Published in M. Massimi (Ed.) *Philosophy and the Sciences for Everyone* (2015), London: Routledge, pp. 86–102

peg.series@gmail.com
mark.sprevak@ed.ac.uk

From intelligent machines to the human brain

Peggy Seriès
University of Edinburgh

Mark Sprevak
University of Edinburgh

6 April 2016

This chapter introduces the idea that computation is a key tool that can help us understand how the human brain works. Recent years have seen a revolution in the kinds of tasks computers can perform. Underlying these advances is the burgeoning field of machine learning, a branch of artificial intelligence, which aims at creating machines that can act without being programmed, learning from data and experience. Rather startlingly, it turns out that the same methods that allow us to make intelligent machines also appear to hold the key to explaining how our brains work. In this chapter, we explore this exciting new field and some of the questions that it raises.

1 Introduction

This chapter introduces the idea that computation is a key tool that can help us understand how the human brain works. Recent years have seen a revolution in the kinds of tasks computers can perform. Underlying these advances is the burgeoning field of machine learning, a branch of artificial intelligence, which aims at creating machines that can act without being programmed, learning from data and experience. Rather startlingly, it turns out that the same methods that allow us to make intelligent machines also appear to hold the key to explaining how our brains work. In this chapter, we explore this exciting new field and some of the questions that it raises.

2 Computations in our head

Intelligent machines are growing in number and complexity around us. Machines search the vast space of the internet to find the precise piece of information we want. Machines read our email to sniff out spam, conversations that are important to us, or possible criminal activity. Machines guess our desires when we shop online, often anticipating what we want before we know it ourselves. Machines recognize human speech and, at least in certain circumstances, offer sensible answers back. Machines pilot aircraft, drive cars, plan missions to space, and explore other planets. Machines predict patterns in the stock market and instigate the movement of trillions of dollars worldwide. Machines read our medical scans and histories to detect the early signs of cancer, heart disease, and stroke.

These are no mean tasks. In many cases, a human would struggle to do at least as well as our best machines. All these tasks have one thing in common: they require intelligent behaviour. They need a machine to follow rules, recognize and generalize patterns, and react rapidly and rationally to new information. Intelligent machines are able to do this courtesy of a brilliant idea: computation. Computation involves solving a problem by following a recipe, or set of instructions. This recipe is called an algorithm. An algorithm tells a machine how to accomplish its task by taking a series of basic steps.

The steps are often simple – for example, adding 1 to a digit, or checking if two digits are the same. In a computation, many simple steps are strung together to achieve complex behaviour.

Alan Turing (1912–54) was one of the originators of our modern notion of computation. Turing was an English mathematician who was obsessed with the idea of creating an intelligent machine. Turing discovered what he believed to be the key to this in one of his mathematical papers, 'On Computable Numbers' (1936/7). In this paper, Turing introduced the idea of a *universal computing machine*. A universal computing machine is a machine that, if given the right instructions, can take on the work of any other computer. The idea that a single machine could replace every other computer sounds, on the face of it, incredible. One might imagine that a universal computing machine would be mind-bogglingly complex if it existed at all. Turing showed a remarkable result: it is relatively easy to build a universal computing machine. He described a machine known as the Universal Turing Machine. The Universal Turing Machine consists of a paper tape and a mechanical head which can read and write marks on the paper tape guided by simple instructions. A Turing machine can reproduce the behaviour, no matter how complex, of any other computing machine. Given the right algorithm – the right instructions – the Universal Turing Machine can solve any task that any other computing machine can

solve. Therefore, creating an intelligent machine should just be a matter of hitting on the right algorithm. In the hands of John von Neumann, Max Newman, John Mauchly, Presper Eckert and others, Turing's idea of a universal machine gave birth to the first generation of general-purpose electronic computers. Today, universal computing machines surround us in the form of PCs, smartphones, and tablets.

The Universal Turing machine is one way to create a universal computing machine, but it is not the only way. After Turing's initial insight, a huge number of different universal machines have been discovered; some are exotic, others mundane. Universal machines are classified into different computational architectures, which include register machines, connectionist networks, quantum computers, DNA computers, and chemical reaction-diffusion computers. These devices work in different ways; they have different methods for producing their behaviour. But they share the same universal property: each can solve any problem that is solvable by a computing machine. One might compare the choice between them to that between methods of transport for getting you from A to B: walking, bicycle, scooter, car, helicopter. Each has the same overall effect: getting you from A to B, given enough time, patience, and money. But some methods make reaching your destination quicker, cheaper, or more convenient than others. There is no computational architecture that is universally 'best'. Some computational architectures are more suited to solving certain problems – their algorithms enable a machine to solve the problem more quickly, cheaply, and conveniently – than others.

