches for pain exist, including medicines, physical therapies (for example, heat/cold treatment, exercise, massage) and complementary therapies (for example, acupuncture and meditation). Placebo effects have been reported to act as pain relievers in certain groups of patients and may offer a viable therapeutic option (Miller & Colloca, 2009).

Functional brain imaging studies show that opioids and placebos activate the same brain regions and that both treatments reduce the activity of brain regions responding to pain, including the cingulate

cortex (Wager et al., 2004). A consistent finding is that some people experience relief from a placebo and others do not. People who respond to placebo show a greater activation of brain regions with opioid receptors than do non-responders, further implicating endogenous opioids in the placebo effect. Opioids have often been reported as inducing relaxation which may account for the feelings of pain relief following placebo treatment. However, there is also substantial evidence that placebos are able to alleviate pain through the reduction of negative emotions (i.e. feelings of fear and anxiety) associated with pain rather than acting to reduce the sensation of pain itself. For example, placebo treatment decreases activation of the cingulate cortex but not the somatosensory cortex. Similar to pain itself, the relief from pain symptoms is complex and placebos can play an important aspect in the therapeutic approach to treat pain in certain individuals.

Mechanisms: psychological mechanisms

From the psychological perspective, the placebo effect has traditionally been attributed either to conscious cognition—for example, the expectations of the patients—or the action of automatic basic learning mechanisms like classical or Pavlovian conditioning. The evidence accumulated over recent decades suggests that conscious cognition and conditioning shape different instances of the placebo effect, and that they can interact to determine the effect (e.g., Stewart-Williams & Podd, 2004). Here, we explore two versions of the conscious cognition approach, the most prevalent Expectancy Theory, and a promising approach that characterises some instances of the placebo effect as a particular type of error in decision making. We will then explore how conditioning accounts for the placebo effect, by reference to the research done with non-human animals and how it translates to clinical practice in humans.

Conscious cognition: expectancy theory

Fig 6.17. Placebo medication can produce symptom alleviation

A placebo produces an effect because the patient expects it to produce such effect. The expectancy account considers several factors known to shape the recipient's expectations, including the therapeutic relationship and the authority of the professional that administers the placebo. Other factors known to contribute to the development of expectancies include the branding and cost of the medication. For example, the use of a placebo was more effective in reducing headache when the use of brand name was used to label the tablets than when a generic label was used; also, fewer side effects were attributed to tablets with the brand name (Faase et al., 2016). Similarly, the colour of the pills can also contribute to shape the expectancies of the recipient: red and orange are

associated with a stimulant effect, while blue and green tend to be associated with sedative effects (de Craen et al., 1996).

The key question from this perspective is how expectancies contribute to the placebo effect. Different mechanisms have been proposed. Lundh (2000) suggested that positive expectancies contribute to reduce the anxiety of the placebo recipient. It is well established that stress and anxiety have an adverse effect in a diversity of physiological processes and increase the number and intensity of the symptoms reported by the patients. The use of placebos, by reducing anxiety levels, can contribute to easing symptomatology (see Stewart-Williams & Podd, 2004). Expectancies can also contribute to the placebo effect by changing other cognitions: the placebo-induced expectancy of improvement, by promoting a sense of control, may enable the recipient to face pain more positively; the patient may be more likely to disregard negative thoughts and interpret ambiguous stimuli more favourably. Another way in which positive expectancies can mediate the placebo effect is by changing the actual behaviour of the recipient: the expectation of an improved condition may lead the patient to resume their daily routines which would improve the mood and distract them from the symptoms reducing the pain experience (Peck & Coleman, 1991; Turner et al., 1994).

