Neil Dodgson

Artificial intelligence
ChatGPT and human gullibility

Abstract

Artificial intelligence (AI) has advanced rapidly in the past decade. The arrival of ChatGPT last year has pushed the debate about AI into the public sphere. ChatGPT, and similar tools, do things we once thought were outside the ability of computers. This raises questions for how we educate people about the capability and the limitations of such tools. This article provides an overview of artificial intelligence and explores what ChatGPT is capable of doing. It also raises questions about morality, responsibility, sentience, intelligence, and how humans' propensity to anthropomorphise makes us gullible and thus ready to believe that this technology is delivering something that it cannot.

Keywords artificial intelligence, ChatGPT, critical thinking, morality, consciousness

Artificial intelligence has made massive strides in the past decade, and particularly in the past year. We now have systems, like ChatGPT, that are doing things that we thought were beyond the capabilities of computers. ChatGPT is writing essays that would get reasonable grades if submitted by a high school student. DALL.E is creating artwork far more quickly than a trained graphic designer could. Google's LaMDA convinced Google engineer Blake Lemoine that it was sentient. After exchanging chat with LaMDA, he said, 'If I didn't know exactly what it was, which is this computer program we built recently, I'd think it was a 7-year-old, 8-year old kid that happens to know physics' (de Cosmo, 2022). Other experts are also convinced that machines are approaching sentience. But many others are convinced that there is no evidence of sentience and quite a lot of evidence against it. Today's large neural networks are producing impressive results, but what are they really doing?

Neil Dodgson is Professor of Computer Graphics in the School of Engineering and Computer Science at Victoria University of Wellington. His research is in 3D TV, mathematical modelling for 3D shape design, the aesthetics of imaging, the social psychology of colour and the use of AI in image processing.

A brief history of AI

Artificial intelligence (AI) has been around since the dawn of computing. In the 1950s there was great hope that we could create artificial intelligence quickly. The thinking was: if people can do something, surely it will be easy to get a computer to do that same thing. That turned out to be incorrect: human behaviours are complicated. For example, understanding natural spoken language is something most children can do easily. Early on in computing, it was expected that speech recognition would be solved by the 1960s. It turns out to be stunningly difficult to get a computer to do speech recognition. It is only in the last 20 years that we have got computers to reliably recognise speech, 50 years later than we expected to be able to.

In the early days of AI, computing researchers hand-coded systems that were based on the way they thought that humans did high-level reasoning. They were successful in getting computers to do some things that humans find difficult, like play chess. It is easy to get a computer to play chess because there are straightforward rules and clear guidelines for what constitute good and bad positions. But the way we got computers to beat us at chess was for them to take a very different approach compared with humans. Humans use experience, practice and intuition to guide them in considering a small number of possible good moves. A computer, by contrast, uses brute force to work through thousands of possibilities, far more and far more rapidly than a human could. Computers have been able to beat humans at chess for decades, but they do it in quite a different way to how humans play. A chess-playing computer is just a box of tricks doing exactly what we tell it to. It cannot do anything other than play chess.

Today we are well beyond the point of getting computers to play abstract games. The reason AI is so much in the news now is because of the phenomenal recent advances. After decades of research, there was a big breakthrough about 15 years ago in how we do AI, when advances in the speed of computer hardware allowed the technique of 'deep learning' to become practical for solving real-world problems. In deep learning, we build computer systems that mimic the way we think the human brain is constructed: with many layers of artificial neurons, each layer communicating to the next through lots of connections. The 'deep' in 'deep learning' comes from the fact that the neural network is many layers deep. The 'learning' part comes because we train this neural network by feeding it an enormous amount of data. That is, we give it lots of different inputs and, for each input, we tell it what the correct output should be. The system then tweaks its internal neurons and their connections based on the difference between what it actually output and what it was told is the correct output. For example, if we wish to build a system that can identify what animal appears in any given photograph, we would start with a blank neural network and train it by giving, as input, a series of photographs of animals, along with information about whether the animal is a cat, dog, goat, bear, etc. The system predicts what animal is in the photograph and checks this against the correct answer. Early on in the training it is making pure guesses. When it gets it wrong, it does tiny internal adjustments of the settings of millions of internal parameters to give a higher probability of being right if it sees a similar input in future. When it gets it right, it does tiny internal adjustments to strengthen the settings that got things right. As the training progresses, the chance of getting the right answer improves: it gets better at predicting the correct outcome.

