Careful how you treat today's AI: It might take revenge in the future

March 13 2019, by Nicholas Agar

Credit: AI-generated image (disclaimer)

Artificial intelligence (AI) systems are becoming more like us. You can ask Google Home to switch off your bedroom lights, much as you might ask your human partner.

When you text inquiries to Amazon online, it's sometimes unclear

whether you're being answered by a human or the company's chatbot technology.

There's clearly a market for machines with human psychological abilities. But we should spare a thought for what we might inadvertently create.

What if we make AI so good at being human that our treatment of it can cause it to suffer? It might feel entitled to take revenge on us.

Machines that 'feel'

With human psychological abilities may come sentience. Philosophers understand sentience as the capacity to suffer and to feel pleasure.

And sentient beings can be harmed. It's an issue raised by the Australian philosopher Peter Singer in his 1975 book Animal Liberation, which asked how we should treat non-human animals. He wrote: "If a being suffers, there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that its suffering be counted equally with the like suffering – insofar as rough comparisons can be made – of any other being."

Singer has devoted a career to speaking up for animals, which are sentient beings incapable of speaking up for themselves.

Speaking up for AI

Researchers in AI are seeking to make an AGI or artificial general intelligence – a machine capable of any intellectual task performed by a human being. AI can already learn, but AGI will be able to perform tasks

beyond that for which it is programmed.

The experts disagree on how far off an AGI is. The US tech inventor Ray Kurzweil expects an AGI soon, maybe 2029. Others think we might have to wait for a century.

But if we are interested in treating sentient beings right, we may not have to wait until the arrival of an AGI.

One of Singer's points is that many sentient beings fall far short of human intelligence. By that argument, AI doesn't have to be as intelligent as a human for it to be sentient.

The problem is there is no straightforward test for sentience.

Sending a human crewed mission to Mars is very challenging, but at least we'll know when we've done it.

Making a machine with feelings is challenging in a more philosophically perplexing way. Because we lack clear criteria for machine sentience, we can't be sure when we've done it.

Look to science fiction

The ambiguity of machine sentience is a feature of several science fiction presentations of AI.

For example, Niska is a humanoid robot, a synth, serving as a sex worker in the TV series Humans. We are told that, unlike most synths, she is sentient.

When Niska is questioned about why she killed a client she explains: "He wanted to be rough."

The human lawyer Laura Hawkins responds: "But, is that wrong if he didn't think you could feel? … Isn't it better he exercises his fantasies with you in a brothel rather than take them out on someone who can actually feel?"

From a human perspective, one could think sexual assault directed against a non-sentient machine is a victimless crime.

But what about a sex robot that has acquired sentience? Niska goes on to explain that she was scared by the client's behaviour towards her. "And I'm sorry I can't cry or … bleed or wring my hands so you know that. But I'm telling you, I was."

Humans is not the only science fiction story to warn of revenge attacks from machines designed to be exploited by humans for pleasure and pain.

In the TV remake of Westworld, humans enter a theme park and kill android hosts with the abandon of Xbox massacres, confident their victims have no hard feelings because they can't have any feelings.

But here again, some hosts have secretly acquired sentience and get payback on their human tormentors.

We're only human

Is it only science fiction? Are sentient machines a long way off? Perhaps. Perhaps not.

But bad habits can take a while to unlearn. We – or rather animals – are still suffering the philosophical hangover of the 17th century French thinker Rene Descartes' terrible idea that animals are mindless automata – lacking in sentience.

If we are going to make machines with human psychological capacities, we should prepare for the possibility that they may become sentient. How then will they react to our behaviour towards them?

Perhaps our behaviour towards non-sentient AI today should be driven by how we would expect people to behave towards any future sentient AI that can feel, that can suffer. How we would expect that future sentient machine to react towards us?

This may be the big difference between machines and the animals that Singer defends. Animals cannot take revenge. But sentient machines just might.

Provided by The Conversation

Citation: Careful how you treat today's AI: It might take revenge in the future (2019, March 13) retrieved 20 April 2024 from https://phys.org/news/2019-03-today-ai-revenge-future.html