Much of the work in the project of creating intelligent machines has focused on which architecture is most suited to solve the task of producing intelligent behaviour. Producing intelligent behaviour is a hard problem no matter which architecture one chooses, but some architectures make that task easier than others. For many years, attention focused on simple rule-based systems not unlike Turing's original machines. In the 1980s, attention shifted to algorithms that do not involve the manipulation of language-like symbols, but instead manipulate distributed patterns of activity in networks of simple nodes inspired by the brain ('connectionist networks'; see Clark 2014, chs 2 and 4). Nowadays, the algorithms that hold most promise for producing intelligent behaviour are those that involve probabilistic representations. These algorithms are characterized by representing not just a range of outcomes, but also the system's *uncertainty* about those outcomes. For example, a computer would not only represent that there is a tiger lurking around the corner, it would also store how probable it thinks this is. The great virtue of probabilistic representations is that they allow a computer to learn easily by following a principle called Bayes' theorem. Bayes' theorem tells a machine how to update its stored knowledge in the light of new data, and how to make intelligent predictions based on its knowledge. As we will see below, this kind of probabilistic reasoning is thought to lie at the heart of many, if not all, cases of intelligent human behaviour.

The story of machine intelligence, including its most recent probabilistic incarnation, is rich and complex (see Russell and Norvig 2010, ch. 1). We are going to put this story to one side to focus on perhaps an even more significant development that has run in parallel. From the earliest days, computation has also suggested a new way of thinking about ourselves. If computation solves the problem of generating intelligent behaviour for machines, then perhaps we humans work in the same way? This idea – that computation not only explains machine intelligence, but also human intelligence – is called the computational theory of mind. The central premise of the computational theory of mind is that intelligent human behaviour is generated by, and should be explained in terms of, computations performed by our brains. In recent years, the computational theory of mind has revolutionized thinking in psychology and neuroscience. Non-computational approaches to the mind exist, and computation may not be able to explain every aspect of our mental lives (conscious feelings are a particularly hard case). Nevertheless, there is widespread agreement that computation provides the most promising story about how humans produce intelligent behaviour.

3 Levels upon levels

Let's try to unpack the idea that computation could help explain intelligent behaviour. Precisely how might computation enable us to do this? In the 1970s, a brilliant young cognitive scientist, David Marr, answered this question. Marr disentangled three ways in which a computational approach could help us understand the brain. Marr asked *which*, *how*, and *why* questions. *Which* computational task does the brain solve? *How* does the brain solve that task? *Why* is this task important for the brain to solve? Marr grouped these questions together in a slightly unusual way. He began by saying that a computational theory should explain how the brain produces intelligent behaviour on one of three different levels.

Marr's first level is the computational level. This level describes *which* computational problem the brain solves and *why* it is important. Imagine that one day you discover a mysterious device in granny's attic. The mysterious device has many flashing lights, buttons, and dials, all of unknown purpose. You recall that granny used the device when she was balancing her cheque book. You play around with the device and you notice a pattern among its lights and dials: if you dial two numbers into the machine, its lights flash out a pattern that could be understood as their sum. Balancing a cheque book requires summing numbers. Therefore, you conjecture, the computational problem that granny's device solves is computing the *addition function*. In Marr's terminology, this is a computational level description of granny's device: a specification of *which* function (addition, subtraction, multiplication, etc.) the device computes. Arriving at this description, as we have just seen, is bound up

with answering a *why* question about the device: why – for what ends – did granny use the device? Without a guess about a device's intended purpose (e.g. balancing a cheque book), there would be no way of picking out of the vast number of things the device does (its many flashing lights and dials) which are relevant to solving its problem. This is why Marr groups the *which* and *why* questions together.

Marr's second level is the algorithmic level. This level concerns *how* a device solves its computational task. An algorithm is a recipe, or set of instructions, which tells the device how to solve its computational task. Many algorithms compute the addition function. Without further investigation, all we know is that granny's device is using one of them. It may be using, for example, a *column-addition* algorithm. This addition algorithm, which many of us learn in school, requires one to work out sums from right to left: first add the units, then the tens, then the hundreds, carrying to the next column when necessary. Alternatively, granny's device may be using a *partial-sums* algorithm. This algorithm requires one to work out the sum from left to right, storing intermediate 'partial sums' along the way: first add the hundreds treating other digits as zero, then add the tens, then the units, then add the three partial sums to get the answer. Different algorithms involve taking different basic steps, or taking those steps in a different order. Some addition algorithms are faster to run than others; some require less memory. All solve the same problem – all calculate the addition function – but some do it quicker, cheaper, or more conveniently given certain resources than others. The algorithm that a device uses is tied to that device's system of *representation*. Column addition only works if a device uses a *positional* system for representing numbers, such as our Arabic decimal numeral system. If granny's device were instead to use Roman numerals to represent numbers, it could not use the column-addition algorithm.