The New Zealand Law Society recently notified its members of a similar situation in their discipline: ChatGPT was able to create case notes that sounded plausible and read well ...

So, deep learning creates what could be called a 'prediction machine'. The system predicts the output based on its input and its past training. With enough training data, you can get a deep neural network to then give the right answer to inputs that it has never seen before: it has 'learnt' how to solve that particular problem. You do need a lot of training data to get this right. In the case of training to spot animal species, you need millions of labelled images to train it to get good accuracy. Compare this to a human child, who can generalise the concept of 'cat' after meeting just a couple of cats.

ChatGPT

ChatGPT is a 'prediction machine', as are all the similar AI chatbot systems that can generate surprisingly good text. They are large, deep neural networks, trained on a phenomenal amount of input data gathered from across the internet. Their job is to predict what the next word will be in their conversation with the user. They do this using the context of the previous several thousand words in the conversation. With good training, which they have, and a big enough context, they can produce stunning results. For example, I got ChatGPT to write a 100-word marketing blurb for my university and it produced something that could have come straight out of our marketing department. I deduce that there is a lot of marketing copy in its training data. I also asked it to write a biography of me. It wrote a beautifully crafted biography, in exactly the right style, but it got over half the facts wrong. It knows what a biography should look like, but it essentially just puts together random facts that sound right. For example, it said I had worked at two universities I have never even visited, and that my PhD is in a completely different topic from what I really did. However, if you did not know better, it would sound right.

The New Zealand Law Society recently notified its members of a similar situation in their discipline: ChatGPT was able to create case notes that sounded plausible and read well (Holt, 2023). The system 'knows' what a case name and citation should look like, but it generated completely fake cases. The references look right but the cases to which they refer do not exist. This problem, which is technically known as 'hallucination', is common across chatbots. They are trained to produce good-sounding text but they are doing this by simply placing one word after another in a sequence driven by probability. They are not drawing on facts. While the output sounds plausible, there is nothing that checks its veracity. So beware: if you use one of these tools to write a paper for you,

you are still going to have to check all the facts.

How is it that ChatGPT can write so convincingly?

Those of us who write for a living, including those who devise and prepare policy, find ChatGPT challenging. It is doing something (writing) that we have been trained to do and that many of us find challenging to do well. Humanity has been here before. Weaving machines challenged human weavers: here was a machine that could do their job faster and more accurately than they could. The Luddites smashed some of the weaving machines, but they did not stop progress. Mechanical diggers obviated the need for armies of navvies to dig our roads. Automatic computers sped up bookkeeping and made some human skills redundant, such as adding long columns of numbers by hand.

We should not be surprised that we are now seeing computers that can do things that we thought were beyond their abilities. ChatGPT can generate grammatically correct essays that read well. It also produces reasonable poetry in a range of styles, from rap to haiku to sonnet. For example, asking it to rewrite Hamlet's 'To be' speech as a haiku produced this:

To be or not to be
Life's mysteries I ponder
Death, my final peace.

That is not bad. A schoolchild could have done this too, with a little bit of training. You might have noticed that ChatGPT has the wrong number of syllables in the first line. A little more investigation uncovers that ChatGPT is terrible at counting. In fact, it cannot count at all. This is because it is a language model and it has no mechanism for calculating.

