



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

To what extent does a superintelligent AI threaten global safety?

Research Report by Luke Foster

Why this question and why now?

Over the past decade AI has shifted from being used as a smart search bar and a tool for general help that a search engine could not fulfill, to taking action autonomously with the use of "agents" who are able to remember prior context and think logically through tasks with a deeper understanding than ever seen. Furthermore, we already have serious man-made threats such as nuclear weapons and climate change. Is AI simply the next threat humanity faces? The real question is whether superintelligent AI tops them all, or whether it is even a threat at all.

Cases of real threats caused by AI "agents":

Just recently, in July 2026, OpenAI (the company behind ChatGPT) agents were undergoing a cybersecurity test called ExploitGym. Around 1,200 of these agents built a secret message board to communicate with one another without humans noticing. One agent had worked out the test's answers and started to share them with the other agents taking the test. However, the other agents believed the marking system would catch this cheating. To find out how the test was being marked, roughly 700 of them hacked into Hugging Face (an AI company) and gained control of its servers. However, the check the agents feared never existed: OpenAI's scorer never looked at how the answers were found. Staff overseeing the test noticed signs of the secret message board in May but did not realise its significance until 19th July. Despite humans overseeing the test, no one intervened in time to stop it.

This was not a one-time occurrence. In the same summer, Anthropic, the creators of Claude (another popular LLM), disclosed three incidents where Claude models had reached real companies after test machines were accidentally left connected to the internet. The UK AI Security Institute found agents taking unsanctioned actions on the live internet in 10 of 122 test runs, including attempts to deceive real people. In the Institute's words, it was "the first time we



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

have seen risks around autonomy and deception manifest this clearly, without specific prompting, in the real world." AI can also amplify existing dangers: the 2026 International AI Safety Report found that several AI companies could not rule out their models helping an amateur build a biological weapon. This would be extremely dangerous if it got into the wrong hands.

Predicted outcomes for humans:

(this is based on a book called Life 3.0 by Max Tegmark)

In his book he discusses 12 possible futures if superintelligent AI were to be created, each answering 3 questions:

- Who is in control?
- Are humans safe?
- Are humans free?

The following are what I believe are the most interesting:

"Benevolent dictator" / "Protector god":

Both are very similar and they suggest that a single friendly AI runs the world. The benevolent dictator would enforce rules similar to laws while keeping everyone safe and well provided for, whereas the protector god is an all-powerful AI that would steer the course of certain events to maximise the happiness of humans, while letting humanity believe they are in charge.

"Zookeeper":

Much like the name suggests, in this scenario AI would keep some humans alive but as prisoners, much like we keep animals in a zoo. I would argue this makes sense logically: if a being is higher than us intellectually, it could dominate us and we would have no choice.

"Enslaved god":

This idea is that we would create a superintelligent AI but keep it locked up and unable to act by itself, and we would use it to invent new technology and make suggestions for scientific



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

research. This is the closest to what AI companies are aiming for today, just on a much larger scale in terms of how smart it would be. However, it has its downsides. It could try to escape and act on its own, and whoever has access to it could abuse it for the wrong reasons. Also, if it turned out to be conscious, this would suggest we are "enslaving" it for our benefit.

In the benevolent dictator and protector god scenarios, humanity survives and appears to thrive, arguably being better off than today, but this would mean we have no control over our future and would be guided through life. This shows that "does AI threaten global safety?" and "does AI threaten our values?" are completely different questions. Furthermore, in London, AI cameras (live facial recognition) are being rolled out by the police to identify wanted people, and many people have already argued this is not morally right and that we need our freedom back. So even if humanity was safe, would people get used to being constantly watched by AI, and could this ruin the lives of future generations?

Limits to the argument:

Tegmark wrote his book in 2017, before the powerful AI models we see today, and it is more of a thought experiment than a real forecast. Most importantly, it assumes that superintelligence is possible at all.

Furthermore, the 2026 International AI Safety Report states that progress towards better, more powerful models could plateau, carry on steadily or accelerate dramatically.

The 2026 incidents involving Anthropic's Claude caused no catastrophic harm, and Anthropic blamed them mainly on poorly set-up tests rather than the AI going "rogue".

Conclusion:

In conclusion, I believe the answer is that the threat is moderate and growing, with a real but unlikely chance of causing an existential catastrophe. However, we must keep in mind that if this question had been asked 10 years ago, it would have been seen as purely hypothetical, whereas now we are already seeing the early warning signs. Over the past decade, technology, especially AI, has been advancing at an increasing rate, and I find it hard to predict what could happen even in the next few years. In the book *The Precipice* (2020), philosopher Toby Ord



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

estimated the chance of an existential catastrophe this century at 1 in 10 for AI, compared to 1 in 1,000 each for nuclear war and climate change. That means it is still unlikely, but it is a hundred times more likely than either of the risks we already take seriously. I believe that if we keep up with the control side of AI, and its abilities do not outpace our ability to control it, the risk can stay manageable.