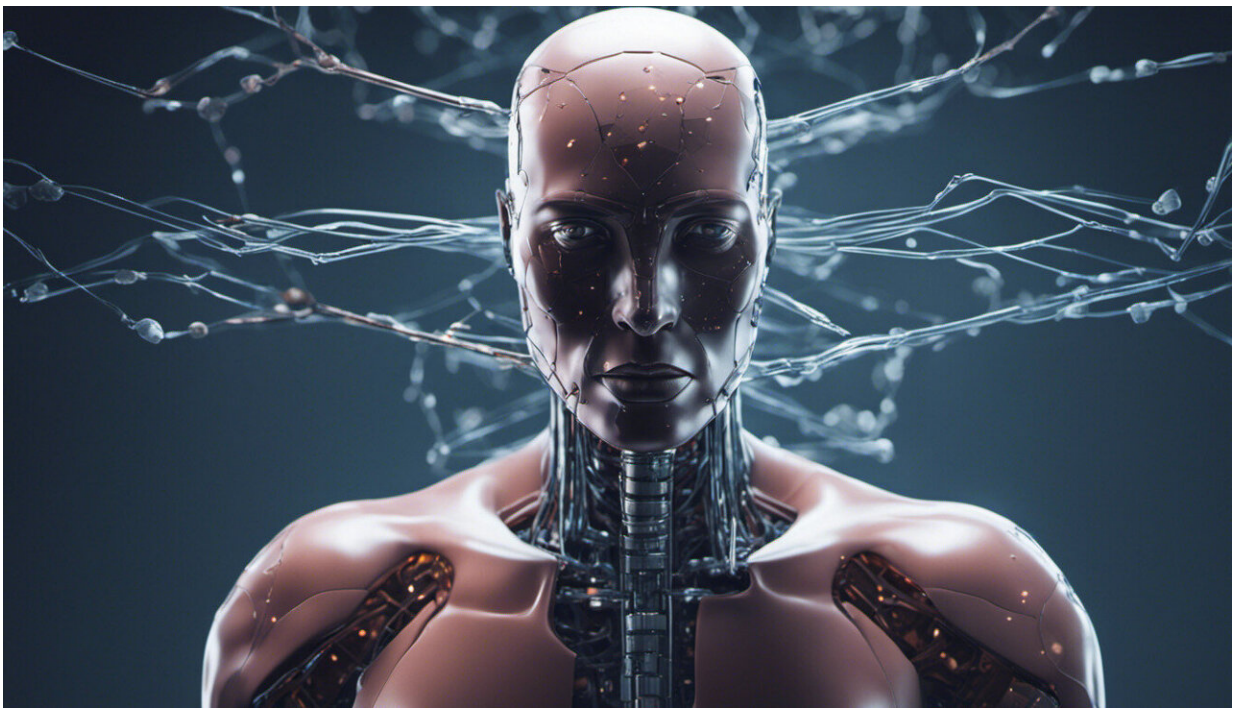
AI will soon become impossible for humans to comprehend—the story of neural networks tells us why

March 31 2023, by David Beer



Credit: AI-generated image (disclaimer)

In 1956, during a year-long trip to London and in his early 20s, the mathematician and theoretical biologist Jack D. Cowan visited Wilfred Taylor and his strange new "learning machine". On his arrival he was baffled by the "huge bank of apparatus" that confronted him. Cowan

could only stand by and watch "the machine doing its thing." The thing it appeared to be doing was performing an "associative memory scheme"—it seemed to be able to learn how to find connections and retrieve data.

It may have looked like clunky blocks of circuitry, soldered together by hand in a mass of wires and boxes, but what Cowan was witnessing was an early analog form of a neural network—a precursor to the most advanced artificial intelligence of today, including the much discussed ChatGPT with its ability to generate written content in response to almost any command. ChatGPT's underlying technology is a neural network.