How do we know which algorithm granny's device uses? A first step would be to look at the resources granny's device has: which basic steps can it take, how much memory does it have, how fast can it execute a single step? Once we know the basic ingredients, we can work out which algorithm it uses. Probe granny's device by giving it a large range of addition problems. Measure how fast it solves addition problems and which kinds of errors it is prone to make. Its distinctive performance profile – speed and susceptibility to errors across different problems – will reveal how it combines basic instructions into a particular algorithm.

Marr's third level is the implementation level. Suppose we are satisfied that granny's device computes the addition function using the partial-sums algorithm. We still do not know how the nuts and bolts inside granny's device correspond to steps in the algorithm. Imagine we open up granny's device. Inside we might find different things. We might find gears and cogwheels, electronic components, little pens and pieces of paper, or perhaps a complex and confusing jumble of all three. Marr's

implementation level description describes how a device's physical hardware maps onto steps in its algorithm. An implementation-level description pinpoints which parts of the machine are *functionally significant*: which physical parts are relevant, and in what way, to running the algorithm. In an electronic PC, components that are functionally significant include electrical wires and silicon chips: these implement steps in the PC's algorithms. In contrast, the colour of the circuit boards, or the noise the cooling fan makes, are not functionally significant.

How do we give an implementation-level description of a device? One strategy would be to watch the device's workings in action. We could observe physical changes inside granny's device when we give it an addition problem and try to infer how those physical changes map onto steps in the algorithm. Or, we might actively intervene on the physical components inside granny's device – for example, by changing its wiring – and see how this affects its performance. If we damage or replace a physical component, how does that affect the device's speed or susceptibility to errors? By using a combination of these two strategies, we can arrive at a description of the role that each physical component plays.

Marr's three levels are not meant to be independent. The computational, algorithmic, and implementation levels inform one another when we use computation to explain a device's behaviour. Nevertheless, Marr's three levels represent three distinct ways that computation can contribute to explaining a device's behaviour. One might offer a computational theory of an unknown device as a computational-level description (which function does the device compute and why?), as an algorithmic description (how does the device compute its function?) or as an implementation description (how does the physical activity in the device map onto its method for computing its function?). When you encounter a computational theory in science, it is worth asking yourself which of Marr's three levels that theory aims to address.

The situation that we faced with granny's mysterious device is not unlike that which cognitive scientists encounter with the human brain. Cognitive scientists want to know which computations the brain performs, which algorithms the brain uses, and which bits of the brain are significant to performing its computation. Computational theories in cognitive science are offered at each of Marr's three levels. The techniques in cognitive science for answering these questions are not dissimilar to those we saw for granny's device. In order to arrive at a computational-level description, cognitive scientists try to understand the ecological purpose of behaviour (what behaviour produced by the brain aims to achieve). In order to arrive at an algorithmic-level description, cognitive scientists try to understand the basic steps that the brain can perform, the speed that it can perform them in, and how basic steps can be combined to produce algorithms that match human reaction times and susceptibility to errors. In order to arrive at an implementation-level description, cognitive

scientists watch the brain, using a variety of experimental techniques (fMRI, EEG, single-cell recording), and see how performance is affected when some of the brain's resources are damaged or temporarily disabled (for example, by stroke or by drugs).

The big difference between granny's device and the human brain is that brains are vastly more complex. The human brain is one of the most complex objects in the universe. It contains a hundred billion neurons, and a mindbogglingly complicated web of close to a quadrillion connections. The brain performs not one, but many, computations simultaneously, each one a great deal more complex than the addition function. Unravelling a computational description of the human brain is a daunting task. Yet it is a project on which significant inroads have already been made.

4 The brain: a guessing machine?

Recently, a new hypothesis has emerged regarding the type of computations that the brain might perform: the idea is that the brain might be working like a probabilistic machine, using statistical knowledge to guide perception and decision-making. At the origin of this idea is the recognition that we live in a world of uncertainty. Our environment is often ambiguous or noisy, and our sensory receptors are limited. For example, the structure of the world is 3D but our retinas are only 2D, so our brains need to 'reconstruct' the 3D structure from two 2D images coming from the eyes. Often, multiple interpretations are possible. In this context, the best our brain can do is to try to guess at what is present in the world and what best action to take.

Hermann von Helmholtz (1821–94) is often credited with understanding this. Studying the human eye and judging it to be a very imperfect optical instrument, von Helmholtz proposed that visual perception was the result of what he called an 'unconscious inference' carried out by the brain. Through this automatic process, the brain would complete missing information and construct hypotheses about the visual environment, which would then be accepted as our immediate reality.