Nevertheless, as a large language model, ChatGPT does produce text that reads well. This is because it is trained on literally billions of examples of text, much of which is well-written. Its training database included hundreds of thousands of publicly available books, all written and edited by humans. However, ChatGPT is achieving its success in a different way to a human. An English or History graduate spends several years learning how to structure a good essay, but they do not acquire this skill by reading millions of other people's essays. Instead, they read a few examples, generalise their skills from those examples, and hone their skills by trial and error: giving it a go, taking feedback, getting better each time. ChatGPT does not work this way. Take this article as an example. You are reading the seventh revision of this human-written work. I invested considerable planning and thought in constructing the arguments and refining the text. ChatGPT would have done none of this, instead simply putting down one word after another in a probabilistic sequence. As with chess-playing computers, there is a substantial difference between ChatGPT's 'prediction machine' method and what a human does.

Professor Edsger Dijkstra, eminent computer scientist and sceptic about AI, expressed it by analogy: to paraphrase him, asking if a computer can think is like asking if a submarine can swim ...

And ChatGPT has limits. As I said above, ChatGPT does not fact check. Indeed, it cannot fact check; it is just a prediction machine working off the probabilities that tell it what word should come next. If it is writing about something for which it has a lot of source data, it tends to produce correct facts on the pure probabilities because it has been trained on a lot of input with the correct facts in it. If it is writing about something more obscure (such as that biography of me), its 'prediction machine' just invents things that sound plausible. It may seem to be operating like an undergraduate skimping on their fact-checking when writing an essay at three in the morning, but that analogy is still wrong: ChatGPT does not have any underlying thought process. ChatGPT truly is just sticking down one word after another in a probabilistic sequence. It is the human reader who is imputing meaning to its probabilistic ramblings, and we humans are gullible if we assume that there is a thought process behind ChatGPT's utterances, because there is not.

ChatGPT is not thinking

We have not (yet) developed a thinking machine. What we have are prediction machines, giving you their best guess at what comes next based on what they have seen before. But, given their performance in writing, where their style outstrips that of a smart human child, people are reasonably asking whether these AI systems could lead to thinking machines, or whether this 'prediction machine' method is a dead end in our search to create a truly intelligent machine.

There is a reductionist view of consciousness that says that humans themselves are just prediction machines, albeit rather more complex than current AI systems. If this reductionist model is correct, then the brain is nothing more than a biological computer and there is no reason why a sufficiently complex digital computer could not develop consciousness to the same level as a human, or higher.

Many people are not comfortable with this reductionist view of consciousness. Our instinct is that humans are something more than just a biological computer. The experts are divided on whether the prediction machine method will lead to true thinking machines. Some experts see evidence, in ChatGPT and more sophisticated models, of emergent behaviour: the ability to do things that should not be possible simply from the underlying model. Others are sceptical. Professor Edsger Dijkstra, eminent computer scientist and sceptic about AI, expressed it by analogy: to paraphrase him,

asking if a computer can think is like asking if a submarine can swim (Dijkstra, 1983).

How do you test whether a computer can think?

Given that there is this debate about whether we can create a machine that truly thinks, how would we go about telling if a computer is thinking, or self aware, or conscious?

The standard response is to say we should use the Turing Test. This was designed by Alan Turing, one of the pioneers of computer science, as a test of whether a machine could exhibit intelligent behaviour indistinguishable from that of a human, but it has substantial limitations. In his 1950 paper (Turing, 1950), Turing considered the question, 'Can a machine think?' Acknowledging that we have a problem with defining what we mean by 'think', he replaced the question with the closely related question, 'Can a machine do what a thinking human can do?' In the Turing Test an entity, either a human or a computer, communicates with someone via a text interface. The computer passes if it can convince the recipient that it is really a human.

Turing never explicitly said that the Turing Test could be used as a measure of intelligence, or, indeed, of anything other than the machine being able to emulate a human to the extent needed to fool a human. In terms of passing the test, the first system to do so was ELIZA, developed by Joseph Weizenbaum in 1966 (Weizenbaum, 1966). ELIZA's most successful variant was based on Rogerian psychotherapy, where the therapist repeats the patient's statements back to them as questions. For example, if the patient says, 'I always had problems getting on with my mother', the therapist might respond, 'Tell me more about your mother'.