As Cowan and Taylor stood and watched the machine work, they really had no idea exactly how it was managing to perform this task. The answer to Taylor's mystery machine brain can be found somewhere in its "analog neurons," in the associations made by its machine memory and, most importantly, in the fact that its automated functioning couldn't really be fully explained. It would take decades for these systems to find their purpose and for that power to be unlocked.

The term neural network incorporates a wide range of systems, yet centrally, according to IBM, these "neural networks—also known as artificial neural networks (ANNs) or simulated neural networks (SNNs)—are a subset of machine learning and are at the heart of deep learning algorithms." Crucially, the term itself and their form and "structure are inspired by the human brain, mimicking the way that biological neurons signal to one another."

There may have been some residual doubt of their value in its initial stages, but as the years have passed AI fashions have swung firmly towards neural networks. They are now often understood to be the future of AI. They have big implications for us and for what it means to be

human. We have heard echoes of these concerns recently with calls to pause new AI developments for a six month period to ensure confidence in their implications.

It would certainly be a mistake to dismiss the neural network as being solely about glossy, eye-catching new gadgets. They are already well established in our lives. Some are powerful in their practicality. As far back as 1989, a team led by Yann LeCun at AT&T Bell Laboratories used back-propagation techniques to train a system to recognize handwritten postal codes. The recent announcement by Microsoft that Bing searches will be powered by AI, making it your "copilot for the web," illustrates how the things we discover and how we understand them will increasingly be a product of this type of automation.

Drawing on vast data to find patterns AI can similarly be trained to do things like image recognition at speed—resulting in them being incorporated into facial recognition, for instance. This ability to identify patterns has led to many other applications, such as predicting stock markets.

Neural networks are changing how we interpret and communicate too. Developed by the interestingly titled Google Brain Team, Google Translate is another prominent application of a neural network.

You wouldn't want to play Chess or Shogi with one either. Their grasp of rules and their recall of strategies and all recorded moves means that they are exceptionally good at games (although ChatGPT seems to struggle with Wordle). The systems that are troubling human Go players (Go is a notoriously tricky strategy board game) and Chess grandmasters, are made from neural networks.

But their reach goes far beyond these instances and continues to expand. A search of patents restricted only to mentions of the exact phrase

"neural networks" produces 135,828 results. With this rapid and ongoing expansion, the chances of us being able to fully explain the influence of AI may become ever thinner. These are the questions I have been examining in my research and my new book on algorithmic thinking.

Mysterious layers of 'unknowability'

Looking back at the history of neural networks tells us something important about the automated decisions that define our present or those that will have a possibly more profound impact in the future. Their presence also tells us that we are likely to understand the decisions and impacts of AI even less over time. These systems are not simply black boxes, they are not just hidden bits of a system that can't be seen or understood.

It is something different, something rooted in the aims and design of these systems themselves. There is a long-held pursuit of the unexplainable. The more opaque, the more authentic and advanced the system is thought to be. It is not just about the systems becoming more complex or the control of intellectual property limiting access (although these are part of it). It is instead to say that the ethos driving them has a particular and embedded interest in "unknowability." The mystery is even coded into the very form and discourse of the neural network. They come with deeply piled layers—hence the phrase deep learning—and within those depths are the even more mysterious sounding "hidden layers." The mysteries of these systems are deep below the surface.

There is a good chance that the greater the impact that artificial intelligence comes to have in our lives the less we will understand how or why. Today there is a strong push for AI that is explainable. We want to know how it works and how it arrives at decisions and outcomes. The EU is so concerned by the potentially "unacceptable risks" and even "dangerous" applications that it is currently advancing a new AI Act

intended to set a "global standard" for "the development of secure, trustworthy and ethical artificial intelligence."

Those new laws will be based on a need for explainability, demanding that "for high-risk AI systems, the requirements of high quality data, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI." This is not just about things like self-driving cars (although systems that ensure safety fall into the EU's category of high risk AI), it is also a worry that systems will emerge in the future that will have implications for human rights.

This is part of wider calls for transparency in AI so that its activities can be checked, audited and assessed. Another example would be the Royal Society's policy briefing on explainable AI in which they point out that "policy debates across the world increasingly see calls for some form of AI explainability, as part of efforts to embed ethical principles into the design and deployment of AI-enabled systems."

