Technological risks are not the end of the world

Jack Stilgoe
Department of Science and Technology Studies, University College London, London, UK.
j.stilgoe@ucl.ac.uk

There's a scene in the movie *Oppenheimer* in which the protagonist is trying to explain to General Groves, his military overseer, the hazards of their endeavour. Groves asks Oppenheimer, "Are you saying there's a chance that when we push that button, we destroy the world?" The physicist says, "The chances are near zero." When Groves, understandably alarmed, asks for clarification, Oppenheimer responds, "What do you want from theory alone?"

It's a compelling set-up. An invention of unprecedented power; the product of scientific genius in the context of a desperate, money-no-object race. The Promethean creator is among a select few with expert insight into the technology's world-ending potential and the moral clarity required to weigh risks and responsibilities.

Artificial Intelligence (AI) is currently having an Oppenheimer moment. As the technology's leaders have attracted money and attention, they have also sought to reassure us they are engaging responsibly with the technology's hazards. A year ago, many of them signed a one-sentence open letter that read, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

For a recent BBC radio documentary, I had the opportunity to speak to proponents and critics of this idea. To put my cards on the table, I was, and I remain, unconvinced that there is a risk of AI wiping out humanity. But as a sociologist I wanted to know why the idea seems to be spreading so quickly and how it is affecting the debate about AI. I began by visiting Geoffrey Hinton, so-called 'godfather of AI' and the first signatory of that open letter.

Sociologists have found that, when it comes to science and innovation, distance normally lends enchantment. Those on the fringes of innovation may see a technology as magical, but the people who see it up close understand the messy reality. With AI, even the people nearest the technology seem in thrall to it.. Hinton explained to me his surprise at the giant leaps made by the large language models that his research has helped enable: "it's very exciting. It's very nice to see all this work coming to fruition. But it's also scary." He, like other AI researchers, cannot fully explain how the machines do what they do and is troubled by the implications. Last year, Hinton stepped down from his role at Google and chose to speak out about what he saw as the existential dangers of AI.

The idea that AI models pose an existential threat will strike many people as the stuff of science fiction. But it has been a part of AI research since the beginning. In a lecture in 1951, Alan Turing said,

> "*it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines*

dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control."

The people who worry about losing control of AI are often unclear about the means of our possible extermination, but the general idea is that, once outsmarted, humans would be seen by a superintelligent AI as surplus to requirements and, through accident or design, disposed of.

Shazeda Ahmed and colleagues have studied the recent emergence of a so-called 'AI Safety' community who prioritise 'x-risk'. Much of their financial support has come from Silicon Valley philanthropists who have calculated that the elimination of not just all current humans but the possibility of any future humans overshadows other concerns. The money has created an increasingly powerful network of lobbyists, redirected the attention of philosophers and computer scientists and fuelled already-enthusiastic online forums. Think tanks have sought to strengthen claims by asking computer scientists to forecast the arrival of an imagined superhuman AI. At research conferences, some prizes for papers on existential risk have dwarfed those for other research questions. One competition funded by the Center for AI Safety had a total prizewinners' pot of $500,000.

This is not just about the hobbies of billionaires or the spreading of online memes. It also about the shaping of policy priorities. AI's existential risk has been given credence by governments around the world. In the run-up to last year's AI Safety Summit in the UK, Rishi Sunak chose to highlight "the risk that humanity could lose control of AI completely".

Michael Wooldridge, a computer scientist from Oxford University told me that the trouble with this view of risk is that it "sucks all the oxygen out of the room", making discussion of other, more pressing concerns impossible. Another of my interviewees, social scientist Kate Crawford, has tried to combat this tendency, highlighting issues such as misinformation, copyright breaches, and unsustainable uses of energy and rare minerals in the development and use of AI.

AI companies would rather we didn't pay attention to the clear downsides of their technology. The paradox is that a focus on the end of the world is oddly convenient. A hypothetical apocalypse is, as Divya Siddarth from the Collective Intelligence Project told me, "a clean risk". It is all-or-nothing, absolving innovators from having to engage with the messy inequities that are produced by their technologies.

An existential risk scenario presumes that the relevant political struggle is between humans and their robot creations, not between humans and other humans. Existential risk sidesteps questions of who the winners and losers will be. We are being asked to trust that the experts currently in charge of AI will not just identify the risks that matter, but save us from them.

Political scientist Philip Tetlock and his team asked a group of 'superforecasters' – people with a knack for prediction – for their assessment of the risk of human extinction from AI before the end of the century (some x-risk people call this "p(doom)"). Their calculation was 0.38%. AI experts averaged 3% and self-appointed existential risk experts were at 4.75%. I don't think the precise percentages are important (the philosopher Alfred North Whitehead would have called it the fallacy of misplaced concreteness), but the differences of opinion

are interesting. One way of interpreting the gap is that the AI and x-risk experts, like Oppenheimer, might know something that they are struggling to communicate to the rest of us. The other interpretation is that, faced with extreme uncertainty and intense societal interest, the experts are imagining risks in ways that fit their expertise. Having spoken to some of the leading voices in this debate, I take the latter view.

We have been here before. Other overhyped new technologies have been accompanied by parables of doom. In 2000, Bill Joy warned in a Wired cover article that "the future doesn't need us" and that nanotechnology would inevitably lead to "knowledge-enabled mass destruction". John Seely Brown and Paul Duguid's criticism at the time was that "Joy can see the juggernaut clearly. What he can't see – which is precisely what makes his vision so scary – are any controls." Existential risks tell us more about their purveyors' lack of faith in human institutions than about the actual hazards we face. As Divya Siddarth explained to me, a belief that "the technology is smart, people are terrible and no one's going to save us" will tend towards catastrophizing.

Geoffrey Hinton is hopeful that, at a time of political polarisation, existential risks offer a way of building consensus. He told me, "It's something we should be able to collaborate on because we all have the same payoff". But it is a counsel of despair. Real policy collaboration is impossible if a technology and its problems are imagined in ways that disempower policymakers. The risk is that, if we build regulations around a future fantasy, we lose sight of where the real power lies and give up on the hard work of governing the technology in front of us.

At the end of the movie, *Oppenheimer* realises in a conversation with Albert Einstein that the real danger comes not from nuclear weapons going wrong, but from the technology working exactly as intended. 'Theory alone' has not prepared Oppenheimer for the real risks of nuclear proliferation. With AI, there are some signs that the risk debate is becoming more grounded. Last year's White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence carefully avoided any mention of existential risks. Policymakers now need to push back against AI hype. In the coming years, AI will pose countless challenges for regulators. But thankfully they won't be the end of the world.