This idea of the brain as a 'guessing machine' has been formalized in recent years taking ideas from machine learning and statistics. It is proposed that the brain works by constantly forming hypotheses or 'beliefs' about what is present in the world and the actions to take, and by evaluating those hypotheses based on current evidence and prior knowledge. Those hypotheses can be described mathematically as conditional probabilities, denoted $P(\text{hypothesis} \mid \text{data})$, which means: the probability of the hypothesis given the data, where 'data' represents the signals available to our senses. Statisticians have shown that the best way to compute those probabilities is to use Bayes' theorem, named after Thomas Bayes (1701–61):

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \tag{1}$$

Bayes' theorem is of fundamental importance in statistics: it is sometimes considered to be 'to the theory of probability what Pythagoras's theorem is to geometry' (Jeffreys 1973, p. 31). Using Bayes' theorem to update beliefs is called Bayesian inference. For example, suppose you are trying to figure out whether it is going to rain today. The data available might be the dark clouds that you can observe by the window. Bayes' theorem states that we can get the probability $P(\text{hypothesis} \mid \text{data})$, which we call the *posterior probability*, by multiplying two other probabilities:

- $P(\text{data} \mid \text{hypothesis})$: our knowledge about the probability of the data given the hypothesis (e.g. 'how probable is it that the clouds look the way they do now, when you actually know it is going to rain?'), which is called the likelihood
- $P(\text{hypothesis})$: called the prior probability, which represents our knowledge about the hypothesis before we collect any new information – here for example the probability that it is going to rain in a day, independently of the shape of the cloud, a number which would be very different whether you live in Edinburgh or Los Angeles.

The denominator, $P(\text{data})$, is only there to ensure the resulting probability is comprised between 0 and 1, and can often be disregarded in the computations. In the visual system, in a similar way, a hypothesis could be about the presence of a given object ('is there a bear running after me?'), or about the value of a given stimulus ('the speed of this bear is 30 km/h'), while the data is the noisy visual inputs. The critical assumptions about Bayesian inference as a model of how the brain works are:

- The uncertainty of the environment is taken into account and manipulated in the brain by always keeping track of the probabilities of the different possible interpretations; The brain has developed (through development and experience) an internal model of the world in the form of prior beliefs and likelihoods that can be consulted to predict and interpret new situations;
- The brain combines new evidence with prior beliefs in a principled way, through the application of Bayes' theorem;
- Because currently developed intelligent machines also work in that way – learning from data to make sense of their noisy or ambiguous inputs and updating beliefs – we can get inspiration from machine learning algorithms to understand how this could be implemented in the brain.

5 Multisensory integration as evidence for Bayesian inference

A good model is one that makes predictions. Can Bayesian inference as a model of cognition make predictions that can be tested? This has been the aim of a lot of experimental and theoretical work in the last fifteen years. How our brain combines information from the different senses (for example vision and audition, or vision and touch) is often considered strong evidence for Bayesian inference. Suppose for example that you are walking in the forest, and fear that someone, or an animal, is following you. You can dimly see and hear a rustling of the leaves of the trees. How do you figure out where the animal is located? Do you use one sensory modality more than the other, for example only vision, or both? How does this depend on the reliability of the information available to each of the senses? Similarly, when someone is talking and you can hear the sound of their voice and see the movement of their lips, how do you combine the visual and auditory information to make sense of what they are saying?

Bayesian inference predicts that the best way to do this is to combine the information from both modalities, while weighting the information from each modality according to its reliability. For example, if the visual information is much clearer than the auditory information, it should have much more influence on your experience. This can lead to illusions in situations where there is a conflict between the two modalities, and one modality is much more reliable than the other.

In a lot of situations, it seems that the Bayesian predictions are qualitatively correct. This can be seen for example in the phenomenon known as the McGurk effect, which illustrates how the brain combines information from vision and audition (Figure 6.1A). This effect was discovered by accident. McGurk and his research assistant MacDonald were conducting a study on language perception in infants. They asked a technician to dub the sound of a phoneme (e.g. (Ba)) over a video that was showing lip movements for another phoneme (e.g. (Ga)). They discovered that one would then perceive a third phoneme, different from the one spoken or mouthed in the video, e.g. (Da). Even if we know about this effect, we continue perceiving (Da). This shows that our brain automatically and unconsciously integrates visual and auditory information in our perception of speech, creating a new 'mixture' that might be very different from the initial sources of information.

