If it quacks like a duck ...

AI has captured the semantics of language, but does it have a mind?

MELANIE MITCHELL

THESE STRANGE NEW MINDS
How AI learned to talk and what it means
CHRISTOPHER SUMMERFIELD
384pp. Viking. £22.

In November 2022, OpenAI released ChatGPT, a large language model (LLM) that can converse in flawless natural language, answer questions, solve maths problems and logic puzzles, write essays and poetry, and explain complex ideas. Similar models have since been released by other AI companies. Shortly after ChatGPT's launch, the neuroscientist Terrence Sejnowski wrote:

> Something is beginning to happen that was not expected even a few years ago. A threshold was reached, as if a space alien suddenly appeared that could communicate with us in an eerily human way ... Some aspects of their behavior appear to be intelligent, but if it's not human intelligence, what is the nature of their intelligence?

Everyone connected to the field of AI is still grappling with this question, with no consensus on the answer. Are LLMs analogous to individual human minds (or, per Sejnowski, those of space aliens) – minds that think, reason and perhaps have their own beliefs, goals and intentions? Or is it misguided to frame an LLM as an intelligent agent; are they instead "cultural technologies", accessible information-organization systems that are more like libraries and the internet than individual minds? These framings exemplify different positions on a spectrum of views in the AI community, which has become increasingly polarized over how to think about these surprising and confusing artefacts.

The early pioneers of AI intended machines to be human-like in their intelligence, and were optimistic about the timeline for this being achieved. In 1965, the future Nobel laureate Herbert Simon predicted that "Machines will be capable, within twenty years, of doing any work that a man can do". In 1970, Marvin Minsky, the founder of MIT's AI Lab, was even more sanguine: "In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, have a fight".

This early optimism was soon dashed by the disappointing results of AI in the real world. In 1973, the UK government solicited a report on the state of AI, which turned out to be devastatingly negative about the prospects of general machine intelligence. The report's author put it this way: "Always there may be some people who try to make us think we can see that old general-purpose robot shimmering there on the horizon, but he's a mirage". Similar negative sentiments led to the first "AI winter", a period of low expectations (and low funding) for AI research and commercialization. In subsequent years, the field cycled between optimistic springs and gloomy winters, the latter reflecting the AI pioneer John McCarthy's admission that "AI was harder than we thought". While there were endless debates as to which methods – logic-like rules or brain-like neural networks – would succeed in yielding "true" AI, none of the proposed approaches produced anything close to human-level performance except in circumscribed domains such as playing chess or detecting spam emails.

All this started to change around 2010, with what is called the "deep learning revolution". Deep learning refers to a machine learning approach in which large amounts of data are used to train "deep" neural networks. Neural networks are AI systems with structures loosely inspired by biological brains: simulated "neurons" are linked to one another via weighted connections and are arranged in hierarchical layers. The more layers, the "deeper" the network. Typically, the network is trained to map an input (eg a word or an image) to a "correct" output (eg the sound of the word or the name of an object in the image). Such training usually requires a large dataset of examples, which are used to tune the network's weighted connections to values that will produce correct outputs.

Deep neural networks have been around since the 1970s, but not until the 2010s was the field able to exploit the nexus of large amounts of available training data (via the world wide web), powerful parallel computing chips and new methods for effectively training neural networks. Suddenly, deep learning worked much better than any AI method in the past, in domains including, among others, computer vision, speech recognition and machine translation.

Another big advance occurred in 2017 with the invention of the "transformer" network, a type of deep neural network that is particularly suited to learning sequences, such as those of words in a text. Transformers now form the basis for all of today's large language models, which are initially trained simply to predict the next word in a given sequence. The training sequences come from web pages, newspapers, books, text messages, video meeting transcripts and any other digitized text that AI companies can get their hands on. Somehow, via training on the vast repository of written human communication and through their enormous size (billions or trillions of weighted connections, or "parameters"), LLMs have developed general abilities for using language. But even though these systems are engineered by humans and trained on human-generated language, their scale and complexity make it difficult to understand exactly how they achieve their sophisticated behaviour, even for the engineers who created them.

In *These Strange New Minds: How AI learned to talk and what it means*, the cognitive neuroscientist Chris Summerfield surveys the philosophical landscape, social impacts and potential dangers of LLMs. In the first parts of the book, he describes the intellectual history of neural networks and provides readers with an intuitive account of what LLMs are and how they are trained.

