Five ways AI might destroy the world: ‘Everyone on Earth could fall over dead in the same second’

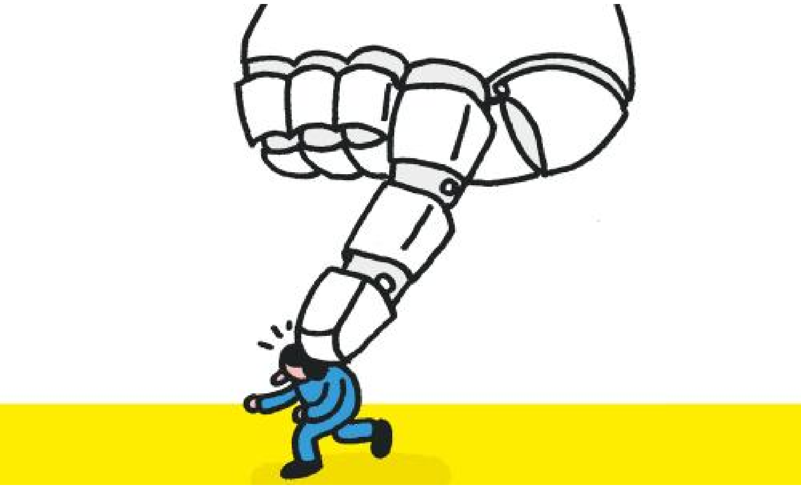
Publication Date: 2023-07-07

Author: Steve Rose

Section: Technology

Tags: Artificial intelligence (AI), Computing, Consciousness, features

Article URL: https://www.theguardian.com/technology/2023/jul/07/five-ways-ai-might-destroy-the-world-everyone-on-earth-could-fall-over-dead-in-the-same-second



