theguardian.com

Why your brain is not a computer

Matthew Cobb

24-30 minutes

We are living through one of the greatest of scientific endeavours – the attempt to understand the most complex object in the universe, the brain. Scientists are accumulating vast amounts of data about structure and function in a huge array of brains, from the tiniest to our own. Tens of thousands of researchers are devoting massive amounts of time and energy to thinking about what brains do, and astonishing new technology is enabling us to both describe and manipulate that activity.

We can now make a mouse remember something about a smell it has never encountered, turn a bad mouse memory into a good one, and even use a surge of electricity to change how people perceive faces. We are drawing up increasingly detailed and complex functional maps of the brain, human and otherwise. In some species, we can change the brain's very structure at will, altering the animal's behaviour as a result. Some of the most profound consequences of our growing mastery can be seen in our ability to enable a paralysed person to control a robotic arm with the power of their mind.

Every day, we hear about new discoveries that shed light on how brains work, along with the promise – or threat – of new

technology that will enable us to do such far-fetched things as read minds, or detect criminals, or even be uploaded into a computer. Books are repeatedly produced that each claim to explain the brain in different ways.

And yet there is a growing conviction among some neuroscientists that our future path is not clear. It is hard to see where we should be going, apart from simply collecting more data or counting on the latest exciting experimental approach. As the German neuroscientist Olaf Sporns has put it: “Neuroscience still largely lacks organising principles or a theoretical framework for converting brain data into fundamental knowledge and understanding.” Despite the vast number of facts being accumulated, our understanding of the brain appears to be approaching an impasse.

In 2017, the French neuroscientist Yves Frégnac focused on the current fashion of collecting massive amounts of data in expensive, large-scale projects and argued that the tsunami of data they are producing is leading to major bottlenecks in progress, partly because, as he put it pithily, “big data is not knowledge”.

“Only 20 to 30 years ago, neuroanatomical and neurophysiological information was relatively scarce, while understanding mind-related processes seemed within reach,” Frégnac wrote. “Nowadays, we are drowning in a flood of information. Paradoxically, all sense of global understanding is in acute danger of getting washed away. Each overcoming of technological barriers opens a Pandora’s box by revealing hidden variables, mechanisms and nonlinearities, adding new levels of complexity.”

The neuroscientists Anne Churchland and Larry Abbott have also emphasised our difficulties in interpreting the massive amount of data that is being produced by laboratories all over the world: “Obtaining deep understanding from this onslaught will require, in addition to the skilful and creative application of experimental technologies, substantial advances in data analysis methods and intense application of theoretic concepts and models.”

There are indeed theoretical approaches to brain function, including to the most mysterious thing the human brain can do – produce consciousness. But none of these frameworks are widely accepted, for none has yet passed the decisive test of experimental investigation. It is possible that repeated calls for more theory may be a pious hope. It can be argued that there is no possible single theory of brain function, not even in a worm, because a brain is not a single thing. (Scientists even find it difficult to come up with a precise definition of what a brain is.)

As observed by Francis Crick, the co-discoverer of the DNA double helix, the brain is an integrated, evolved structure with different bits of it appearing at different moments in evolution and adapted to solve different problems. Our current comprehension of how it all works is extremely partial – for example, most neuroscience sensory research has been focused on sight, not smell; smell is conceptually and technically more challenging. But the way that olfaction and vision work are different, both computationally and structurally. By focusing on vision, we have developed a very limited understanding of what the brain does and how it does it.

The nature of the brain – simultaneously integrated and composite – may mean that our future understanding will inevitably be

fragmented and composed of different explanations for different parts. Churchland and Abbott spelled out the implication: “Global understanding, when it comes, will likely take the form of highly diverse panels loosely stitched together into a patchwork quilt.”