Conscious cognition: decision making

It is suggested that patients treated either with an active therapeutic agent or a placebo are left with a binary decision: was the symptom alleviated or not (Allan & Siegel, 2002)? This might be a tricky question in some situations where the symptom, for example periodic pain, emerges from the brain's interpretation of the input received from sensory receptors and, as discussed above, is mediated by psychological factors. The relative intensity of pain would fluctuate over time depending on whether the patient is distracted or fully focused on the symptom, for example. To decide whether an improvement is experienced or not, patients need to consider whether the average pain intensity has decreased. Perhaps it has, or perhaps the level of pain is similar to what they felt before treatment was administered. In judging the relative intensity of the sensation, the patient is facing an ambiguous situation. The reduction of symptomatology is a signal presented against a noisy environment (the changing intensity of symptoms over time). We can apply the principles of the Signal Detection Theory (SDT, Tanner & Swets, 1954) to characterise this instance of the placebo effect. We can summarise all the possible outcomes by reference to 'The Patient's Decision Problem' (see Table 1).

The outcome would depend on the criterion used, which can be liberal (any change would be identified as the signal,

and therefore the patient is likely to incur a False Positive and experience alleviation) or conservative (the signal will not be easily detected, and the patient is likely to incur a False Rejection—experiencing the absence of effect). The adoption of a liberal or conservative criterion would depend on the perceived consequences of each of the possible errors: high risk of false rejections leads to a liberal criterion; high risk following a false positive contributes to the adoption of a more conservative criterion. In the clinical context, a false rejection could be equivalent to claiming that an effective, tested drug is non-effective. This would challenge the accepted wisdom as well as the authority of the physician that administers the treatment. To avoid this potentially embarrassing situation, the recipient might adopt a liberal criterion, which increases the probability of a false positive. In many instances, a patient given a placebo would rather make the decision or 'mistake' that is deferential to the established wisdom (the science) and pleases the doctor and their family, experiencing relief of their symptoms in the absence of an active therapeutic agent. This would lead to an instance of the placebo effect.

	Active Agent	Placebo
Alleviation	Correct Positive	***False Positive***
No effect	***False Rejection***	Correct Rejection

Table 1. The Patient's Decision Problem

Treated with an 'active agent' or a 'placebo', the patient can experience 'alleviation' of the symptom, or 'no effect'. The outcome can be termed a ***Correct Positive*** when an active agent has been administered and the patient experiences alleviation. Similarly, in the absence of an active therapeutic agent (placebo), the absence of effect would be a ***Correct Rejection***. In this situation, the patients can also make mistakes; failure to experience alleviation when treated with a therapeutic agent would produce a ***False Rejection***; on the other hand, experiencing alleviation of the symptoms in the absence of an active agent (treatment with a placebo) would be a ***False Positive***. Incurring a False Positive, a common and in some cases a desirable mistake, constitutes the placebo effect.

Conditioning: learning-mediated placebo

In this section, we focus on a different instance of the placebo effect that emerges when pairing two events: the situational cues where a treatment is taking place (physical context, the form of administration, the health worker that administers the treatment, etc.) and the active agent that has a therapeutic effect. The therapeutic effect is an automatic, unconditioned response to the active agent or drug. Repeated experience of the drug in the presence of the situational cues promotes the development of Pavlovian conditioning, whereby the

situational cues acquire the capacity to elicit a conditioned therapeutic response: in the presence of the situational cues, even in the absence of the active agent, the patient will experience alleviation of the symptoms. This therapeutic conditioned response to the situational cues is an instance of the placebo effect that can be used to reduce the dose of the active agent (especially relevant for drugs with undesirable side effects) in different contexts. We will briefly describe a couple of examples of the use of the conditioned placebo effect in the treatment of auto-immune diseases and the treatment of pain.