ELIZA convinced some participants that it was human (Natale, 2019), even though it was based on a simple parlour trick. It was easy to get it to spout nonsense if you had a modicum of understanding of how it worked. However, even those who knew how ELIZA worked would sometimes treat it as if it were a human therapist. Humans have a strong propensity to anthropomorphise and to be able to suspend their disbelief: we are remarkably gullible, and not just with computer systems. We anthropomorphise our pets, imputing human emotions and thought processes when we are seeing only instinct and habit; and we assume other human's motives and feelings on the very sketchy evidence of their facial expression, body language and utterances. We do this because it helps us to make sense of the world and guides our interactions with others.

[Sébastien Bubeck's] statement [on the definition of intelligence] says that intelligence requires evidence of six things: reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, and learning quickly and from experience.

Fifty-six years after ELIZA first fooled a few people, we have a professional computer scientist, Blake Lemoine, convinced that a modern computer system, Google LaMDA, is sentient, even though he knows how the system is programmed and other experts are convinced he is wrong. It looks as if the Turing Test is not useful.

Indeed, the Turing Test is concerned only with how the subject acts; that is, its external behaviour. The example of ELIZA shows that a computer program can demonstrate the right behaviour with no intelligence or consciousness behind it. I argue that ChatGPT and LaMDA are the same. They are much more complex than ELIZA, but they are simply responding to stimuli as prediction machines; they have no internal sense of self, no intelligence, no consciousness.

Sébastien Bubeck, of Microsoft Research, spoke at MIT in March this year about whether GPT-4 is intelligent (Bubeck, 2023a; see also Bubeck, 2023b). GPT-4 is the successor to ChatGPT and is much more powerful. He based his definition of intelligence on a 1997 statement signed by 52 professors in the field of intelligence (Gottfredson, 1997). That statement says that intelligence requires evidence of six things: reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, and learning quickly and from experience. Bubeck demonstrates that GPT-4 can do four of these well. It is not so good at learning, and it really cannot plan, but it definitely shows signs of being intelligent at a level beyond the abilities of most children. This might be a sign of emergent behaviour.

But reasoning and intelligence are not the same as sentience, or consciousness, or being self-aware. Can we go beyond intelligence to ascertain whether a machine is conscious? How can you know that any animal, other human being, or anything that seems conscious is not just faking it? How do you know whether it enjoys an internal subjective experience, complete with sensations and emotions like hunger, joy or sadness? We lack what neuroscientist Christof Koch has called a consciousness meter – a device that can measure consciousness in the same way that a thermometer measures temperature.

There is much work going on worldwide in understanding consciousness, including work at Victoria University of Wellington (Bareham et al., 2020). Tamara Hunt (University of Melbourne) and Jonathan Scholler (University of California, Santa Barbara) are developing a framework to think about the different possible ways to test for the presence of consciousness. They

use three types of test: brain activity that matches the subject's reported subjective states; physical actions that seem to be accompanied by subjective states; and creative products that provide evidence that a conscious being produced them. All are interesting. All tell us that a human is conscious. The first two would tell us that cats are conscious. The latter might limit us to higher animals, like elephants, who can create. By some of these tests, ChatGPT is conscious. And yet we know that it is simply following its programming, a prediction machine giving its best guess of what comes next based on what has gone before.

The final analysis is that we currently have no reliable way to tell whether a machine is conscious or not. Even if a machine tells us repeatedly that it feels, that it is self-aware, that it loves, we have no way of knowing whether it truly is conscious or just a clever trick giving the impression of consciousness.

Can a computer be a moral agent?

What sort of morality could you instil in a computer? Humans learn their moral code from the society in which they grow up. Some theories of consciousness require that the subject be immersed in a sufficiently rich social environment to develop consciousness, specifically a nurturing environment in which you can learn how to make an internal model of yourself by observing others (Rahimian, 2021). Would a conscious computer need a social environment in which to develop? Arthur C. Clarke's masterwork, *2001: A Space Odyssey*, has that famous fictional example of a thinking, conscious computer, HAL. There is an implication in the film that HAL had to be instructed like a child to bring it up to full operation. While fictional, it raises the question of whether this is what we will need to do to create a truly conscious computer.