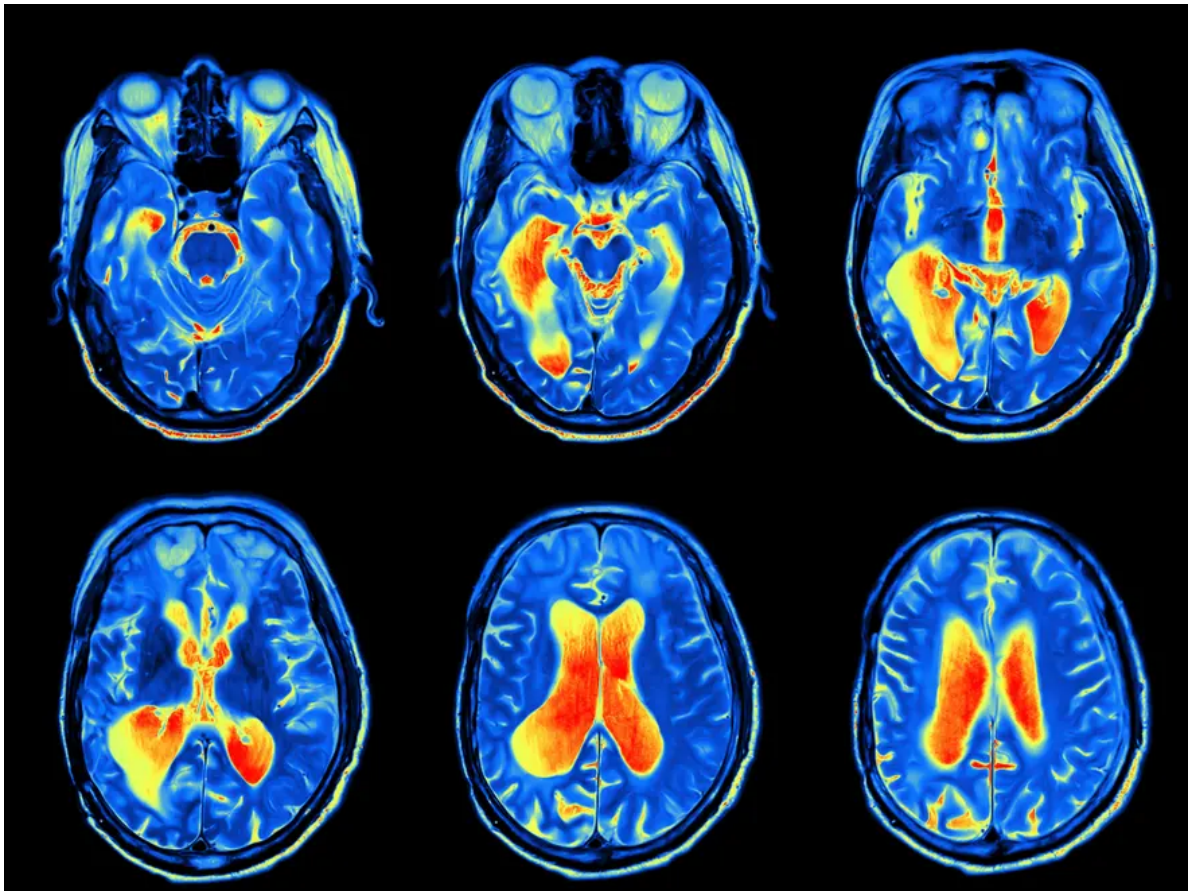
For more than half a century, all those highly diverse panels of patchwork we have been working on have been framed by thinking that brain processes involve something like those carried out in a computer. But that does not mean this metaphor will continue to be useful in the future. At the very beginning of the digital age, in 1951, the pioneer neuroscientist Karl Lashley argued against the use of any machine-based metaphor.

“Descartes was impressed by the hydraulic figures in the royal gardens, and developed a hydraulic theory of the action of the brain,” Lashley wrote. “We have since had telephone theories, electrical field theories and now theories based on computing machines and automatic rudders. I suggest we are more likely to find out about how the brain works by studying the brain itself, and the phenomena of behaviour, than by indulging in far-fetched physical analogies.”

This dismissal of metaphor has recently been taken even further by the French neuroscientist Romain Brette, who has challenged the most fundamental metaphor of brain function: coding. Since its inception in the 1920s, the idea of a neural code has come to dominate neuroscientific thinking – more than 11,000 papers on the topic have been published in the past 10 years. Brette’s fundamental criticism was that, in thinking about “code”, researchers inadvertently drift from a technical sense, in which there is a link between a stimulus and the activity of the neuron, to

a representational sense, according to which neuronal codes represent that stimulus.

The unstated implication in most descriptions of neural coding is that the activity of neural networks is presented to an ideal observer or reader within the brain, often described as “downstream structures” that have access to the optimal way to decode the signals. But the ways in which such structures actually process those signals is unknown, and is rarely explicitly hypothesised, even in simple models of neural network function.



MRI scan of a brain. Photograph: Getty/iStockphoto

The processing of neural codes is generally seen as a series of linear steps – like a line of dominoes falling one after another. The brain, however, consists of highly complex neural networks that are interconnected, and which are linked to the outside world to effect action. Focusing on sets of sensory and processing neurons

without linking these networks to the behaviour of the animal misses the point of all that processing.