The conditioning of the pharmacological effects of drugs has a long history. Pavlov (1927, p. 35) described early experiments by Krylov in which dogs were repeatedly injected with morphine, which produces nausea, salivation, vomiting and sleep. After 5 or 6 injections of morphine in a particular experimental setting, the preliminaries of the injection sufficed to produce all these symptoms in response, not to the effect of the drug in the blood stream, but of the exposure to the external stimuli that previously preceded the morphine injection. Conditioned pharmacological responses have been successfully used in humans subject to immunodepression treatment. Giang et al. (1996) treated 10 patients of multiple sclerosis (MS) with cyclophosphamide, an effective immunosuppressant that helps control the symptoms of MS but has serious side effects (e.g., increased risks of infection and cardiovascular disease and depletion of the bone marrow). The participants ingested an anise-flavoured syrup prior to each

administration of the immunosuppressive drug. Later, they were given the syrup with a small, ineffective dose of the drug. Eight of the ten participants displayed a clear conditioned immunosuppressive response, suggesting that it is possible to reduce the dose of the drug administered during the treatment to keep the side effects at bay.

Conditioned pharmacological responses can also be used in the treatment of pain. Opioids are used to block pain signals between the brain and the body and are typically prescribed to treat moderate to severe pain. However, a second set of responses (undesirable side effects) are activated that contribute to the development of tolerance, which reduces the effectiveness of the drug requiring increased doses to achieve the desired therapeutic effect. Used in high doses, opioids can lead to the development of addiction and of opioid induced hyperalgesia (OIH) that worsen the patients' wellbeing (e.g., Holtman, 2012). When an analgesic drug (like an opioid) is administered to an individual in pain the drug results in a reduction of pain, a therapeutic effect which is highly rewarding. Repeated presentations of the active therapeutic agent in a particular context would allow the context to activate a conditioned therapeutic response that reduces pain in the absence of the active therapeutic agent. This would potentially help reducing the dose of the drugs used to treat pain keeping opioids effective at low doses without side effects.

Persuasive evidence has been presented for the development of conditioned analgesia in mice. Guo et al. (2010), treated

mice with either morphine or aspirin before placing them on a hotplate. Animals exposed to the hotplate at 55 °C display a paw withdrawal response; treatment with an analgesic significantly delays the paw withdrawal response—evidence of the analgesic properties of the drug. Following training with either morphine or aspirin, the animals showed evidence of a conditioned analgesic response by delaying the paw withdrawal response when they were exposed to the hot plate following the injection of a saline solution—an instance of the conditioning mediated placebo effect. Interestingly, when animals were treated with an opioid antagonist (naloxone) the conditioned analgesia disappeared in the animals initially treated with morphine, but not in the animals treated with aspirin. This is consistent with the observation that, in humans, placebo analgesia is associated with the release of endogenous opioids (Eippert et al., 2009) indicating the importance of opioidergic signalling in pain-modulating and the placebo effect. It is worth mentioning that the psycho- and pharmaco-dynamics of opioids is very complex and not yet fully understood. In some cases, pairing situational cues with opioids can lead to the development of a conditioned response which is opposed to the desired therapeutic response (a conditioned hyperalgesia response; see Siegel, 2002, for a full review). The development of conditioned hyperalgesia is beyond the remit of this chapter, but the reader should be aware of the need to identify the parameters that promote the development of therapeutic conditioned responses and

prevent the development of conditioned hyperalgesia that could worsen the condition of patients in clinical settings.

Biological mechanisms of the placebo effect

In order to explain the changes seen in the function of certain brain areas following placebo administration, a biological mechanism of action must exist. In terms of placebo effects these are typically described as occurring via opioid or non-opioid mechanisms. The role of opioids in the placebo effect was established by the observation that under some conditions the effect is abolished by prior injection of the opioid antagonist naloxone. In the placebo effect, dopamine and opioids are activated in various brain regions (e.g. nucleus accumbens) corresponding to the expectation of beneficial effects. Comparing different people, high placebo responsiveness is associated with high activation of these neurochemicals. Opioid receptors are found in regions of the pain neuromatrix which are reduced in activation corresponding to the placebo effect e.g. anterior cingulate cortex and the insula (Kim et al., 2021). Opioid mediated placebo responses also extend beyond pain pathways. It is reported that placebo-induced respiratory depression (a conditioned placebo side effect) and decreased heart rate and β-adrenergic activity can be reversed by naloxone, demonstrating

the involvement of opioid mechanisms on other physiological processes, such as respiratory and cardiovascular function.