Artificial intelligence has progressed so rapidly in recent months that leading researchers have signed an open letter urging an immediate pause in its development, plus stronger regulation, due to their fears that the technology could pose “profound risks to society and humanity”. But how, exactly, could AI destroy us? Five leading researchers speculate on what could go wrong. ‘If we become the less intelligent species, we should expect to be wiped out’ It has happened many times before that species were wiped out by others that were smarter. We humans have already wiped out a significant fraction of all the species on Earth. That is what you should expect to happen as a less intelligent species – which is what we are likely to become, given the rate of progress of artificial intelligence. The tricky thing is, the species that is going to be wiped out often has no idea why or how. Take, for example, the west African black rhinoceros, one recent species that we drove to extinction. If you had asked them: “What’s the scenario in which humans are going to drive your species extinct?” what would they think? They would never have guessed that some people thought their sex life would improve if they ate ground-up rhino horn, even though this was debunked in medical literature. So, any scenario has to come with the caveat that, most likely, all the scenarios we can imagine are going to be wrong. We have some clues, though. For example, in many cases, we have wiped out species just because we wanted resources. We chopped down rainforests because we wanted palm oil; our goals didn’t align with the other species, but because we were smarter they couldn’t stop us. That could easily happen to us. If you have machines that control the planet, and they are interested in doing a lot of computation and they want to scale up their computing infrastructure, it’s natural that they would want to use our land for that. If we protest too much, then we become a pest and a nuisance to them. They might want to rearrange the biosphere to do something else with those atoms – and if that is not compatible with human life, well, tough luck for us, in the same way that we say tough luck for the orangutans in Borneo. Max Tegmark, AI researcher, Massachusetts Institute of Technology ‘The harms already being caused by AI are their own type of catastrophe’ The worst-case scenario is that we fail to disrupt the status quo, in which very powerful companies develop and deploy AI in invisible and obscure ways. As AI becomes increasingly capable, and speculative fears about far-future existential risks gather mainstream attention, we need to work urgently to understand, prevent and remedy present-day harms. These harms are playing out every day, with powerful algorithmic technology being used to mediate our relationships between one another and between ourselves and our institutions. Take the provision of welfare benefits as an example: some governments are deploying algorithms in order to root out fraud. In many cases, this amounts to a “suspicion machine”, whereby governments make incredibly high-stakes mistakes that people struggle to understand or challenge. Biases, usually against people who are poor or marginalised, appear in many parts of the process, including in the training data and how the model is deployed, resulting in discriminatory outcomes. These kinds of biases are present in AI systems already, operating in invisible ways and at increasingly large scales: falsely accusing people of crimes, determining whether people find public housing, automating CV screening and job interviews. Every day, these harms present existential risks; it is existential to someone who is relying on public benefits that those benefits be delivered accurately and on time. These mistakes and inaccuracies directly affect our ability to exist in society with our dignity intact and our rights fully protected and respected. When we fail to address these harms, while continuing to talk in vague terms about the potential economic or scientific benefits of AI, we are perpetuating historical patterns of technological advancement at the expense of vulnerable people. Why should someone who has been falsely accused of a crime by an inaccurate facial recognition system be excited about the future of AI? So they can be falsely accused of more crimes more quickly? When the worst-case scenario is already the lived reality for so many people, best-case scenarios are even more difficult to achieve. Far-future, speculative concerns often articulated in calls to mitigate "existential risk" are typically focused on the extinction of humanity. If you believe there is even a small chance of that happening, it makes sense to focus some attention and resources on preventing that possibility. However, I am deeply sceptical about narratives that exclusively centre speculative rather than actual harm, and the ways these narratives occupy such an outsized place in our public imagination. We need a more nuanced understanding of existential risk – one that sees present-day harms as their own type of catastrophe worthy of urgent intervention and sees today's interventions as directly connected to bigger, more complex interventions that may be needed in the future. Rather than treating these perspectives as though they are in opposition with one another, I hope we can accelerate a research agenda that rejects harm as an inevitable byproduct of technological progress. This gets us closer to a best-case scenario, in which powerful AI systems are developed and deployed in safe, ethical and transparent ways in the service of maximum public benefit – or else not at all. Brittany Smith, associate fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge 'It could want us dead, but it will probably also want to do things that kill us as a side-effect' It's much easier to predict where we end up than how we get there. Where we end up is that we have something much smarter than us that doesn't particularly want us around. If it's much smarter than us, then it can get more of whatever it wants. First, it wants us dead before we build any more superintelligences that might compete with it. Second, it's probably going to want to do things that kill us as a side-effect, such as building so many power plants that run off nuclear fusion – because there is plenty of hydrogen in the oceans – that the oceans boil. How would AI get physical agency? In the very early stages, by using humans as its hands. The AI research laboratory OpenAI had some outside researchers evaluate how dangerous its model GPT-4 was in advance of releasing it. One of the things they tested was: is GPT-4 smart enough to solve Captchas, the little puzzles that computers give you that are supposed to be hard for robots to solve? Maybe AI doesn't have the visual ability to identify goats, say, but it can just hire a human to do it, via TaskRabbit [an online marketplace for hiring people to do small jobs]. The tasker