By viewing the brain as a computer that passively responds to inputs and processes data, we forget that it is an active organ, part of a body that is intervening in the world, and which has an evolutionary past that has shaped its structure and function. This view of the brain has been outlined by the Hungarian neuroscientist György Buzsáki in his recent book The Brain from Inside Out. According to Buzsáki, the brain is not simply passively absorbing stimuli and representing them through a neural code, but rather is actively searching through alternative possibilities to test various options. His conclusion – following scientists going back to the 19th century – is that the brain does not represent information: it constructs it.

The metaphors of neuroscience – computers, coding, wiring diagrams and so on – are inevitably partial. That is the nature of metaphors, which have been intensely studied by philosophers of science and by scientists, as they seem to be so central to the way scientists think. But metaphors are also rich and allow insight and discovery. There will come a point when the understanding they allow will be outweighed by the limits they impose, but in the case of computational and representational metaphors of the brain, there is no agreement that such a moment has arrived. From a historical point of view, the very fact that this debate is taking place suggests that we may indeed be approaching the end of the computational metaphor. What is not clear, however, is what would replace it.

Scientists often get excited when they realise how their views have been shaped by the use of metaphor, and grasp that new

analogies could alter how they understand their work, or even enable them to devise new experiments. Coming up with those new metaphors is challenging – most of those used in the past with regard to the brain have been related to new kinds of technology. This could imply that the appearance of new and insightful metaphors for the brain and how it functions hinges on future technological breakthroughs, on a par with hydraulic power, the telephone exchange or the computer. There is no sign of such a development; despite the latest buzzwords that zip about – blockchain, quantum supremacy (or quantum anything), nanotech and so on – it is unlikely that these fields will transform either technology or our view of what brains do.

One sign that our metaphors may be losing their explanatory power is the widespread assumption that much of what nervous systems do, from simple systems right up to the appearance of consciousness in humans, can only be explained as emergent properties – things that you cannot predict from an analysis of the components, but which emerge as the system functions.

In 1981, the British psychologist Richard Gregory argued that the reliance on emergence as a way of explaining brain function indicated a problem with the theoretical framework: “The appearance of ‘emergence’ may well be a sign that a more general (or at least different) conceptual scheme is needed … It is the role of good theories to remove the appearance of emergence. (So explanations in terms of emergence are bogus.)”

This overlooks the fact that there are different kinds of emergence: weak and strong. Weak emergent features, such as the movement of a shoal of tiny fish in response to a shark, can be understood in

terms of the rules that govern the behaviour of their component parts. In such cases, apparently mysterious group behaviours are based on the behaviour of individuals, each of which is responding to factors such as the movement of a neighbour, or external stimuli such as the approach of a predator.

This kind of weak emergence cannot explain the activity of even the simplest nervous systems, never mind the working of your brain, so we fall back on strong emergence, where the phenomenon that emerges cannot be explained by the activity of the individual components. You and the page you are reading this on are both made of atoms, but your ability to read and understand comes from features that emerge through atoms in your body forming higher-level structures, such as neurons and their patterns of firing – not simply from atoms interacting.

Strong emergence has recently been criticised by some neuroscientists as risking “metaphysical implausibility”, because there is no evident causal mechanism, nor any single explanation, of how emergence occurs. Like Gregory, these critics claim that the reliance on emergence to explain complex phenomena suggests that neuroscience is at a key historical juncture, similar to that which saw the slow transformation of alchemy into chemistry. But faced with the mysteries of neuroscience, emergence is often our only resort. And it is not so daft – the amazing properties of deep-learning programmes, which at root cannot be explained by the people who design them, are essentially emergent properties.

Interestingly, while some neuroscientists are discombobulated by the metaphysics of emergence, researchers in artificial intelligence revel in the idea, believing that the sheer complexity of modern computers, or of their interconnectedness through the internet, will

lead to what is dramatically known as the singularity. Machines will become conscious.

There are plenty of fictional explorations of this possibility (in which things often end badly for all concerned), and the subject certainly excites the public's imagination, but there is no reason, beyond our ignorance of how consciousness works, to suppose that it will happen in the near future. In principle, it must be possible, because the working hypothesis is that mind is a product of matter, which we should therefore be able to mimic in a device. But the scale of complexity of even the simplest brains dwarfs any machine we can currently envisage. For decades – centuries – to come, the singularity will be the stuff of science fiction, not science.