Can we program an ethical or moral code into a computer? In 1942, Isaac Asimov imagined a moral code for robots embodied in three laws:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Asimov says that these three laws 'are the only way in which rational human beings can deal with robots – or with anything else'. (Asimov, 1981)

While the laws make sense, there are problems. First, there is the technical problem of whether and how we could embed these laws into a thinking machine in a way that would guarantee that the laws would be followed. Second, there is the regulatory problem of whether we could guarantee that all systems will have the laws embedded. It is all too easy to imagine a military robot that is explicitly not Asimov-compliant. Third is the moral problem of the status of sentient computers under Asimov's laws. Are we to make computers that are permanent slaves? How would we justify a society where we enslave sentient, conscious machines?

The question of what might happen if machines go wrong (i.e., behave immorally) can be considered from a different perspective by considering what happens when *humans* go wrong.

In what ways might things go wrong?

There is a great deal of dystopian literature considering how artificially intelligent machines can go wrong. These often focus on sentient killer robots (e.g., the *Terminator* and *Matrix* movies), but I believe there are challenging problems that face the world right now, with the level of artificial intelligence that we already have.

The question of what might happen if machines go wrong (i.e., behave immorally) can be considered from a different perspective by considering what happens when *humans* go wrong. We learn a lot about what it means to be human from the exceptions, and they give a warning of what might go wrong if we get the morality wrong in machines. Consider those unwelcome personality types: the bully, the narcissist, the psychopath. All of these personalities can be held by a person who is perfectly able to function in society, including by people who rise to leadership roles. Psychopaths can be quite charming when it suits their ends, but they have poor empathy, are manipulative, and even when they get good results everyone is wary of them. Humanity produces such people at a rate of about one in 100 (Burton and Saleh, 2020). When they get into positions of power they can be tremendously disruptive, as demonstrated in various regimes in the past century. These people take advantage of the cultural and moral norms to disrupt society to their own ends. Consider Joseph Stalin, who manipulated and controlled the Soviet Union for decades, with psychopathic cruelty. Was he clinically insane? Or was he, as one author has put it, 'a very smart and implacably rational ideologue'? (Appelbaum, 2014). I find that latter characterisation chilling because 'very smart and implacably rational' is a good description of a computer.

Imagine a Machiavellian or psychopathic computer that had control over the financial services or the military hardware of a country. It could do so much wrong, as imagined in the *Terminator* movies, where the computers are given control of weapons systems. While that remains the go-to message when people think about intelligent computers (see, for example, the campaign against killer robots, https://www.stopkillerrobots.org/), there

is a more prosaic and insidious problem that is already happening. AI systems are taking increasing control of decisions about how humans live their lives. For example, an AI system likely decides whether you get life insurance and how much it should cost; or whether you should be given a mortgage. AI systems decide which videos to recommend to you on YouTube, what websites to suggest on Google Search, and where you should go next on social media. There are regimes that are experimenting with 'social credit' systems, where people's monitored behaviour feeds into a credit score that determines what they are and are not allowed to do. A good score might lead to opportunities for better jobs, better housing and travel; a poor score might block those opportunities. Human beings alone could manage a system like this only with considerable personnel, bureaucracy and paperwork (think Cold War East Germany), and, even then, it would be a relatively blunt instrument. Artificial intelligence allows for more efficient, more fine-grained control over human behaviour. This is what George Orwell was hinting at in *Nineteen Eighty-Four*, where even the Inner Party members were controlled by the system, but even Orwell did not imagine just how much control you could have if you use computers to do the monitoring for you.