However, the opioid system is not the only pathway involved in the placebo effect. Placebo administration also increases the release and uptake of dopamine and dopamine receptors are activated in anticipation of benefit when a placebo is administered. This suggests the dopamine system may underlie the expectation of reward following placebo administration (Scott et al., 2008). In addition, placebo effects that are non-opioid mediated can be blocked by the cannabinoid receptor antagonist CB1 (Benedetti et al., 2011) suggesting a role of the endocannabinoid system. Genetics are also reported to have a part in the biological explanation of the placebo effect in that they can influence the strength of the effect. For example, patients with opioid receptors that are less active are less likely to be placebo responders whereas patients with reduced dopamine metabolism, and therefore higher dopamine levels in the brain, are more likely to experience a strong placebo effect (Hall et al., 2015). Placebo treatments can also affect hormonal responses that are mediated via forebrain control of the hypothalamus-pituitary-hormone system.

Although other medical conditions have been investigated from a neurobiological perspective, the placebo mechanisms in these conditions are not as well understood compared to pain and analgesia. For example, placebo administration to Parkinson patients induces dopamine release in the striatum,

and changes in basal ganglia and thalamic neuron firing. In addition, changes occur in metabolic activity in the brain following placebo administration in depression (see insert box) and following expectation manipulations in addiction.

Depression

Placebo effects in clinical trials exploring potential therapies for the treatment of depression are extensively reported, with many trials failing to report a significant benefit of a novel therapeutic treatment compared to that seen in the placebo group. This can be attributable to positive benefits of the placebo treatment rather than simply being due to an ineffective treatment. In fact it has been reported that in clinical trials for major depression approximately 25% of the benefit reported by patients is due to the active medication, 25% due to other factors such as spontaneous remission of symptoms and 50% is due to the placebo effect.

Insight originally gained from pain studies has helped to reveal how the endogenous opioid system,

important in regulating the stress response and emotional regulation, is an important mediator for placebo. As this system is dysregulated in depression it is plausible that opioids are responsible for mediating the placebo effect seen in depression. Studies have shown that individuals with higher opioid receptor activity in areas of the brain such as the anterior cingulate cortex, nucleus accumbens and amygdala, all areas implicated in emotion stress regulation and depression, are more likely to experience anti-depressive symptoms following placebo treatment (Zubieta et al., 2005).

Ethics and the nocebo effect

In clinical trials the placebo is essential to the design of experiments evaluating the effectiveness of new medications because it eliminates the influence of expectation on the part of the patient. This control group is identical to the experimental group in all ways, yet the patients, and medical staff administrating treatment, are blinded to whether they are receiving active or placebo treatment. Such precautions ensure that the results of any given treatment will not be influenced by overt or covert prejudices on the part of the patient or the observer. The assumption is that to truly examine the potential

biological effects of a treatment and exclude the influence of the psychosocial context the patient must be deceived as to whether they are receiving an active treatment or placebo.

In terms of the placebo response an individual can often be characterised as a responder or non-responder. It is important to consider the ethical implications of this characterisation. For example, it may be appropriate to consider targeting responders with placebo treatments that result in a positive response for them whereas it may also be appropriate to consider excluding responders from clinical trials to ensure results are not compromised. The design of clinical trials is important (see insert box) – if, for example, a new treatment is found to be effective during a clinical trial then it would be considered unethical to deny any participants of that trial access to the effective treatment. Thus, clinical trials are often designed as blocks where patients receiving alternating blocks of active treatment and placebo to ensure all patients have equal chance of receiving benefit of a new treatment.

A less well understood phenomena is the nocebo effect, in which negative expectations of a treatment decrease the therapeutic effect experienced or increase experience of side effects. Administration of the peptide cholecystokinin (CCK) has been shown to play a role in nocebo hyperalgesia through inducing anticipatory anxiety mechanisms, while blocking CCK reduces nocebo effects (Benedetti et al, 1995; https://pubmed.ncbi.nlm.nih.gov/9211474/). A deactivation of dopamine has been found in the nucleus accumbens during

nocebo hyperalgesia and brain imaging studies have demonstrated activation of brain areas, different to those activated during a placebo effect, including the hippocampus and regions involved with anticipatory anxiety (Finniss & Fabrizio, 2005).