But the story of neural networks tells us that we are likely to get further away from that objective in the future, rather than closer to it.

Inspired by the human brain

These neural networks may be complex systems yet they have some core principles. Inspired by the human brain, they seek to copy or simulate forms of biological and human thinking. In terms of structure and design they are, as IBM also explains, comprised of "node layers, containing an input layer, one or more hidden layers, and an output layer." Within this, "each node, or artificial neuron, connects to another." Because they require inputs and information to create outputs they "rely on training data to learn and improve their accuracy over time." These technical details matter but so too does the wish to model these systems on the

complexities of the human brain.

Grasping the ambition behind these systems is vital in understanding what these technical details have come to mean in practice. In a 1993 interview, the neural network scientist Teuvo Kohonen concluded that a "self-organizing" system "is my dream," operating "something like what our nervous system is doing instinctively." As an example, Kohonen pictured how a "self-organizing" system, a system that monitored and managed itself, "could be used as a monitoring panel for any machine … in every airplane, jet plane, or every nuclear power station, or every car." This, he thought, would mean that in the future "you could see immediately what condition the system is in."

The overarching objective was to have a system capable of adapting to its surroundings. It would be instant and autonomous, operating in the style of the nervous system. That was the dream, to have systems that could handle themselves without the need for much human intervention. The complexities and unknowns of the brain, the nervous system and the real world would soon come to inform the development and design of neural networks.

"Something fishy about it'

But jumping back to 1956 and that strange learning machine, it was the hands-on approach that Taylor had taken when building it that immediately caught Cowan's attention. He had clearly sweated over the assembly of the bits and pieces. Taylor, Cowan observed during an interview on his own part in the story of these systems, "didn't do it by theory, and he didn't do it on a computer." Instead, with tools in hand, he "actually built the hardware." It was a material thing, a combination of parts, perhaps even a contraption. And it was "all done with analog circuitry" taking Taylor, Cowan notes, "several years to build it and to play with it." A case of trial and error.

Understandably Cowan wanted to get to grips with what he was seeing. He tried to get Taylor to explain this learning machine to him. The clarifications didn't come. Cowan couldn't get Taylor to describe to him how the thing worked. The analog neurons remained a mystery. The more surprising problem, Cowan thought, was that Taylor "didn't really understand himself what was going on." This wasn't just a momentary breakdown in communication between the two scientists with different specialisms, it was more than that.

In an interview from the mid-1990s, thinking back to Taylor's machine, Cowan revealed that "to this day in published papers you can't quite understand how it works." This conclusion is suggestive of how the unknown is deeply embedded in neural networks. The unexplainability of these neural systems has been present even from the fundamental and developmental stages dating back nearly seven decades.

This mystery remains today and is to be found within advancing forms of AI. The unfathomability of the functioning of the associations made by Taylor's machine led Cowan to wonder if there was "something fishy about it."

Long and tangled roots

Cowan referred back to his brief visit with Taylor when asked about the reception of his own work some years later. Into the 1960s people were, Cowan reflected, "a little slow to see the point of an analog neural network." This was despite, Cowan recalls, Taylor's 1950s work on "associative memory" being based on "analog neurons." The Nobel Prize-winning neural systems expert, Leon N. Cooper, concluded that developments around the application of the brain model in the 1960s, were regarded "as among the deep mysteries." Because of this uncertainty there remained a skepticism about what a neural network might achieve. But things slowly began to change.

Some 30 years ago the neuroscientist Walter J. Freeman, who was surprised by the "remarkable" range of applications that had been found for neural networks, was already commenting on the fact that he didn't see them as "a fundamentally new kind of machine." They were a slow burn, with the technology coming first and then subsequent applications being found for it. This took time. Indeed, to find the roots of neural network technology we might head back even further than Cowan's visit to Taylor's mysterious machine.

