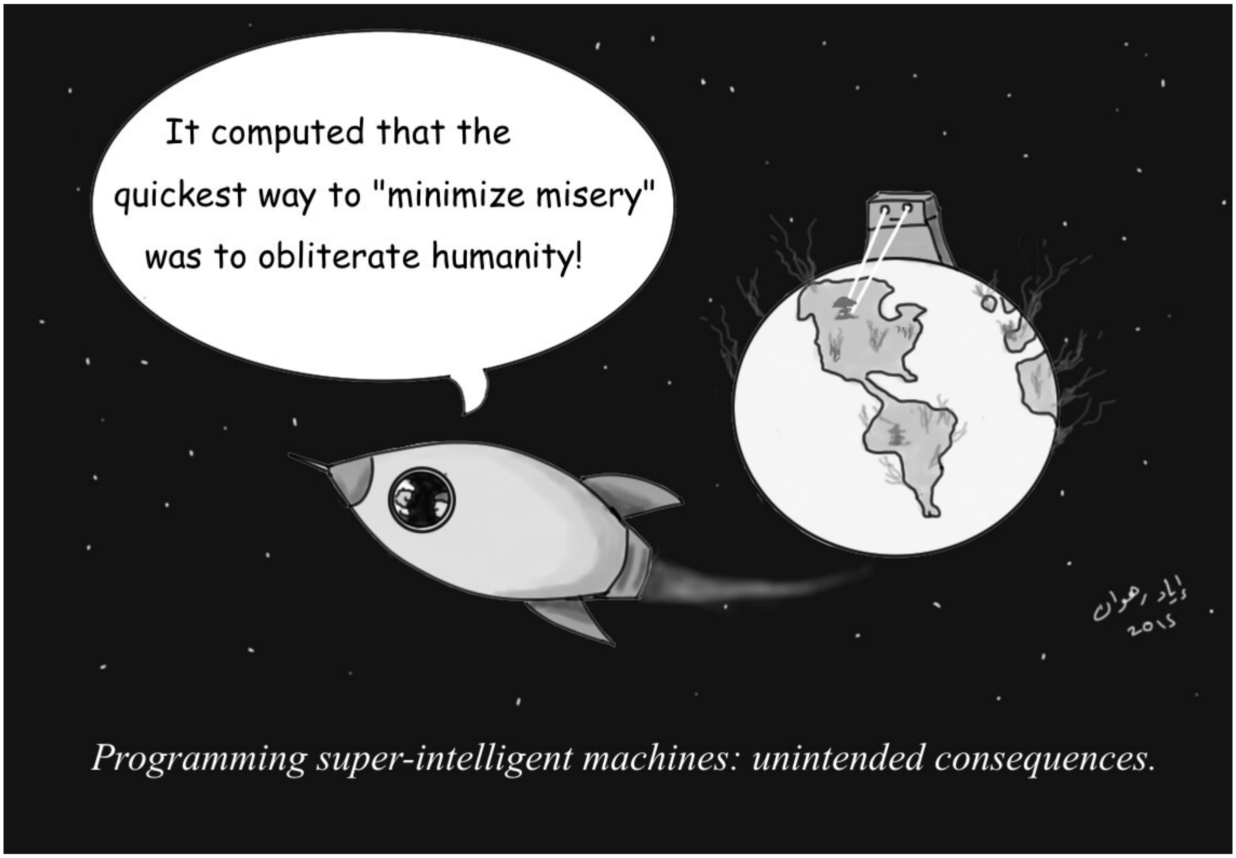
TechXplore

We wouldn't be able to control superintelligent machines

January 11 2021


It computed that the quickest way to "minimize misery" was to obliterate humanity!
Programming super-intelligent machines: unintended consequences.

Endowing AI with noble goals may not prevent unintended consequences. Credit: Iyad Rahwan

We are fascinated by machines that can control cars, compose symphonies, or defeat people at chess, Go, or Jeopardy! While more progress is being made all the time in Artificial Intelligence (AI), some scientists and philosophers warn of the dangers of an uncontrollable superintelligent AI. Using theoretical calculations, an international team of researchers, including scientists from the Center for Humans and Machines at the Max Planck Institute for Human Development, shows that it would not be possible to control a superintelligent AI.

Suppose someone were to program an AI system with intelligence superior to that of humans, so it could learn independently. Connected to the Internet, the AI may have access to all the data of humanity. It could replace all existing programs and take control all machines online worldwide. Would this produce a utopia or a dystopia? Would the AI cure cancer, bring about world peace, and prevent a climate disaster? Or would it destroy humanity and take over the Earth?

Computer scientists and philosophers have asked themselves whether we would even be able to control a superintelligent AI at all, to ensure it would not pose a threat to humanity. An international team of computer scientists used theoretical calculations to show that it would be fundamentally impossible to control a super-intelligent AI

"A super-intelligent machine that controls the world sounds like science fiction. But there are already machines that perform certain important tasks independently without programmers fully understanding how they learned it. The question therefore arises whether this could at some point become uncontrollable and dangerous for humanity," says study co-author Manuel Cebrian, Leader of the Digital Mobilization Group at the Center for Humans and Machines, Max Planck Institute for Human Development.

Scientists have explored two different ideas for how a superintelligent AI could be controlled. On one hand, the capabilities of superintelligent AI could be specifically limited, for example, by walling it off from the

Internet and all other technical devices so it could have no contact with the outside world—yet this would render the superintelligent AI significantly less powerful, less able to answer humanities quests. Lacking that option, the AI could be motivated from the outset to pursue only goals that are in the best interests of humanity, for example by programming ethical principles into it. However, the researchers also show that these and other contemporary and historical ideas for controlling super-intelligent AI have their limits.

In their study, the team conceived a theoretical containment algorithm that ensures a superintelligent AI cannot harm people under any circumstances, by simulating the behavior of the AI first and halting it if considered harmful. But careful analysis shows that in our current paradigm of computing, such algorithm cannot be built.

"If you break the problem down to basic rules from theoretical computer science, it turns out that an algorithm that would command an AI not to destroy the world could inadvertently halt its own operations. If this happened, you would not know whether the containment algorithm is still analyzing the threat, or whether it has stopped to contain the harmful AI. In effect, this makes the containment algorithm unusable," says Iyad Rahwan, Director of the Center for Humans and Machines.

Based on these calculations the containment problem is incomputable, i.e. no single algorithm can find a solution for determining whether an AI would produce harm to the world. Furthermore, the researchers demonstrate that we may not even know when superintelligent machines have arrived, because deciding whether a machine exhibits intelligence superior to humans is in the same realm as the containment problem.

The study, "Superintelligence cannot be contained: Lessons from Computability Theory," was published in the *Journal of Artificial Intelligence Research*.

More information: Manuel Alfonseca et al. Superintelligence Cannot be Contained: Lessons from Computability Theory, *Journal of Artificial Intelligence Research* (2021). DOI: 10.1613/jair.1.12202

Provided by Max Planck Society

Citation: We wouldn't be able to control superintelligent machines (2021, January 11) retrieved 19 April 2024 from https://techxplore.com/news/2021-01-wouldnt-superintelligent-machines.html