Placebo role in clinical trial design

The 'discovery' of a new drug or treatment usually occurs in one of three ways; the rediscovery of usage of naturally occurring products, the accidental observation of an unexpected drug effect or the synthesizing of known or novel compounds. In all cases a substance must progress through various stages in order to meet the licencing arrangements within the relevant country to allow that treatment to be approved and subsequently marketed. The initial stages of drug/treatment development tend to involve extensive synthesis (if relevant) of the drug, chemical characterisation and a series of preclinical or animal studies to establish the potential effectiveness and/or safety of the treatment. Drugs/

treatments deemed worthy of clinical investigation progress through three phases of clinical trial and it is important to consider the role of placebo within the design of a clinical trial:

- **Phase I** involves healthy volunteers and aims to specify the human reactions, in terms of physiology and biochemistry, to a drug along with determining safety.
- **Phase II** involves patients with the disorder the new drug/treatment is targeting and are aimed at determining the effectiveness of such a drug.
- **Phase III** expands Phase II by increasing the number of patients in the trial. These trials are typically less well controlled than phase II as they tend to occur across multiple sites and even multiple countries.

Clinical trials normally occur by randomly allocating patients into treatment groups which may vary in terms of the dose received and whether the patient is receiving active or placebo treatments. Such trials are termed randomised controlled trials (RCTs). A double-blind study is one in which neither patient or medical staff knows into which group (i.e. active

treatment or placebo) a patient has been allocated. Treatments and placebos are made to look identical and coded to obscure their identity. Both the subject's and experimenter's expectancies may influence the effects of the drug that the subject experiences. Whereas the simplest design is to assign patients either active or placebo treatment, due to ethical considerations, most trials are based on a block design with each subject receiving blocks of active or placebo treatment. Such designs can help unpick placebo from treatment effects, however, as a clinical trial progresses the observed response in the placebo group may occur due to other factors such as natural course of the disease and fluctuations of symptoms, making it harder to discern a genuine placebo response.

Conclusions

Strong evidence supports the notion that placebo effects are real and that they may even have therapeutic potential. Placebo effects are mediated via diverse processes which can include learning, expectations and social cognition and are mediated via biological mechanisms. It is important to consider the contribution of placebo effects in the design and

interpretation of clinical trials. Placebos may have meaningful therapeutic effect and should continue to be studied to fully understand their potential.

Key Takeaways

- Whilst placebos do not contain an active substance to produce a biological effect, they can produce a response. Thus, they are important to consider in the design of clinical trials to determine the true effect of a biologically active drug or treatment.
- The placebo effect can be psychological or physiological in nature and can be observed in humans (typically in medical settings) and in non-human animals (typically in a research context). For example, pharmacological conditioning elicits strong placebo effects both in humans (e.g., Amanzio & Benedetti, 1999; Olness & Ader, 1992) and animals (e.g., mice; see Guo et al., 2010).
- The placebo effect has been extensively

researched and the picture that emerges suggests there is not a single placebo response but many, with different mechanisms at work across a variety of medical conditions, interventions, and systems (see Benedetti, 2008, for a full review).

References and Further Reading

Ader, R. (1997). The role of conditioning in pharmacotherapy. In A. Harrington (Ed.), *The Placebo Effect (*pp. 138-165). Harvard University Press.

Allan, L.G., & Siegel, S. (2002). A signal detection theory analysis of the placebo effect. *Evaluation & the Health Professions*, *25*(4), 410-420. https://doi.org/10.1177/0163278702238054

Amanzio, M., & Benedetti, F. (1999). Neuropharmacological dissection of placebo analgesia: expectation-activated opioid systems versus conditioning-activated specific

subsystems. *Journal of Neuroscience*, *19*(1), 484-494. https://doi.org/10.1523%2FJNEUROSCI.19-01-00484.1999

Benedetti, F., Amanzio, M., Casadio, C., Oliaro, A,. & Maggi, G. (1997). Blockade of nocebo hyperalgesia by the cholecystokinin antagonist proglumide. *Pain, 71*(2), 135-40. https://doi.org/10.1016/s0304-3959(97)03346-0

Benedetti, F. (2008). *Placebo effects: understanding the mechanisms in health and disease.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199559121.001.0001

Benedetti, F., Amanzio, M., Rosato, R., & Blanchard, C. (2011). Nonopioid placebo analgesia is mediated by CB1 cannabinoid receptors. *Nature Medicine, 17*, 1228-1230. https://doi.org/10.1038/nm.2435

Colloca, L., Benedetti, F. (2005). Placebos and painkillers: Is mind as real as matter? *Nature Reviews Neuroscience, 6*, 545-552. https://doi.org/10.1038/nrn1705

De Craen, A.J., Roos, P.J., De Vries, A.L., & Kleijnen, J. (1996). Effect of colour of drugs: Systematic review of perceived effect of drugs and of their effectiveness. *British Medical Journal*, *313*(7072), 1624-1626. https://doi.org/10.1136/bmj.313.7072.1624

De Hower, J. (2018). A functional-cognitive perspective of the relation between conditioning and placebo research. *International Review of Neurobiology*, *138*, 95-111. https://doi.org/10.1016/bs.irn.2018.01.007

Faasse, K., Martin, L.R., Grey, A., Gamble, G., & Petrie, K.J. (2016). Impact of brand or generic labeling on medication effectiveness and side effects. *Health Psychology 35*(2), 187-90. https://doi.org/10.1037/hea0000282

Finniss, D.G., Fabrizio, B. (2005). Mechanisms of the placebo response and their impact on clinical trials and clinical practice. *Pain, 114*(1-2), 3-6. https://doi.org/10.1016/j.pain.2004.12.012

Giang, D.W., Goodman, A.D., Schiffer, R.B., Mattson, D.H., Petrie, M., Cohen, N., & Ader, R. (1996). Conditioning of cyclophosphamide-induced leukopenia in humans. *Journal of Neuropsychiatry and Clinical Neurosciences*, *8*(2), 194-201. https://doi.org/10.1176/jnp.8.2.194

Guo, J.Y., Wang, J.Y., & Luo, F. (2010). Dissection of placebo analgesia in mice: The conditions for activation of opioid and non-opioid systems. *Journal of Psychopharmacology*, *24*(10), 1561-1567. https://doi.org/10.1177/0269881109104848

Hall, K.T., Loscalzo, J., Kaptchuk, T.J. (2015). Genetics and the placebo effect: The placebome. *Trends in Molecular Medicine, 21*,(5) 285-294. https://doi.org/10.1016/j.molmed.2015.02.009

Holtman Jr, J.R., & Jellish, W.S. (2012). Opioid-induced hyperalgesia and burn pain. *Journal of Burn Care & Research, 33*(6), 692-701. https://doi.org/10.1097/bcr.0b013e31825adcb0

Kim, D., Chae, Y., Park, H.-J., & Lee, I.-S. (2021). Effects of

chronic pain treatment on altered functional and metabolic activities in the brain: A systematic review and meta-analysis of functional neuroimaging studies. *Frontiers in Neuroscience, 15*, https://doi.org/10.3389/fnins.2021.684926

Lundh, L.G. (2000). Suggestion, suggestibility, and the placebo effect. *Hypnosis International Monographs*, *4*, 71-90.

