Google's Go victory shows AI thinking can be unpredictable, and that's a concern

March 18 2016, by Jonathan Tapson, Western Sydney University

Humans have been taking a beating from computers lately. The 4-1 defeat of Go grandmaster Lee Se-Dol by Google's AlphaGo artificial intelligence (AI) is only the latest in a string of pursuits in which technology has triumphed over humanity.

Self-driving cars are already less accident-prone than human drivers, the TV quiz show Jeopardy! is a lost cause, and in chess humans have fallen so woefully behind computers that a recent international tournament was won by a mobile phone.

There is a real sense that this month's human vs AI Go match marks a turning point. Go has long been held up as requiring levels of human intuition and pattern recognition that should be beyond the powers of number-crunching computers.

AlphaGo's win over one of the world's best players has reignited fears over the pervasive application of deep learning and AI in our future – fears famously expressed by Elon Musk as "our greatest existential threat ".

We should consider AI a threat for two reasons, but there are approaches we can take to minimise that threat.

The first problem is that AI is often trained using a combination of logic and heuristics, and reinforcement learning.

The logic and heuristics part has reasonably predictable results: we program the rules of the game or problem into the computer, as well as some human-expert guidelines, and then use the computer's number-crunching power to think further ahead than humans can.

This is how the early chess programs worked. While they played ugly chess, it was sufficient to win.

The problem of reinforcement learning

Reinforcement learning, on the other hand, is more opaque.

We have the computer perform the task – playing Go, for example – repetitively. It tweaks its strategy each time and learns the best moves from the outcomes of its play.

In order not to have to play humans exhaustively, this is done by playing the computer against itself. AlphaGo has played millions of games of Go – far more than any human ever has.

The problem is the AI will explore the entire space of possible moves and strategies in a way humans never would, and we have no insight into the methods it will derive from that exploration.

In the second game between Lee Se-Dol and AlphaGo, the AI made a move so surprising – "not a human move" in the words of a commentator – that Lee Se-Dol had to leave the room for 15 minutes to recover his composure.

This is a characteristic of machine learning. The machine is not constrained by human experience or expectations.

Until we see an AI do the utterly unexpected, we don't even realise that

we had a limited view of the possibilities. AIs move effortlessly beyond the limits of human imagination.

In real-world applications, the scope for AI surprises is much wider. A stock-trading AI, for example, will re-invent every single method known to us for maximising return on investment. It will find several that are not yet known to us.

Unfortunately, many methods for maximising stock returns – bid support, co-ordinated trading, and so on – are regarded as illegal and unethical price manipulation.

How do you prevent an AI from using such methods when you don't actually know what its methods are? Especially when the method it's using, while unethical, may be undiscovered by human traders – literally, unknown to humankind?

It's farcical to think that we will be able to predict or manage the worst-case behaviour of AIs when we can't actually imagine their probable behaviour.

The problem of ethics

This leads us to the second problem. Even quite simple AIs will need to behave ethically and morally, if only to keep their operators out of jail.

Unfortunately, ethics and morality are not reducible to heuristics or rules.

Consider Philippa Foot's famous trolley problem:

A trolley is running out of control down a track. In its path are five people who have been tied to the track by a mad philosopher.

Fortunately, you could flip a switch, which will lead the trolley down a different track to safety. Unfortunately, there is a single person tied to that track.

Should you flip the switch or do nothing?

What would you expect – or instruct – an AI to do?

In some psychological studies on the trolley problem, the humans who choose to flip the switch have been found to have underlying emotional deficits and score higher on measures of psychopathy – defined in this case as "a personality style characterised by low empathy, callous affect and thrill-seeking".

This suggests an important guideline for dealing with AIs. We need to understand and internalise that no matter how well they imitate or outperform humans, they will never have the intrinsic empathy or morality that causes human subjects to opt not to flip the switch.

Morality suggests to us that we may not take an innocent life, even when that path results in the greatest good for the greatest number.

Like sociopaths and psychopaths, AIs may be able to learn to imitate empathetic and ethical behaviour, but we should not expect there to be any moral force underpinning this behaviour, or that it will hold out against a purely utilitarian decision.

A really good rule for the use of AIs would be: "Would I put a sociopathic genius in charge of this process?"

There are two parts to this rule. We characterise AIs as sociopathic, in the sense of not having any genuine moral or empathetic constraints. And we characterise them as geniuses, and therefore capable of actions

that we cannot foresee.

Playing chess and Go? Maybe. Trading on the stock market? Well, one Swiss study found stock market traders display similarities to certified psychopaths, although that's not supposed to be a good thing.

But would you want an AI to look after your grandma, or to be in charge of a Predator drone?

There are good reasons why there is intense debate about the necessity for a human in the loop in autonomous warfare systems, but we should not be blinded to the potential for disaster in less obviously dangerous domains in which AIs are going to be deployed.

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation

Citation: Google's Go victory shows AI thinking can be unpredictable, and that's a concern (2016, March 18) retrieved 25 April 2024 from https://phys.org/news/2016-03-google-victory-ai-unpredictable.html