To protect us from the risks of advanced artificial intelligence, we need to act now

January 25 2019, by Paul Salmon, Peter Hancock And Tony Carden

Credit: AI-generated image (disclaimer)

Artificial intelligence can play chess, drive a car and diagnose medical issues. Examples include Google DeepMind's AlphaGo, Tesla's self-driving vehicles, and IBM's Watson.

This type of artificial intelligence is referred to as Artificial Narrow

Intelligence (ANI) – non-human systems that can perform a specific task. We encounter this type on a daily basis, and its use is growing rapidly.

But while many impressive capabilities have been demonstrated, we're also beginning to see problems. The worst case involved a self-driving test car that hit a pedestrian in March. The pedestrian died and the incident is still under investigation.

The next generation of AI

With the next generation of AI the stakes will almost certainly be much higher.

Artificial General Intelligence (AGI) will have advanced computational powers and human level intelligence. AGI systems will be able to learn, solve problems, adapt and self-improve. They will even do tasks beyond those they were designed for.

Importantly, their rate of improvement could be exponential as they become far more advanced than their human creators. The introduction of AGI could quickly bring about Artificial Super Intelligence (ASI).

While fully functioning AGI systems do not yet exist, it has been estimated that they will be with us anywhere between 2029 and the end of the century.

What appears almost certain is that they will arrive eventually. When they do, there is a great and natural concern that we won't be able to control them.

The risks associated with AGI

There is no doubt that AGI systems could transform humanity. Some of the more powerful applications include curing disease, solving complex global challenges such as climate change and food security, and initiating a worldwide technology boom.

But a failure to implement appropriate controls could lead to catastrophic consequences.

Despite what we see in Hollywood movies, existential threats are not likely to involve killer robots. The problem will not be one of malevolence, but rather one of intelligence, writes MIT professor Max Tegmark in his 2017 book Life 3.0: Being Human in the Age of Artificial Intelligence.

It is here that the science of human-machine systems – known as Human Factors and Ergonomics – will come to the fore. Risks will emerge from the fact that super-intelligent systems will identify more efficient ways of doing things, concoct their own strategies for achieving goals, and even develop goals of their own.

Imagine these examples:

- an AGI system tasked with preventing HIV decides to eradicate the problem by killing everybody who carries the disease, or one tasked with curing cancer decides to kill everybody who has any genetic predisposition for it
- an autonomous AGI military drone decides the only way to guarantee an enemy target is destroyed is to wipe out an entire community
- an environmentally protective AGI decides the only way to slow or reverse climate change is to remove technologies and humans that induce it.

These scenarios raise the spectre of disparate AGI systems battling each other, none of which take human concerns as their central mandate.

Various dystopian futures have been advanced, including those in which humans eventually become obsolete, with the subsequent extinction of the human race.

Others have forwarded less extreme but still significant disruption, including malicious use of AGI for terrorist and cyber-attacks, the removal of the need for human work, and mass surveillance, to name only a few.

So there is a need for human-centred investigations into the safest ways to design and manage AGI to minimise risks and maximise benefits.

How to control AGI

Controlling AGI is not as straightforward as simply applying the same kinds of controls that tend to keep humans in check.

Many controls on human behaviour rely on our consciousness, our emotions, and the application of our moral values. AGIs won't need any of these attributes to cause us harm. Current forms of control are not enough.

Arguably, there are three sets of controls that require development and testing immediately:

1. the controls required to ensure AGI system designers and developers create safe AGI systems
2. the controls that need to be built into the AGIs themselves, such as "common sense", morals, operating procedures, decision-rules, and so on

3. the controls that need to be added to the broader systems in which AGI will operate, such as regulation, codes of practice, standard operating procedures, monitoring systems, and infrastructure.

Human Factors and Ergonomics offers methods that can be used to identify, design and test such controls well before AGI systems arrive.

For example, it's possible to model the controls that exist in a particular system, to model the likely behaviour of AGI systems within this control structure, and identify safety risks.

This will allow us to identify where new controls are required, design them, and then remodel to see if the risks are removed as a result.

In addition, our models of cognition and decision making can be used to ensure AGIs behave appropriately and have humanistic values.

Act now, not later

This kind of research is in progress, but there is not nearly enough of it and not enough disciplines are involved.

Even the high-profile tech entrepreneur Elon Musk has warned of the "existential crisis" humanity faces from advanced AI and has spoken about the need to regulate AI before it's too late.

The next decade or so represents a critical period. There is an opportunity to create safe and efficient AGI systems that can have far reaching benefits to society and humanity.

At the same time, a business-as-usual approach in which we play catch-up with rapid technological advances could contribute to the extinction

of the human race. The ball is in our court, but it won't be for much longer.

Provided by The Conversation