OpenAI leaders call for regulation to prevent AI destroying humanity

Publication Date: 2023-05-24

Author: Alex Hern

Section: Technology

Tags: Artificial intelligence (AI), Computing, ChatGPT, OpenAI, Consciousness, news

Article URL: https://www.theguardian.com/technology/2023/may/24/openai-leaders-call-regulation-prevent-ai-destroying-humanity

The leaders of the ChatGPT developer OpenAI have called for the regulation of “superintelligent” AIs, arguing that an equivalent to the International Atomic Energy Agency is needed to protect humanity from the risk of accidentally creating something with the power to destroy it. In a short note published to the company’s website, co-founders Greg Brockman and Ilya Sutskever and the chief executive, Sam Altman, call for an international regulator to begin working on how to “inspect systems, require audits, test for compliance with safety standards, [and] place restrictions on degrees of deployment and levels of security” in order to reduce the “existential risk” such systems could pose. “It’s conceivable that within the next 10 years, AI systems will exceed expert skill level in most domains, and carry out as much productive activity as one of today’s largest corporations,” they write. “In terms of both potential upsides and downsides, superintelligence will be more powerful than other technologies humanity has had to contend with in the past. We can have a dramatically more prosperous future; but we have to manage risk to get there. Given the possibility of existential risk, we can’t just be reactive.” In the shorter term, the trio call for “some degree of coordination” amongcompanies working on the cutting-edge of AI research, in order to ensure the development of ever-more powerful models integrates smoothly with society while prioritising safety. That coordination could come through a government-led project, for instance, or through a collective agreement to limit growth in AI capability. Researchers have been warning of the potential risks of superintelligence for decades, but as AI development has picked up pace those risks have become more concrete. The US-based Center for AI Safety (CAIS), which works to “reduce societal-scale risks from artificial intelligence”, describes eight categories of “catastrophic” and “existential” risk that AI development could pose. While some worry about a powerful AI completely destroying humanity, accidentally or on purpose, CAIS describes other more pernicious harms. A world where AI systems are voluntarily handed ever more labour could lead to humanity “losing the ability to self-govern and becoming completely dependent on machines”, described as “enfeeblement”; and a small group of people controlling powerful systems could “make AI a centralising force”, leading to “value lock-in”, an eternal caste system between ruled and rulers. OpenAI’s leaders say those risks mean “people around the world should democratically decide on the bounds and defaults for AI systems”, but admit that “we don’t yet know how to design such a mechanism”. However, they say continued development of powerful systems is worth the risk. “We believe it’s going to lead to a much better world than what we can imagine today (we are already seeing early examples of this in areas like education, creative work, and personal productivity),” they write. They warn it could also be dangerous to pause development. “Because the upsides are so tremendous, the cost to build it decreases each year, the number of actors building it is rapidly increasing, and it’s inherently part of the technological path we are on. Stopping it would require something like a global surveillance regime, and even that isn’t guaranteed to work. So we have to get it right.”