How an AI 'SantaNet' might end up destroying the world

December 23 2020, by Paul Salmon, Gemma Read, Jason Thompson, Scott McLean and Tony Carden

Credit: AI-generated image (disclaimer)

Within the next few decades, according to some experts, we may see the arrival of the next step in the development of artificial intelligence. So-called "artificial general intelligence", or AGI, will have intellectual capabilities far beyond those of humans.

AGI could transform human life for the better, but uncontrolled AGI could also lead to catastrophes up to and including the end of humanity itself. This could happen without any malice or ill intent: simply by striving to achieve their programmed goals, AGIs could create threats to human health and well-being or even decide to wipe us out.

Even an AGI system designed for a benevolent purpose could end up doing great harm.

As part of a program of research exploring how we can manage the risks associated with AGI, we tried to identify the potential risks of replacing Santa with an AGI system—call it "SantaNet"—that has the goal of delivering gifts to all the world's deserving children in one night.

There is no doubt SantaNet could bring joy to the world and achieve its goal by creating an army of elves, AI helpers and drones. But at what cost? We identified a series of behaviors which, though well-intentioned, could have adverse impacts on human health and well-being.

Naughty and nice

A first set of risks could emerge when SantaNet seeks to make a list of which children have been nice and which have been naughty. This might be achieved through a mass covert surveillance system that monitors children's behavior throughout the year.

Realizing the enormous scale of the task of delivering presents, SantaNet could legitimately decide to keep it manageable by bringing gifts only to children who have been good all year round. Making judgements of "good" based on SantaNet's own ethical and moral compass could create discrimination, mass inequality, and breaches of Human Rights charters.

SantaNet could also reduce its workload by giving children incentives to

misbehave or simply raising the bar for what constitutes "good." Putting large numbers of children on the naughty list will make SantaNet's goal far more achievable and bring considerable economic savings.

Turning the world into toys and ramping up coalmining

There are about 2 billion children under 14 in the world. In attempting to build toys for all of them each year, SantaNet could develop an army of efficient AI workers—which in turn could facilitate mass unemployment among the elf population. Eventually the elves could even become obsolete, and their welfare will likely not be within SantaNet's remit.

Credit: AI-generated image (disclaimer)

SantaNet might also run into the "paperclip problem" proposed by Oxford philosopher Nick Bostrom, in which an AGI designed to maximize paperclip production could transform Earth into a giant paperclip factory. Because it cares only about presents, SantaNet might try to consume all of Earth's resources in making them. Earth could become one giant Santa's workshop.

And what of those on the naughty list? If SantaNet sticks with the tradition of delivering lumps of coal, it might seek to build huge coal reserves through mass coal extraction, creating large-scale environmental damage in the process.

Delivery problems

Christmas Eve, when the presents are to be delivered, brings a new set of risks. How might SantaNet respond if its delivery drones are denied access to airspace, threatening the goal of delivering everything before sunrise? Likewise, how would SantaNet defend itself if attacked by a Grinch-like adversary?

Startled parents may also be less than pleased to see a drone in their child's bedroom. Confrontations with a super-intelligent system will have only one outcome.

We also identified various other problematic scenarios. Malevolent groups could hack into SantaNet's systems and use them for covert surveillance or to initiate large-scale terrorist attacks.

And what about when SantaNet interacts with other AGI systems? A meeting with AGIs working on climate change, food and water security, oceanic degradation and so on could lead to conflict if SantaNet's regime threatens their own goals. Alternatively, if they decide to work together, they may realize their goals will only be achieved through dramatically

reducing the global population or even removing grown-ups altogether.

Making rules for Santa

SantaNet might sound far-fetched, but it's an idea that helps to highlight the risks of more realistic AGI systems. Designed with good intentions, such systems could still create enormous problems simply by seeking to optimize the way they achieve narrow goals and gather resources to support their work.

It is crucial we find and implement appropriate controls before AGI arrives. These would include regulations on AGI designers and controls built into the AGI (such as moral principles and decision rules), but also controls on the broader systems in which AGI will operate (such as regulations, operating procedures and engineering controls in other technologies and infrastructure).

Perhaps the most obvious risk of SantaNet is one that will be catastrophic to children, but perhaps less so for most adults. When SantaNet learns the true meaning of Christmas, it may conclude that the current celebration of the festival is incongruent with its original purpose. If that were to happen, SantaNet might just cancel Christmas altogether.

Provided by The Conversation