Figure 6.1. The brain naturally combines information coming from different senses. (A) The McGurk effect. When the sound of the phoneme (Ba) is dubbed over a video showing lip movements for the phoneme (Ga), we perceive a third phoneme, (Da). (Copyright image: R. Guski, *Wahrnehmen: Ein Lehrbuch* (Stuttgart: Kohlhammer, 1996)) (B) Ventriloquism is an extreme example where visual information completely overwhelms auditory information: because the source of the sound is

quite uncertain, while visual information about the puppet's moving lips is clear, we end up perceiving that it is the puppet which is talking and not the ventriloquist. (The ventriloquist Edgar Bergen in *Stage Door Canteen*) (C) Experimental paradigm used by Ernst and Banks (2002) to test the predictions of the Bayesian approach. Participants had to estimate the height of a ridge based on visual and touch information. (Copyright image: *Nature Publishing Group*)

Sometimes, though, when vision is much more reliable than audition, it can completely dominate our perceptual judgements. This can be seen in the compelling illusion of ventriloquism (Figure 6.1B). Originally used as a religious practice in ancient Greece (the ventriloquist was thought to be able to talk with the dead), ventriloquism is defined as the art of projecting one's voice so that it seems to come from another source, as from a puppet. This is a case of 'visual capture': because the origin of the sound is uncertain, but the lips of the puppet can be clearly perceived, one attributes the origin of the sound to the visual inputs, i.e. the mouth of the puppet.

The previous examples show qualitatively that, in everyday life, the brain combines signals coming from different senses in a way that depends on their uncertainty. In the laboratory, researchers can perform much more precise measurements about the validity of the Bayesian predictions. In a seminal paper published in *Nature* in 2002, Marc Ernst and Martin Banks reported the results of an experiment where human subjects were required to make discrimination judgements about 3D shapes (Figure 6.1C). Subjects had to judge which of two sequentially presented ridges was taller. There were three types of trials. First, the subjects could only touch the ridge. Then, they had only visual information: they could only see the ridge. Finally, subjects had both types of information at the same time: they could both touch and see the ridge simultaneously. A different amount of noise was added to the visual stimuli so as to manipulate the reliability of the visual cue. Ernst and Banks measured the smallest difference in the height of the ridge that subjects could reliability detect (aka the 'discrimination threshold') based first on only visual information, then based only on touch information. From these, they could predict quantitatively the performance of subjects for the condition where both visual and touch cues were present, under the assumption that subjects would integrate information from the two cues in a Bayesian way. They found that measured performance was in fact very similar to the Bayesian prediction and concluded that human observers were 'Bayesian optimal'. Since then, this result has been replicated in many different laboratories, using different modalities (for example vision and audition). It is commonly considered as evidence that the brain combines information from different sources in a way similar to a Bayesian machine.

6 Visual illusions and Bayesian priors

The Bayesian model not only predicts how simultaneous signals need to be combined, but also how to include prior knowledge. According to Bayes' rule, such knowledge can be represented as a prior probability, which would serve as a summary of all previous experience, and which should be multiplied with the new information, the likelihood (see Equation 1). Recently, a number of researchers have tried to explore this idea: if the brain uses prior beliefs, what are those? And how do they influence perception?

Intuitively, it is when sensory data is limited or ambiguous that we rely on our prior knowledge. For example, if we wake up in the middle of the night and need to walk in total darkness, we automatically use our prior knowledge of the environment, or of similar environments, to guide our path. Mathematically, similarly, Bayes' theorem indicates that prior distributions should have maximum impact in situations of strong uncertainty. Thus, a good way to discover the brain's expectations or assumptions is to study perception or cognition in situations where the current sensory inputs or the 'evidence' is very limited. Studying such situations reveals that our brains make automatic assumptions all the time.

Visual illusions are great examples of this. Consider Figure 6.2A for example: despite this image being 2D, we automatically have an impression of depth. Are those shapes bumps (i.e. convex shapes) or dimples (i.e. concave shapes)? You might not be aware of it, but perceiving one dimple in the middle of bumps is consistent with assuming that light comes from the top of the image. Turning the page upside down would lead to the opposite percept (seeing a bump in a middle of dimples). The prior assumption that light comes 'from above' has been extensively studied. It is known to play a crucial role in how we view shapes that project a shadow. The fact that the brain uses this assumption makes sense of course, since light usually comes from the sun, above us.

Figure 6.2. The brain naturally combines visual information with prior assumptions, leading sometimes to visual illusions. (A) Example of the 'light-from-above' prior. Are those shapes bumps or dimples? Perceiving one dimple in the middle of bumps is consistent with assuming that light comes from the top of the image. Turning the page upside down would lead to the opposite percept (seeing a bump in a middle of dimples). (Copyright image: R. Champion and W. Adams, *Journal of Vision* 7(13) (2007), article 10) (B) *The Ames room illusion*. Here, the brain assumes that the room is boxshaped and thus infers that the height of the people is vastly different. In reality, the room is trapezoidal and the ceiling is not horizontal. (Copyright image: Tom Pringle). (C) *The hollow mask illusion*. The interior of a mask is perceived as a normal face, with the nose sticking out instead of sticking in. (Image taken by

Peggy Series. Copyright object: Camera Obscura, Edinburgh)