A related view of the nature of consciousness turns the brain-as-computer metaphor into a strict analogy. Some researchers view the mind as a kind of operating system that is implemented on neural hardware, with the implication that our minds, seen as a particular computational state, could be uploaded on to some device or into another brain. In the way this is generally presented, this is wrong, or at best hopelessly naive.

The materialist working hypothesis is that brains and minds, in humans and maggots and everything else, are identical. Neurons and the processes they support – including consciousness – are the same thing. In a computer, software and hardware are separate; however, our brains and our minds consist of what can best be described as wetware, in which what is happening and where it is happening are completely intertwined.

Imagining that we can repurpose our nervous system to run different programmes, or upload our mind to a server, might sound scientific, but lurking behind this idea is a non-materialist view going back to Descartes and beyond. It implies that our minds are somehow floating about in our brains, and could be transferred into a different head or replaced by another mind. It would be possible to give this idea a veneer of scientific respectability by posing it in terms of reading the state of a set of neurons and writing that to a new substrate, organic or artificial.

But to even begin to imagine how that might work in practice, we would need both an understanding of neuronal function that is far beyond anything we can currently envisage, and would require unimaginably vast computational power and a simulation that precisely mimicked the structure of the brain in question. For this to be possible even in principle, we would first need to be able to fully model the activity of a nervous system capable of holding a single state, never mind a thought. We are so far away from taking this first step that the possibility of uploading your mind can be dismissed as a fantasy, at least until the far future.

For the moment, the brain-as-computer metaphor retains its dominance, although there is disagreement about how strong a metaphor it is. In 2015, the roboticist Rodney Brooks chose the computational metaphor of the brain as his pet hate in his contribution to a collection of essays entitled This Idea Must Die. Less dramatically, but drawing similar conclusions, two decades earlier the historian S Ryan Johansson argued that “endlessly debating the truth or falsity of a metaphor like ‘the brain is a computer’ is a waste of time. The relationship proposed is metaphorical, and it is ordering us to do something, not trying to

tell us the truth."

On the other hand, the US expert in artificial intelligence, Gary Marcus, has made a robust defence of the computer metaphor: "Computers are, in a nutshell, systematic architectures that take inputs, encode and manipulate information, and transform their inputs into outputs. Brains are, so far as we can tell, exactly that. The real question isn't whether the brain is an information processor, per se, but rather how do brains store and encode information, and what operations do they perform over that information, once it is encoded."

Marcus went on to argue that the task of neuroscience is to "reverse engineer" the brain, much as one might study a computer, examining its components and their interconnections to decipher how it works. This suggestion has been around for some time. In 1989, Crick recognised its attractiveness, but felt it would fail, because of the brain's complex and messy evolutionary history – he dramatically claimed it would be like trying to reverse engineer a piece of "alien technology". Attempts to find an overall explanation of how the brain works that flow logically from its structure would be doomed to failure, he argued, because the starting point is almost certainly wrong – there is no overall logic.

Reverse engineering a computer is often used as a thought experiment to show how, in principle, we might understand the brain. Inevitably, these thought experiments are successful, encouraging us to pursue this way of understanding the squishy organs in our heads. But in 2017, a pair of neuroscientists decided to actually do the experiment on a real computer chip, which had a real logic and real components with clearly designed functions. Things did not go as expected.

The duo – Eric Jonas and Konrad Paul Kording – employed the very techniques they normally used to analyse the brain and applied them to the MOS 6507 processor found in computers from the late 70s and early 80s that enabled those machines to run video games such as Donkey Kong and Space Invaders.

First, they obtained the connectome of the chip by scanning the 3510 enhancement-mode transistors it contained and simulating the device on a modern computer (including running the games programmes for 10 seconds). They then used the full range of neuroscientific techniques, such as “lesions” (removing transistors from the simulation), analysing the “spiking” activity of the virtual transistors and studying their connectivity, observing the effect of various manipulations on the behaviour of the system, as measured by its ability to launch each of the games.

Despite deploying this powerful analytical armoury, and despite the fact that there is a clear explanation for how the chip works (it has “ground truth”, in technospeak), the study failed to detect the hierarchy of information processing that occurs inside the chip. As Jonas and Kording put it, the techniques fell short of producing “a meaningful understanding”. Their conclusion was bleak: “Ultimately, the problem is not that neuroscientists could not understand a microprocessor, the problem is that they would not understand it given the approaches they are currently taking.”