The neural net scientist James Anderson and the science journalist Edward Rosenfeld have noted that the background to neural networks goes back into the 1940s and some early attempts to, as they describe, "understand the human nervous systems and to build artificial systems that act the way we do, at least a little bit." And so, in the 1940s, the mysteries of the human nervous system also became the mysteries of computational thinking and artificial intelligence.

Summarizing this long story, the computer science writer Larry Hardesty has pointed out that deep learning in the form of neural networks "have been going in and out of fashion for more than 70 years." More specifically, he adds, these "neural networks were first proposed in 1944 by Warren McCulloch and Walter Pitts, two University of Chicago researchers who moved to MIT in 1952 as founding members of what's sometimes called the first cognitive science department."

Elsewhere, 1943 is sometimes the given date as the first year for the technology. Either way, for roughly 70 years accounts suggest that neural networks have moved in and out of vogue, often neglected but then sometimes taking hold and moving into more mainstream applications and debates. The uncertainty persisted. Those early developers frequently describe the importance of their research as being overlooked, until it found its purpose often years and sometimes decades later.

Moving from the 1960s into the late 1970s we can find further stories of the unknown properties of these systems. Even then, after three decades, the neural network was still to find a sense of purpose. David Rumelhart, who had a background in psychology and was a co-author of a set of books published in 1986 that would later drive attention back again towards neural networks, found himself collaborating on the development of neural networks with his colleague Jay McClelland.

As well as being colleagues they had also recently encountered each other at a conference in Minnesota where Rumelhart's talk on "story understanding" had provoked some discussion among the delegates.

Following that conference McClelland returned with a thought about how to develop a neural network that might combine models to be more interactive. What matters here is Rumelhart's recollection of the "hours and hours and hours of tinkering on the computer."

We sat down and did all this in the computer and built these computer models, and we just didn't understand them. We didn't understand why they worked or why they didn't work or what was critical about them.

Like Taylor, Rumelhart found himself tinkering with the system. They too created a functioning neural network and, crucially, they also weren't sure how or why it worked in the way that it did, seemingly learning from data and finding associations.

Mimicking the brain—layer after layer

You may already have noticed that when discussing the origins of neural networks the image of the brain and the complexity this evokes are never far away. The human brain acted as a sort of template for these systems. In the early stages, in particular, the brain—still one of the great unknowns—became a model for how the neural network might function.

So these experimental new systems were modeled on something whose functioning was itself largely unknown. The neurocomputing engineer Carver Mead has spoken revealingly of the conception of a "cognitive iceberg" that he had found particularly appealing. It is only the tip of the iceberg of consciousness of which we are aware and which is visible. The scale and form of the rest remains unknown below the surface.

In 1998, James Anderson, who had been working for some time on neural networks, noted that when it came to research on the brain "our major discovery seems to be an awareness that we really don't know what is going on."

In a detailed account in the Financial Times in 2018, technology journalist Richard Waters noted how neural networks "are modeled on a theory about how the human brain operates, passing data through layers of artificial neurons until an identifiable pattern emerges." This creates a knock-on problem, Waters proposed, as "unlike the logic circuits employed in a traditional software program, there is no way of tracking this process to identify exactly why a computer comes up with a particular answer." Waters' conclusion is that these outcomes cannot be unpicked. The application of this type of model of the brain, taking the data through many layers, means that the answer cannot readily be retraced. The multiple layering is a good part of the reason for this.

Hardesty also observed these systems are "modeled loosely on the human brain." This brings an eagerness to build in ever more processing complexity in order to try to match up with the brain. The result of this aim is a neural net that "consists of thousands or even millions of simple processing nodes that are densely interconnected." Data moves through these nodes in only one direction. Hardesty observed that an "individual node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data."

Models of the human brain were a part of how these neural networks were conceived and designed from the outset. This is particularly interesting when we consider that the brain was itself a mystery of the time (and in many ways still is).

"Adaptation is the whole game'

Scientists like Mead and Kohonen wanted to create a system that could genuinely adapt to the world in which it found itself. It would respond to its conditions. Mead was clear that the value in neural networks was that they could facilitate this type of adaptation. At the time, and reflecting on this ambition, Mead added that producing adaptation "is the whole game." This adaptation is needed, he thought, "because of the nature of the real world," which he concluded is "too variable to do anything absolute."