Similarly, we seem to expect objects to be symmetrical, to change smoothly in space and time, orientations to be more frequently horizontal or vertical and angles to look like perpendicular corners (Figure 6.2B). We also expect objects to bulge outward more than inward (i.e. to be convex shapes, like balloons or pears), that background images are coloured in a uniform way, that objects move slowly or not at all, that the gaze of other people is directed towards us, and that faces correspond to convex surfaces. The latter is illustrated by the classic illusion known as the 'hollow-mask illusion' where a concave mask of a face (i.e. the interior of a mask) appears as a normal convex face, with the nose sticking outward instead of inward (Figure 6.2C). Here, the bias towards perceiving faces as bulging outward is so strong that it counters depth cues, such as shading and shadows, as well as 3D cues that the brain receives by comparing the information available from both eyes, which signal that the object is hollow. As for the McGurk effect, knowing about the illusion doesn't help: the interpretation chosen by the brain is automatic and unconscious and can't be modulated voluntarily.

Why would the brain use such assumptions to interpret the visual inputs? These prior assumptions make sense because most objects in the world conform to those expectations: light usually comes from 'above' (the sun), noses are always sticking outward, most objects are static or move only slowly, etc. On average, using such prior assumptions thus leads to the best possible guess about the environment. This is why they can be thought of as being 'optimal'. However, in situations of strong uncertainty and where objects don't conform to the average statistics, such assumptions can lead to illusions: we then perceive reality as being more similar to our expectations than it really is. Objects seem slower, more symmetrical, and smoother in space and time, etc.

The Bayesian approach helps in formalizing these ideas. A seminal example of this is the work of Yair Weiss and colleagues. These researchers were interested in the expectation that objects are static or move slowly (which they called the 'slow speed prior'). They postulated that this prior belief could elegantly explain many motion illusions: for example the fact that a line moving behind a circular window (aka 'aperture') is always perceived as moving perpendicular to its orientation ('the aperture problem') or that the perceived direction of motion of a diamond-shaped object (aka 'rhombus') depends on how bright it is compared with the background (i.e. its contrast level). Using a simple Bayesian model where visual information is combined with a prior expectation for slow motion, Yair Weiss and colleagues have shown that they could explain a variety of illusions that had only been explained by independent models previously. They thus offered the idea that visual illusions were not due to the limitations of a collection of imperfect hacks that the brain

would use, as commonly thought, or to 'the result of sloppy computation by various components in the visual system' or, but 'rather a result of a coherent computational strategy that is optimal under reasonable assumptions.' They finally concluded that, because they correspond to the brain making very sensible assumptions in a situation of uncertainty, visual illusions could be viewed, paradoxically, as 'optimal percepts'.

A lot of important questions remain. Where do those prior beliefs come from? How are they learned? Are they the same for everybody? Do they depend on the experience of individuals? Can we 'unlearn' the fact that light tends to come from above, or that faces are convex? These questions are the focus of current research. Experimental work combined with Bayesian modelling shows that our brain creates prior expectations all the time, unconsciously and automatically incorporating long-term and recent experience to make sense of the world. For example, after a few minutes of exposure to an environment where some visual features are more frequent than others (for example, objects tend to move in a given direction), we will expect these features to occur again. As a result, we will be more likely to perceive them even when they are not there, or to think that other features are more similar to what we expect than they really are. It has also been shown that the brain can update our 'long-term' prior beliefs that light comes from above or that objects move slowly if we are placed in environments where lights come from below or where objects move quickly. This shows that the brain constantly revises its assumptions and updates its internal model of the environment.

Researchers have also found ways to quantitatively measure the priors used by individuals, and in some cases compared such priors with the statistics of the environment. In general, it appears that the assumptions that people use conform *qualitatively* to the statistics of the world, but that *quantitatively* there is a lot of variability between individuals. This has generated some debates around the notion of optimality: the way the human brain works can be considered as 'optimal' in the type of computation it is trying to perform (i.e. an approximation of Bayesian inference, given the noisy signals it receives) but not always 'optimal' in that the beliefs and internal models it uses can be slightly different from how things really are in the world.

7 Mental disorders as deficits in Bayesian inference

The idea that the brain functions like a probabilistic machine is not restricted to perception, but has been applied to all domains of cognition. For example, the Bayesian approach may have promising application for the field of psychiatry. It is still very early to say whether this approach will be helpful for understanding

mental illness, and there are many competing approaches, which are not all mutually exclusive. However, recent research shows that Bayesian models could potentially help in quantifying differences between different groups (e.g. healthy vs ill) and identifying whether such differences come from using different internal models, for example different prior beliefs, or from different learning or decision strategies. Ultimately, this may help drug discovery.

