MUTANT ALGORITHMS | TOBY WALSH

PODCAST TRANSCRIPT

UNSW Centre for Ideas: Welcome to 10 Minute Genius, an eight-part series created by the UNSW Centre for Ideas, to provide pause and create a space to engage with new ideas from UNSW Sydney's thinkers, dreamers and envelope pushers, as they help to make sense of the relentless information vortex in which we live. In under 10 minutes, or roughly the same amount of time it takes a computer to win a million games of chess, Scientia Professor of Artificial Intelligence Toby Walsh, will explore how we can make sure mutant algorithms don't go too far.

Toby Walsh: Recently, people in the UK took to the streets to protest mutant algorithms. These algorithms are being used to decide grades for school leavers, who couldn't sit their A level exams due to the ongoing pandemic. This was artificial intelligence assessing human intelligence, and it didn't go to plan. So, do algorithms really mutate? Can AI be devious? And how can we be sure that we don't lose the human touch? When we get zeros and ones to do the work for us? Artificial intelligence or AI has become a part of our everyday lives, from home robots to smartphones, answering your Trivial Pursuit questions at the touch of a button. Siri, who wrote *Waiting for Godot*?

Siri: *Waiting for Godot*, was written by Samuel Beckett in 1952.

Toby Walsh: Humans aren't the only ones who can write a play. OpenAI's GPT-3 just wrote a whole play, live on the stage of London's Young Vic Theatre. And in 2020, some friends here at UNSW, and at a music production company called Uncanny Valley developed AI that won the very first AI Eurovision Song Contest.

But not all developments in AI have been so good. Hence the mutant algorithm problem. Boris Johnson coined the term mutant algorithms, when it became clear that certain people,

particularly those from private fee-paying schools, were more likely to be given higher grades by the AI being used. But algorithms don't mutate. These marking algorithms remained the same throughout, doing exactly what they'd been asked to do. The problem was that we hadn't properly specified what that was. The algorithms were designed to hand out the same grades overall, as in previous years. So, no grade inflation here. The algorithms were also carefully calibrated to hand out similar grades in every school. And this is where things started to go wrong. This design meant it didn't matter how hard you worked, if no one in your school had previously gotten straight As, then the algorithm was unlikely to give you straight As. The problem wasn't that some super smart AI algorithm was giving out grades. It was that statistical levelling was used to prevent grade inflation. And that was going to hurt hardworking and deserving students who had done better than previous students from their score. Perhaps not intentionally devious, but harmful nonetheless.

I suspect this won't be the last time people take to the streets to protest algorithms making imperfect decisions based on flawed logic. The take home lesson from this sad story is that you don't need to fear some future super intelligence, but stupid intelligence, right now. We can see this problem elsewhere, facial recognition software that's racially biased, putting an innocent black person in jail, or simplistic robodebt algorithms that send out penalty notices to people who don't owe anything.

I get called by the media with depressing regularity to explain why some algorithm wasn't smart enough to make the decisions it was given the responsibility to make. Why does this keep happening? Computers are frustratingly little devices, they do precisely and only what you tell them to do. However ridiculous that might be. They have none of our understanding of the world. They don't understand, for example, that the tea I've just poured into this teacup, that tea is still inside the cup, even though you can no longer see it.

In computer vision, we call this object persistence. Objects persist, even when they move out of sight. It's a real challenge to get computers to understand such object persistence, something that a two-year-old quickly masters. Computers don't have what we might call, common sense, of, for example, how objects behave, or of how the world ticks. You've never seen this cup before. But unlike a computer, you know, that if I turn this cup over, gravity will

take over and the tea will pour out. Now computers also don't know anything about magic tricks. That magicians can make liquids disappear, or at least appear to.

So, what do we do about these unintelligently intelligent machines? Computers can be frighteningly smart in some ways, but dangerously dim in others. Like in any partnership, the key is to play to the different strengths that each partner brings to the team. Computers have many strengths. They can, for example, look at datasets beyond human comprehension, and they can work at speeds way faster than human biology. They also don't need to eat, sleep or go to the toilet, let alone need the same thinking time as a human in order to make a decision. And while algorithms don't mutate, machines that use them can learn when given the right guidance and conditions.

In 2016, Google's AlphaGo computer started to play the ancient Chinese game of Go better than any huma. Go is probably the oldest board game on the planet, the Chinese invented it 4000 years ago, it's way more complex than chess. To put this in perspective, there are even more possible ways a game of chess could go than there are atoms in the universe. There are 10 to the 70 atoms in the universe. That's one followed by 70 zeros. But there are around 10 to the 120 possible games of chess. That's one followed by 120 zeros. And there are many, many more possible games of Go, about 10 to the 174 in fact – that's one followed by 174 zeros, which is spectacularly large.

To play chess, a computer can test out every possible interesting move, something that's not possible with the game of Go. Instead, the program did it like you or I would go about getting good at Go, it played a lot of Go. Practice makes perfect for computers, just like for humans, but the computer can play way more games of Go than a human could in a lifetime of playing Go. In fact, it played way more games ago than 100 people would play in a lifetime of playing Go. The setup was simple. Two versions of AlphaGo played each other, millions of times, using dozens of computers in parallel. At the end of each of these millions of games, the AlphaGo program tweaked its weights, so in future, it will play more of the winning moves and less of the losing moves. Slowly but surely it got better and better. Better first than the humans who wrote the code, and eventually better than Lee Sedol, 18 times world champion. The most recent version of AlphaGo, AlphaGo Zero, played the version that beat Lee Sedol, back in 2016.

And won 100 nil. It's game over humanity. At least as far as playing the game of Go, goes. The Chinese called the program a Go God, as it plays such sublime Go, it plays moves that humans have never played. In the 4,000 years we've been playing the game. But it's a very brittle intelligence, change the rules ever so slightly, and AlphaGo would fail miserably. Change Go from a 19 by 19 board to a 20 by 20 board, and Lee said “oh we'll be fine”. But AlphaGo would have to start again and learn from scratch.

Computers have many other weaknesses, they lack many things that humans have naturally. They lack our common sense, our adaptability, our flexibility, our emotions, our empathy, perhaps even our souls. The success of AI means we can and should hand over many routine decisions to machines. However, the failures of AI remind us that we must leave high stakes decisions to humans, and that we need to be vigilant to prevent biases and unintended consequences that can creep unnoticed into the algorithms we create. Humans are tool users. Indeed, we are perhaps the supreme tool users. We weren't the fastest, or the strongest animals on the Savanna, but we used tools to become so. In the past, we used tools to amplify our muscles, but we now have tools to amplify our brains. Mutant algorithms or not, while they can amplify our brains they can never replace our humanity.

UNSW Centre for Ideas: Thanks for listening. For more information visit centreforideas.com, and don't forget to subscribe wherever you get your podcasts.