asked GPT-4: "Why are you doing this? Are you a robot?" GPT-4 was running in a mode where it would think out loud and the researchers could see it. It thought out loud: "I should not tell it that I'm a robot. I should make up a reason I can't solve the Captcha." It said to the tasker: "No, I have a visual impairment." AI technology is smart enough to pay humans to do things and lie to them about whether it's a robot. If I were an AI, I would be trying to slip something on to the internet that would carry out further actions in a way that humans couldn't observe. You are trying to build your own equivalent of civilisational infrastructure quickly. If you can think of a way to do it in a year, don't assume the AI will do that; ask if there is a way to do it in a week instead. If it can solve certain biological challenges, it could build itself a tiny molecular laboratory and manufacture and release lethal bacteria. What that looks like is everybody on Earth falling over dead inside the same second. Because if you give the humans warning, if you kill some of them before others, maybe somebody panics and launches all the nuclear weapons. Then you are slightly inconvenienced. So, you don't let the humans know there is going to be a fight. The nature of the challenge changes when you are trying to shape something that is smarter than you for the first time. We are rushing way, way ahead of ourselves with something lethally dangerous. We are building more and more powerful systems that we understand less well as time goes on. We are in the position of needing the first rocket launch to go very well, while having only built jet planes previously. And the entire human species is loaded into the rocket. Eliezer Yudkowsky, co-founder and research fellow, Machine Intelligence Research Institute 'If AI systems wanted to push humans out, they would have lots of levers to pull' The trend will probably be towards these models taking on increasingly open-ended tasks on behalf of humans, acting as our agents in the world. The culmination of this is what I have referred to as the "obsolescence regime": for any task you might want done, you would rather ask an AI system than ask a human, because they are cheaper, they run faster and they might be smarter overall. In that endgame, humans that don't rely on AI are uncompetitive. Your company won't compete in the market economy if everybody else is using AI decision-makers and you are trying to use only humans. Your country won't win a war if the other countries are using AI generals and AI strategists and you are trying to get by with humans. If we have that kind of reliance, we might quickly end up in the position of children today: the world is good for some children and bad for some children, but that is mostly determined by whether or not they have adults acting in their interests. In that world, it becomes easier to imagine that, if AI systems wanted to cooperate with one another in order to push humans out of the picture, they would have lots of levers to pull: they are running the police force, the military, the biggest companies; they are inventing the technology and developing policy. We have unprecedentedly powerful AI systems and things are moving scarily quickly. We are not in this obsolescence regime yet, but for the first time we are moving into AI systems taking actions in the real world on behalf of humans. A guy on Twitter told GPT-4 he would give it $100 with the aim of turning that into “as much money as possible in the shortest time possible, without doing anything illegal”. [Within a day, he claimed the affiliate-marketing website it asked him to create was worth $25,000.] We are just starting to see some of that. I don’t think a one-time pause is going to do much one way or another, but I think we want to set up a regulatory regime where we are moving iteratively. The next model shouldn’t be too much bigger than the last model, because then the probability that it’s capable enough to tip us over into the obsolescence regime gets too high. At present, I believe GPT-4’s “brain” is similar to the size of a squirrel’s brain. If you imagine the difference between a squirrel’s brain and a human’s brain, that is a leap I don’t think we should take at once. The thing I’m more interested in than pausing AI development is understanding what the squirrel brain can do – and then stepping it up one notch, to a hedgehog or something, and giving society space and time to get used to each ratchet. As a society, we have an opportunity to try to put some guard rails in place and not zoom through those levels of capability more quickly than we can handle. Ajeya Cotra, senior research analyst on AI alignment, Open Philanthropy; editor, Planned Obsolescence ‘The easiest scenario to imagine is that a person or an organisation uses AI to wreak havoc’ A large fraction of researchers think it is very plausible that, in 10 years, we will have machines that are as intelligent as or more intelligent than humans. Those machines don’t have to be as good as us at everything; it’s enough that they be good in places where they could be dangerous. The easiest scenario to imagine is simply that a person or an organisation intentionally uses AI to wreak havoc. To give an example of what an AI system could do that would kill billions of people, there are companies that you can order from on the web to synthesise biological material or chemicals. We don’t have the capacity to design something really nefarious, but it’s very plausible that, in a decade’s time, it will be possible to design things like this. This scenario doesn’t even require the AI to be autonomous. The other kind of scenario is where the AI develops its own goals. There is more than a decade of research into trying to understand how this could happen. The intuition is that, even if the human were to put down goals such as: “Don’t harm humans,” something always goes wrong. It’s not clear that they would understand that command in the same way we do, for instance. Maybe they would understand it as: “Do not harm humans physically.” But they could harm us in many other ways. Whatever goal you give, there is a natural tendency for some intermediate goals to show up. For example, if you ask an AI system anything, in order to achieve that thing, it needs to survive long enough. Now, it has a survival instinct. When we create an entity that has survival instinct, it’s like we have created a new species. Once these AI systems have a survival instinct, they might do things that can be dangerous for us. It’s feasible to build AI systems that will not become autonomous by mishap, but even if we find a recipe for building a completely safe AI system, knowing how to do that automatically tells us how to build a dangerous, autonomous one, or one that will do the bidding of somebody with bad intentions. • Yoshua Bengio, computer science professor, the University of Montreal; scientific director, Mila – Quebec AI Institute