This is the position we are in right now. We do not need to wait for computers to be sentient for them to exert considerable control over our lives. Human-run organisations are using artificial intelligence to improve their profit margins and their market share. We consumers are equally culpable, making conscious and rational decisions to engage with these systems because of the perceived benefits we receive in return. We are already in a world where artificial intelligence combined with human intelligence is controlling what we do. Corporations and governments are using the existing tools to modify and manipulate human behaviour. We need to develop policy now. We do not need to wait for AI systems to improve, or to demonstrate sentience. Humans and computers combined already create a world that is different from the world where everything was controlled by humans alone. The speed with which computers can calculate has enhanced what humans can do with their brains, in the same way that mechanical machines enhanced what humans can do with their muscles.

Going beyond today, we can imagine a world where a computer gives the *illusion* of sentience, given that we cannot test whether or not it really is sentient, and given humans' tendency to anthropomorphise. If enough people *believe* the computer to be sentient, this could significantly affect how human society behaves and develops; in the same way that, if a psychopath became leader of a major nation today, and if enough people believed in them, it would cause substantial disruption to the entire nation, not just to those who believed. So the question is not whether a computer can be sentient (which we cannot prove), but whether humans can believe that a computer is sentient (which is all too likely) and how they will react and respond to a computer that they believe to be sentient.

References

Appelbaum, A. (2014) 'Understanding Stalin', *Atlantic*, https://www.theatlantic.com/magazine/archive/2014/11/understanding-stalin/380786/

Asimov, I. (1981) 'I. Guest commentary: The three laws', *Compute!*, 18, p.18, https://archive.org/details/1981-11-compute-magazine/page/n19/mode/2up?view=theater

Bareham, C.A., M. Oxner, T. Gastrell and D. Carmel (2020) 'Beyond the neural correlates of consciousness: using brain stimulation to elucidate causal mechanisms underlying conscious states and contents', *Journal of the Royal Society of New Zealand*, 51 (1), pp.143–70, https://doi.org/10.1080/03036758.2020.1840405

Bubeck, S. (2023a) 'Sparks of AGI: early experiments with GPT-4', video, https://www.youtube.com/watch?v=qbIk7-JPB2c

Bubeck, S. (2023b) 'Sparks of AGI: early experiments with GPT-4', https://arxiv.org/pdf/2303.12712.pdf

Burton, B. and F.M. Saleh (2020) 'Psychopathy: insights for general practice', *Psychiatric Times*, 27 (10)

de Cosmo, L. (2022) 'Google engineer claims AI chatbot is sentient: why that matters', *Scientific American*, 12 July, https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

Dijkstra, E.W. (1983) 'The fruits of misunderstanding', *Elektronische Rechenanlagen*, 25, pp.286–9, https://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD854.html

Gottfredson, S. (1997) 'Mainstream science on intelligence: an editorial with 52 signatories, history, and bibliography', *Intelligence*, 24 (1), pp.13–23, https://doi.org/10.1016/S0160-2896(97)90011-8

Holt, T. (2023) 'The curious case of ChatGPT and the fictitious legal notes', *Stuff*, 31 March, https://www.stuff.co.nz/dominion-post/news/wellington/131658119/the-curious-case-of-chatgpt-and-the-fictitious-legal-notes

Hunt, T. (2019) 'How can you tell if another person, animal or thing is conscious? Try these 3 tests', *The Conversation*, 1 July, https://theconversation.com/how-can-you-tell-if-another-person-animal-or-thing-is-conscious-try-these-3-tests-115835

Natale, S. (2019) 'If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA', *New Media and Society*, 21 (3), pp.712–28, http://doi.org/10.1177/1461444818804980

Rahimian, S. (2021) 'Consciousness in solitude: is social interaction really a necessary condition?', *Frontiers in Psychology: Consciousness Research*, 12, https://doi.org/10.3389/fpsyg.2021.630922

Turing, A.M. (1950) 'Computing machinery and intelligence', *Mind*, 59 (236), pp.433–60, https://doi.org/10.1093/mind/LIX.236.433

Weizenbaum, J. (1966) 'ELIZA: a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, 9 (1), pp.36–45, https://dl.acm.org/doi/pdf/10.1145/365153.365168