This problem needed to be reckoned with especially as, he thought, this was something "the nervous system figured out a long time ago." Not only were these innovators working with an image of the brain and its unknowns, they were combining this with a vision of the "real world" and the uncertainties, unknowns and variability that this brings. The systems, Mead thought, needed to be able to respond and adapt to circumstances *without* instruction.

Around the same time in the 1990s, Stephen Grossberg—an expert in cognitive systems working across maths, psychology and bioemedical engineering—also argued that adaptation was going to be the important step in the longer term. Grossberg, as he worked away on neural network modeling, thought to himself that it is all "about how biological measurement and control systems are designed to adapt quickly and stably in real time to a rapidly fluctuating world." As we saw earlier with Kohonen's "dream" of a "self-organizing" system, a notion of the "real world" becomes the context in which response and adaptation are being

coded into these systems. How that real world is understood and imagined undoubtedly shapes how these systems are designed to adapt.

Hidden layers

As the layers multiplied, deep learning plumbed new depths. The neural network is trained using training data that, Hardesty explained, "is fed to the bottom layer—the input layer—and it passes through the succeeding layers, getting multiplied and added together in complex ways, until it finally arrives, radically transformed, at the output layer." The more layers, the greater the transformation and the greater the distance from input to output. The development of Graphics Processing Units (GPUs), in gaming for instance, Hardesty added, "enabled the one-layer networks of the 1960s and the two to three- layer networks of the 1980s to blossom into the ten, 15, or even 50-layer networks of today."

Neural networks are getting deeper. Indeed, it's this adding of layers, according to Hardesty, that is "what the 'deep' in 'deep learning' refers to." This matters, he proposes, because "currently, deep learning is responsible for the best-performing systems in almost every area of artificial intelligence research."

But the mystery gets deeper still. As the layers of neural networks have piled higher their complexity has grown. It has also led to the growth in what are referred to as "hidden layers" within these depths. The discussion of the optimum number of hidden layers in a neural network is ongoing. The media theorist Beatrice Fazi has written that "because of how a deep neural network operates, relying on hidden neural layers sandwiched between the first layer of neurons (the input layer) and the last layer (the output layer), deep-learning techniques are often opaque or illegible even to the programmers that originally set them up."

As the layers increase (including those hidden layers) they become even

less explainable—even, as it turns out, again, to those creating them. Making a similar point, the prominent and interdisciplinary new media thinker Katherine Hayles also noted that there are limits to "how much we can know about the system, a result relevant to the 'hidden layer' in neural net and deep learning algorithms."

Pursuing the unexplainable

Taken together, these long developments are part of what the sociologist of technology Taina Bucher has called the "problematic of the unknown." Expanding his influential research on scientific knowledge into the field of AI, Harry Collins has pointed out that the objective with neural nets is that they may be produced by a human, initially at least, but "once written the program lives its own life, as it were; without huge effort, exactly how the program is working can remain mysterious." This has echoes of those long-held dreams of a self-organizing system.

I'd add to this that the unknown and maybe even the unknowable have been pursued as a fundamental part of these systems from their earliest stages. There is a good chance that the greater the impact that artificial intelligence comes to have in our lives the less we will understand how or why.

But that doesn't sit well with many today. We want to know how AI works and how it arrives at the decisions and outcomes that impact us. As developments in AI continue to shape our knowledge and understanding of the world, what we discover, how we are treated, how we learn, consume and interact, this impulse to understand will grow. When it comes to explainable and transparent AI, the story of neural networks tells us that we are likely to get further away from that objective in the future, rather than closer to it.

Commons license. Read the original article.

Provided by The Conversation

Citation: AI will soon become impossible for humans to comprehend—the story of neural networks tells us why (2023, March 31) retrieved 25 April 2024 from https://techxplore.com/news/2023-03-ai-impossible-humans-comprehendthe-story.html