Summerfield explains how the language abilities of LLMs serve as strong evidence for a linguistic theory called "distributional semantics", which proposes that the meaning of language can be derived from the statistics of how words occur together in text. For example, we understand the meaning of the word "bill" from its context: if "bird" and "beak" are mentioned nearby, "bill" probably has one meaning; if "plumber" and "cost" are mentioned nearby, it probably has another meaning. Distributional semantics posits that words and phrases can be mapped to a high-dimensional "semantic space"; the position of a word or phrase in that space, and its distance to other words or phrases, are what define its meaning. Creating such a semantic space by learning word co-occurrence statistics from vast amounts of training data is what LLMs do, and in that sense they can be said to have captured something of the semantics of language.

In addition to supporting the theory of distributional semantics, LLMs have had another impact on linguistic science: their success has put the final nail in the coffin of Noam Chomsky's theory of language acquisition. Summerfield describes in detail how Chomsky, the most influential of linguistic theorists, asserted that language is unique to humans and cannot be acquired without some sort of innate mental structure that is predisposed to learn syntax, termed a "universal grammar" and possessed by all humans at birth. LLMs – which lack any engineered structure specific to language – represent a fatal counterexample to the theory, since they have successfully acquired language entirely by learning from data.

While it is indisputable that LLMs have human-level capabilities for generating language, does this mean that we should think of them as "minds", albeit strange ones? This seems to be Summerfield's position, evident in the title of the book. While the author acknowledges that there are no agreed-on definitions of mental terminology such as "thinking", "understanding", and "meaning", he doesn't see any reason why we should not grant mental status to LLMs, even though they have acquired such capacities in ways in different ways to humans. "[M]eaning can be acquired via two different routes", he writes:

> There is the high road of linguistic data, in which we learn that "spider" goes with "web". Then there is the low road of perceptual data, in which we catch sight of an eight-legged insect at the centre of a geometric lattice, glinting in the morning dew. Most people have the luxury of travelling down both routes, and so can learn to connect words with words, objects with objects, words with objects, and objects with words. LLMs that are trained exclusively as chatbots, by contrast, can only travel on the high road – they can only use linguistic data to learn about the world. This means that any thinking or reasoning that they might do will inevitably be very different from our own.

He further argues that LLMs can also be said to possess *knowledge* in a humanlike sense, in spite of their disembodied nature:

> Each of the major language models knows vastly more than each one of the eight billion humans alive, without having ever taken the tiniest peek at the natural world in which we all live.

Such statements seem to put Summerfield in the "LLMs as intelligent agents" camp, as opposed to the "LLMs as cultural technologies" camp; it would be strange to say, for example, that the Library of Congress "knows" more than any human alive, or that it "understands the meaning of its books". "Knowing" and "understanding" are mentalistic terms – they are not appropriate categories for describing libraries or most other cultural technologies.

Why use anthropomorphic mentalistic terms to describe AI systems? Summerfield argues that

Melanie Mitchell is a Professor at the Santa Fe Institute and the author of Artificial Intelligence: A guide for thinking humans, *2019*

such terms are appropriate given the behaviour of these systems:

> We should subject LLMs to the so-called Duck Test: if something swims like a duck and quacks like a duck, then we should assume that it probably is a duck, rather than inventing abstruse arguments to otherwise explain its behaviour.

The problem with this, as the author acknowledges, is that humans have a strong, sometimes misleading tendency to project mental qualities onto anything that communicates with us in fluent natural language. This has been dubbed the "Eliza effect", named after the 1960s chatbot Eliza, which imitated a psychotherapist. Even though Eliza had zero intelligence or understanding – it used templates such as "Tell me more about X", where X was something a user just mentioned – people who chatted with it often believed that it understood them deeply. While today's LLMs are far more sophisticated language users than Eliza, to what extent are we humans similarly falling for an imitation of understanding and intelligence rather than recognizing the "real thing", especially when we don't have consensus about what "the real thing" is?

Summerfield spends considerable space in the book on a debate between what he terms "equivalentists" and "exceptionalists". The former are those who believe that there are no reasons *in principle* that machines cannot be intelligent, or have understanding, beliefs, intentions and so on, whereas the latter insist that there is something special about the minds of humans (and perhaps other biological organisms) that is fundamentally missing in non-biological entities, which means that using mental terms to describe machines will always be a category error.

This debate is real and has been played out in philosophy circles for centuries. Summerfield, however, seems to conflate the "in principle" debate (whether some kind of machine could *ever* think, believe, desire, etc) with a wholly different "in practice" one – whether *today's* LLMs have or could have these qualities. People who argue against the latter are, unfairly I think, painted with the "exceptionalist" brush.

Indeed, the author's insightful descriptions of the many differences between LLMs and humans could be taken as an argument that LLMs remain quite far from our usual notions of what it is to have a mind. While the most recent version of ChatGPT is perhaps "the most complex software object ever made", in the words of the journalist Ross Andersen, it arguably has no sense of self or personal identity, no motivations, no memory of its own experiences (if it could be said to have any "experiences" at all), no feelings or emotional states, no survival imperatives; in short, it lacks many of the attributes usually associated with having a mind. As the philosopher Shannon Vallor put it, "they can answer the questions we choose to ask, paint us pretty pictures, generate deepfake videos and more. But an AI tool is dark inside"

Summerfield maintains that, in spite of all that's missing, the correct way to think about LLMs is as *minds* that engage in *thinking*. "The minds of LLMs are not like ours", he notes, "but they are minds, of sorts, nonetheless – strange new minds, quite unlike anything we have encountered before." Furthermore, "to say that LLMs do not think at all requires a new and rather convoluted definition of what it means to 'think'". Beyond their passing his personal "Duck Test", however, Summerfield does not, in my view, argue convincingly that today's LLMs have earned the right to these terms.

One might ask why it matters how we describe these systems. Who cares whether an LLM is called a "mind that thinks" or a "cultural technology, more like a library or the internet"? It turns out that the way we frame these systems and the metaphors we use to conceptualize them have important consequences for determining what we expect of them (can one, for example, have a romantic relationship with an LLM?) and how we treat them in legal and regulatory decisions (eg a mind that reads a book and thinks about it is not infringing copyright, whereas a library that hosts illegally copied materials is doing so).

> **"It would be strange to say that the Library of Congress 'knows' more than any human alive, or that it understands the meaning of its books**

The final sections of the book focus on the social impacts and dangers of AI. While there are mentions of beneficial applications, in medicine and science, for example, these pages are mostly concerned with harms, both existing and speculated. And the harms are abundant. Humans trusting LLMs with critical tasks in law, medicine and other fields can be burnt by their "confabulations" and other unexpected failures. LLMs exhibit both overt and subtle racial and gender biases; they can and are being used to generate misinformation, to perpetuate scams and to flood the internet and jam up search engines with useless "AI slop". Summerfield surveys these and many other present harms, as well as those we can look forward to in the near future, such as "personalized" AI assistants that can be used to manipulate users, AI agents that can be hacked to obtain sensitive personal information and LLMs masquerading as real people.

Then there are the more speculative, science fiction-like dangers that Summerfield discusses in the final section, entitled "Are We All Doomed?" Here we learn about the "alignment problem": if we ask a (still imaginary) superintelligent AI system to solve a problem, say to fix global warming, how can we ensure that it doesn't destroy us as part of its solution? How can we prevent a self-improving AI system from developing a self-preservation instinct, in service of which the system will inevitably try to amass as much power and as many resources as possible, dispatching humans that get in its way and manipulating other humans to do its bidding? Such scenarios are the focus for a community of "AI safety" researchers concerned about "existential risk". On the other side of the coin are the "effective accelerationists – extreme techno-optimists who believe that the benefits of AI will be so great that society should amplify its development and avoid slowing down progress with, say, government regulations. OpenAI's Sam Altman, for example, wrote that AI investments will eventually lead to a utopia in which "astounding triumphs – fixing the climate, establishing a space colony, and the discovery of all of physics – will eventually become commonplace". Summerfield describes the profound disagreements between these two communities, and with a third community, which he calls the "#AIhypers", who believe that AI progress is overhyped and that we should be more focused on mitigating existing AI harms than on highly speculative scenarios with little evidence to back them.

These Strange New Minds is an entertaining and enlightening work with broad scope. Like the field of AI itself, it brings together philosophy, cognitive science, neuroscience, engineering and sociology in order to make sense of LLMs and their possible impacts. Christopher Summerfield lays out many of the key debates in the field and is clear (if not wholly convincing) about which side he takes on questions of AI thought. The philosopher Bertrand Russell recalled how, in his youth, his grandmother used to dismiss metaphysics whenever it was mentioned with the witticism: "What is mind? No matter. What is matter? Never mind". As this book illustrates, in the age of AI, defining the concept of *mind* will matter enormously. ■