In the study of schizophrenia, for example, recent work reveals that patients with schizophrenia are not as good as healthy subjects at some probabilistic inference tasks. A task that is often used is that of the 'urns', where participants have to guess from which urn comes a bead drawn at random. In the original version of the task, one urn contains 85 per cent red beads and 15 per cent black beads, whereas the other urn contains 15 per cent red beads and 85 per cent black beads. The beads are drawn one after the other from the same urn, and the participants are asked when they have received sufficient information to decide which urn the beads were drawn from.

Schizophrenic patients are more likely to make their decision after a small number of observations (1 or 2 draws) and to report being certain about their decision after only one draw – a tendency to 'jump to conclusions' which could be crucial for the understanding of delusions and paranoia. Modelling work suggests that this behaviour could be explained by the patients' decision process using less information before committing to an answer, or that would be noisier than in controls.

A common idea in psychiatry is also that the internal models used by patients, in particular their prior beliefs, could be different from those of healthy subjects. In the study of schizophrenia, for example, it has been proposed that 'positive symptoms' (hallucination and delusions) could be related to an imbalance between information coming from the senses and prior beliefs or expectations. For example, using the wrong prior expectations could lead to delusions, while having prior expectations that are too strong could lead to hallucinations. In autism, similarly, it has been proposed that the influence of prior expectations might be too weak compared with that of sensory inputs, which could explain that patients feel overwhelmed by a world perceived as being 'too real'.

In the long run, Bayesian modelling could also help diagnosis. Psychiatric disease or personality traits are usually measured using questionnaires or classification such as DSM-V (the *Diagnostic and Statistical Manual of Mental Disorders* used by clinicians and psychiatrists to diagnose psychiatric illnesses). Coupled with behavioural measurements, Bayesian modelling could help identify more quantitatively the internal beliefs people's brains are working with. For example, Aistis Stankevicius and colleagues (2014) have shown that Bayesian models could help measure how optimistic or pessimistic people are, using a simple type of game

where participants are asked to choose between different visual targets. One of the targets is certain: the participants are explicitly told with which probability the target could lead to a reward. The other target is uncertain: the participants have to guess its probability of reward, based on limited previous experience. Optimists expect the uncertain target to be associated with rewards more often than pessimists do, and the amplitude of these expectations can be precisely measured based on the choices of the participants. Such measures could be a complement to the usual questionnaires and have interesting applications in the study of depression.

8 The implementation of probabilities in neural activity

Bayesian models seem to be very useful for describing perception and behaviour at the computational level (the *which* and the *why*, as explained above). How these algorithms are implemented in the brain and relate to neural activity is still an open question and an active area of research. Whether the Bayesian approach can actually make predictions for neurobiology (for example on which parts of the brain would be involved, or how neural activity could represent probabilities) is debated. It is yet unclear whether the Bayesian approach is only useful at the 'computational' level, to describe the computations performed by the brain overall, or whether it can also be useful at the 'implementation level' to predict how those algorithms might be implemented in the neural tissue.

9 Chapter summary

- Intelligent machines that are able to learn from data have become more and more common. To be efficient, such machines need to represent uncertainty in the data, be adaptive and robust. Recently, building machines that represent beliefs in the form of probabilities and update such beliefs using Bayes' theorem has been found to be a particularly successful approach.
- In neuroscience, the idea has emerged that the brain might work in the same way. The brain would represent beliefs in the form of probabilities, and would have developed an internal model of the world in the form of prior beliefs and likelihoods that can be consulted to predict and interpret new situations. The brain would then combine new evidence with prior beliefs in a principled way, through the application of Bayes' theorem.
- Experiments provide support for this idea. When combining multiple sources of integration, the brain does take into account the reliability of each source of information. Moreover, it is clear that the brain works by using prior beliefs in situations of strong uncertainty. The existence of these beliefs can explain

a variety of visual illusions, such as the 'hollow mask illusion' or the 'Ames room illusion'. Experiments also show that beliefs about the environment are updated constantly and automatically and can be quantitatively measured in individuals.

- It is still very early to tell but this approach might have interesting applications in psychiatry. Mental illness might correspond to deficits in Bayesian inference, or to the learning and use of internal models that would be different from that used by healthy controls.

Study questions

1. What is the computational theory of mind? Can you think of mental processes that computation would be good at explaining? Which mental processes may it struggle to explain?
2. What are probabilistic representations and why might they be useful for generating intelligent behaviour?
3. Describe in your own words Marr's three levels. How might information at each level constrain the description at the other levels?
4. Why does Marr group the question of *which* computation a device performs together with the question of *why* the device performs that computation?
5. In your own words, describe what Bayes' theorem is about.
6. Explain in your own words how the ventriloquist illusion works.
7. Can you think of a situation where your perception was influenced by your expectations or prior beliefs, so that you had the impression of perceiving something that was in reality not there? Could you try to explain what went on in your brain?
8. Describe in your own words why the Bayesian approach might give us new ways to understand mental illness.