This sobering outcome suggests that, despite the attractiveness of the computer metaphor and the fact that brains do indeed process information and somehow represent the external world, we still need to make significant theoretical breakthroughs in order to make progress. Even if our brains were designed along logical lines, which they are not, our present conceptual and analytical

tools would be completely inadequate for the task of explaining them. This does not mean that simulation projects are pointless – by modelling (or simulating) we can test hypotheses and, by linking the model with well-established systems that can be precisely manipulated, we can gain insight into how real brains function. This is an extremely powerful tool, but a degree of modesty is required when it comes to the claims that are made for such studies, and realism is needed with regard to the difficulties of drawing parallels between brains and artificial systems.

Current ‘reverse engineering’ techniques cannot deliver a proper understanding of an Atari console chip, let alone of a human brain. Photograph: Radharc Images/Alamy

Even something as apparently straightforward as working out the storage capacity of a brain falls apart when it is attempted. Such calculations are fraught with conceptual and practical difficulties. Brains are natural, evolved phenomena, not digital devices. Although it is often argued that particular functions are tightly localised in the brain, as they are in a machine, this certainty has

been repeatedly challenged by new neuroanatomical discoveries of unsuspected connections between brain regions, or amazing examples of plasticity, in which people can function normally without bits of the brain that are supposedly devoted to particular behaviours.

In reality, the very structures of a brain and a computer are completely different. In 2006, Larry Abbott wrote an essay titled “Where are the switches on this thing?”, in which he explored the potential biophysical bases of that most elementary component of an electronic device – a switch. Although inhibitory synapses can change the flow of activity by rendering a downstream neuron unresponsive, such interactions are relatively rare in the brain.

A neuron is not like a binary switch that can be turned on or off, forming a wiring diagram. Instead, neurons respond in an analogue way, changing their activity in response to changes in stimulation. The nervous system alters its working by changes in the patterns of activation in networks of cells composed of large numbers of units; it is these networks that channel, shift and shunt activity. Unlike any device we have yet envisaged, the nodes of these networks are not stable points like transistors or valves, but sets of neurons – hundreds, thousands, tens of thousands strong – that can respond consistently as a network over time, even if the component cells show inconsistent behaviour.

Understanding even the simplest of such networks is currently beyond our grasp. Eve Marder, a neuroscientist at Brandeis University, has spent much of her career trying to understand how a few dozen neurons in the lobster’s stomach produce a rhythmic grinding. Despite vast amounts of effort and ingenuity, we still cannot predict the effect of changing one component in this tiny

network that is not even a simple brain.

This is the great problem we have to solve. On the one hand, brains are made of neurons and other cells, which interact together in networks, the activity of which is influenced not only by synaptic activity, but also by various factors such as neuromodulators. On the other hand, it is clear that brain function involves complex dynamic patterns of neuronal activity at a population level. Finding the link between these two levels of analysis will be a challenge for much of the rest of the century, I suspect. And the prospect of properly understanding what is happening in cases of mental illness is even further away.

Not all neuroscientists are pessimistic – some confidently claim that the application of new mathematical methods will enable us to understand the myriad interconnections in the human brain. Others – like myself – favour studying animals at the other end of the scale, focusing our attention on the tiny brains of worms or maggots and employing the well-established approach of seeking to understand how a simple system works and then applying those lessons to more complex cases. Many neuroscientists, if they think about the problem at all, simply consider that progress will inevitably be piecemeal and slow, because there is no grand unified theory of the brain lurking around the corner.

There are many alternative scenarios about how the future of our understanding of the brain could play out: perhaps the various computational projects will come good and theoreticians will crack the functioning of all brains, or the connectomes will reveal principles of brain function that are currently hidden from us. Or a theory will somehow pop out of the vast amounts of imaging data we are generating. Or we will slowly piece together a theory (or

theories) out of a series of separate but satisfactory explanations. Or by focusing on simple neural network principles we will understand higher-level organisation. Or some radical new approach integrating physiology and biochemistry and anatomy will shed decisive light on what is going on. Or new comparative evolutionary studies will show how other animals are conscious and provide insight into the functioning of our own brains. Or unimagined new technology will change all our views by providing a radical new metaphor for the brain. Or our computer systems will provide us with alarming new insight by becoming conscious. Or a new framework will emerge from cybernetics, control theory, complexity and dynamical systems theory, semantics and semiotics. Or we will accept that there is no theory to be found because brains have no overall logic, just adequate explanations of each tiny part, and we will have to be satisfied with that. Or –

This is an edited extract from The Idea of the Brain by Matthew Cobb, which will be published in the UK by Profile on 12 March, and in the US by Basic Books on 21 April, and is available at guardianbookshop.com