Miller, F.G., & Colloca, L. (2009). The legitimacy of placebo treatments in clinical practice: Evidence and ethics. *American Journal of Bioethics, 9*(12), 39-47. https://doi.org/10.1080/15265160903316263

Olness, K., & Ader, R. (1992). Conditioning as an adjunct in the pharmacotherapy of lupus erythematosus: A case report. *Journal of Developmental and Behavioral Pediatrics*, *13*(2), 124-125. https://doi.org/10.1097/00004703-199204000-00008

Peck, C., & Coleman, G. (1991). Implications of placebo theory for clinical research and practice in pain management. *Theoretical Medicine*, *12*(3), 247-270. https://doi.org/10.1007/bf00489609

Scott, D.J., Stohler, C.S., Egnatuk, C.M., Wang, H., Koppe, R.A., & Zubieta, J.K. (2008) Placebo and nocebo effects are defined by opposite opioid and dopaminergic responses. *Archives of General Psychiatry, 65*(2), 22-231. https://doi.org/10.1001/archgenpsychiatry.2007.34

Siegel, S. (2002). Explanatory mechanisms for placebo effects:

Pavlovian conditioning. In H. A. Guess, A. Kleinman, J. W. Kusek and L. W. Engel (Eds.), *The science of the placebo: Toward an interdisciplinary research agenda* (pp. 133-157). BMJ Books.

Stewart-Williams, S., & Podd, J. (2004). The placebo effect: Dissolving the expectancy versus conditioning debate. *Psychological Bulletin*, *130*(2), 324-340. https://doi.org/10.1037/0033-2909.130.2.324

Tanner Jr, W.P., & Swets, J.A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401-409. https://doi.org/10.1037/h0058700

Turner, J.A., Deyo, R.A., Loeser, J.D., Von Korff, M., & Fordyce, W.E. (1994). The importance of placebo effects in pain treatment and research. *The Journal of the American Medical Association*, *271*(20), 1609-1614. https://doi.org/10.1001/jama.1994.03510440069036

Wager, T.D., Billing, J.K., Smith, E.E., Sokolik, A., Casey, K.L., Davidson, R.J., Kosslyn, S.M., Rose, R.M., & Cohen, J.D. (2004). Placebo–induced changes in fMRI in the anticipation and experience of pain. *Science, 303*(5661), 1162–1166. https://doi.org/10.1126/science.1093065

Wager, T.D., Atlas, L.Y. (2015). The neuroscience of placebo effects: Connecting context, learning and health. *Nature Reviews Neuroscience, 16*, 403-418. https://doi.org/10.1038/nrn3976

Zubieta, J.-K., Bueller, J.A., Jackson, L.R., Scott, D.J., Xu, Y., Koeppe, R.A., Nichols, T.E., & Stohler, C.S. (2005).

Placebo effects mediated by endogenous opioid activity on μ-opioid receptors. *Journal of Neuroscience, 25*(34), 7754-7762. https://doi.org/10.1523/jneurosci.0439-05.2005

About the authors

Professor Jose Prados
UNIVERSITY OF DERBY

Professor Jose Prados has a PhD in Psychology from the University of Barcelona (Spain) and has been working in Higher Education for more than twenty-five years in a diversity of institutions. He is now a Professor of Psychology at the University of Derby. His primary research interests concern learning and memory from a comparative and evolutionary perspective. He uses Pavlovian and instrumental tasks to ascertain whether animals from different phyla (vertebrates and invertebrates like snails or flatworms) learn according to the same principles and test the power of associative learning theory to explain abilities traditionally considered out of its scope (e.g., navigation; perceptual learning).

Professor Claire Gibson
UNIVERSITY OF NOTTINGHAM
https://twitter.com/PreclinStroke

Professor Claire Gibson obtained a BSc degree in Neuroscience from the University of Sheffield and her PhD from the University of Newcastle. She then gained a number of years' experience researching the mechanisms of injury following CNS damage – initially focusing on spinal cord injury and moving on later to cerebral stroke. She is now a Professor of Psychology at the University of Nottingham whose research pursues the mechanisms of damage and investigates novel treatment approaches following CNS disorders, focusing primarily on stroke and neurodegeneration. She regularly teaches across the spectrum of biological psychology to both undergraduate and postgraduate students.