Introductory further reading

- Clark, A. (2014) *Mindware: An Introduction to Cognitive Science*, 2nd edn, Oxford: Oxford University Press. (A great introduction to the computational approach to the mind.)
- Copeland, B. J. (ed.) (2004) *The Essential Turing*, Oxford: Oxford University Press, chs 9–14. (An anthology of many of Turing's original papers with

excellent introductions and annotations. Turing's original paper on universal computing machines, 'On Computable Numbers', is in chapter 1. Chapters 9–14 give a wonderful overview of Turing's contribution to machine intelligence.)

- Frith, C. (2007) *Making Up the Mind: How the Brain Creates Our Mental World*, Malden, MA: Blackwell. (Highly enjoyable book that introduces many key ideas in current psychology.)
- Russell, S. and Norvig, P. (2010) *Artificial Intelligence: A Modern Approach*, 3rd edn, Upper Saddle River: Pearson. (The classic textbook on artificial intelligence. Chapter 1 has a wonderful and accessible summary of the history of the field.)
- Stone, J. V. (2013) *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*, n.p.: Sebtel Press, http://jim-stone.staff.shef.ac.uk/BookBayes2012/bookbayesch01.pdf. (A very accessible introduction to Bayesian inference and its applications.)
- Vilares, I. and Kording, K. (2011) 'Bayesian models: the structure of the world, uncertainty, behavior, and the brain', *Annals of the New York Academy of Sciences* 1224: 22–39. (An accessible review of the current research using Bayesian models to study the brain.)

Advanced further reading

- Adams, W. J., Graf, E. W. and Ernst, M. O. (2004) 'Experience can change the "light-from-above" prior', *Nature Neuroscience* 7: 1057–8. (A seminal study showing that the prior beliefs used by the brain are constantly updated.)
- Ernst, M. O. and Banks, M. S. (2002) 'Humans integrate visual and haptic information in a statistically optimal fashion', *Nature* 415: 429–33. (A seminal study showing that the way the brain integrates vision and touch is compatible with Bayesian inference.)
- Haugeland, J. (ed.) (1999) *Mind Design II*, Cambridge, MA: MIT Press. (A nice collection of essays on philosophical debates surrounding a computational approach to the mind.)
- Hohwy, J. (2014) *The Predictive Mind*, Oxford: Oxford University Press. (A recent and accessible monograph describing the theory according to which the brain works as a hypothesis-testing machine, one that attempts to minimize the error of its predictions about the sensory inputs it receives from the world.)
- Jeffreys, H. (1973) *Scientific Inference*, 3rd edn, Cambridge: Cambridge University Press.
- Marr, D. (1982) *Vision*, San Francisco: W. H. Freeman. (Marr's best known work, published posthumously, and a classic in cognitive science. Very readable and engaging. Chapter 1 neatly describes Marr's three levels of computa-

tional description.)

- Stankevicius, A., Huys, Q., Kalra, A. and Series, P. (2014) 'Optimism as a prior on the likelihood of future reward', *PLOS Computational Biology* 10. (A recent study showing how Bayesian models can be used to study personality traits and cognitive biases.)
- Weiss, Y., Simoncelli, E. P. and Adelson, E. H. (2002) 'Motion illusions as optimal percepts', *Nature Neuroscience* 5: 598–604. (A very elegant study showing how a variety of motion illusions can be explained by the brain expecting objects to move slowly.)

Internet resources

- Davey, M. (2010) A Turing Machine in Action: The Classic Style [website], http://aturingmachine.com
- Example of an addition algorithm that uses Roman numerals: Turner, L. E. (2007) Roman Arithmetic: When in Rome, Do as the Romans Do!, Southwest Adventist University [online course material], http://turner.faculty.swau.edu/mathematics/materialslibrary/roman/
- The McGurk effect: [BBC] (2010) 'Try the McGurk effect! – Horizon: is seeing believing? – BBC Two', YouTube, 10 November, http://www.youtube.com/watch?v=G-lN8vWm3m0
- The aperture problem: Anonymous (2013) 'The aperture problem', §4 of 'Motion perception', Wikipedia, 6 October, http://en.wikipedia.org/wiki/Motion_perception.
- The rhombus illusion: Weiss, Y. (n.d.) Moving Rhombus Displays, Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University of Jerusalem, http://www.cs.huji.ac.il/~yweiss/Rhombus/rhombus.html
- The Ames room illusion: [Scientific American] (2012) 'What is the Ames illusion? – Instant Egghead 23', YouTube, 11 October, http://www.youtube.com/watch?v=gJhyu6nlGt8
- The hollow mask illusion: eChalk Scientific (2012) 'The rotating mask illusion', YouTube, 20 July, http://www.youtube.com/watch?v=sKa0eaKsdA0.
- A TED talk by Daniel Wolpert (2011), 'The real reason for brains', TED, July [video blog], http://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains