Robot takeover? Not quite. Here’s what AI doomsday would look like

Publication Date: 2023-06-03

Author: Kari Paul

Section: Technology

Tags: Technology, Artificial intelligence (AI), features

Article URL: https://www.theguardian.com/technology/2023/jun/03/ai-danger-doomsday-chatgpt-robots-fears

Alarm over artificial intelligence has reached a fever pitch in recent months. Just this week, more than 300 industry leaders published a letter warning AI could lead to human extinction and should be considered with the seriousness of “pandemics and nuclear war”. Terms like “AI doomsday” conjure up sci-fi imagery of a robot takeover, but what does such a scenario actually look like? The reality, experts say, could be more drawn out and less cinematic – not a nuclear bomb but a creeping deterioration of the foundational areas of society. “I don’t think the worry is of AI turning evil or AI having some kind of malevolent desire,” said Jessica Newman, director of University of California Berkeley’s Artificial Intelligence Security Initiative. “The danger is from something much more simple, which is that people may program AI to do harmful things, or we end up causing harm by integrating inherently inaccurate AI systems into more and more domains of society.” That’s not to say we shouldn’t be worried. Even if humanity-annihilating scenarios are unlikely, powerful AI has the capacity to destabilize civilizations in the form of escalating misinformation, manipulation of human users, and a huge transformation of the labor market as AI takes over jobs. Artificial intelligence technologies have been around for decades, but the speed with which language learning models like ChatGPT have entered the mainstream has intensified longstanding concerns. Meanwhile, tech companies have entered a kind of arms race, rushing to implement artificial intelligence into their products to compete with one another, creating a perfect storm, said Newman. “I am extremely worried about the path we are on,” she said. “We’re at an especially dangerous time for AI because the systems are at a place where they appear to be impressive, but are still shockingly inaccurate and have inherent vulnerabilities.” Experts interviewed by the Guardian say these are the areas they’re most concerned about. Disinformation speeds the erosion of truth In many ways, the so-called AI revolution has been under way for some time. Machine learning underpins the algorithms that shape our social media newsfeeds – technology that has been blamed for perpetuating gender bias, stoking division and fomenting political unrest. Experts warn that those unresolved issues will only intensify as artificial intelligence models take off. Worst-case scenarios could include an eroding of our shared understanding of truth and valid information, leading to more uprisings based on falsehoods – as played out in the 6 January attack on the US Capitol. Experts warn further turmoil and even wars could be sparked by the rise in mis- and disinformation. “It could be argued that the social media breakdown is our first encounter with really dumb AI – because the recommender systems are really just simple machine learning models,” said Peter Wang, CEO and co-founder of the data science platform Anaconda. “And we really utterly failed that encounter.” Wang added that those mistakes could be self-perpetuating, as language learning models are trained on misinformation that creates flawed data sets for future models. This could lead to a “model cannibalism” effect, where future models amplify and are forever biased by the output of past models. Misinformation – simple inaccuracies – and disinformation – false information maliciously spread with the intent to mislead – have both been amplified by artificial intelligence, experts say. Large language models like ChatGPT are prone to a phenomenon called “hallucinations”, in which fabricated or false information is repeated. A study from the journalism credibility watchdog NewsGuard identified dozens of “news” sites online written entirely by AI, many of which contained such inaccuracies. Such systems could be weaponized by bad actors to purposely spread misinformation at a large scale, said Gordon Crovitz and Steven Brill, co-CEOs of NewsGuard. This is particularly concerning in high-stakes news events, as we have already seen with intentional manipulation of information in the Russia-Ukraine war. “You have malign actors who can generate false narratives and then use the system as a force multiplier to disseminate that at scale,” Crovitz said. “There are people who say the dangers of AI are being overstated, but in the world of news information it is having a staggering impact.” Recent examples have ranged from the more benign, like the viral AI-generated image of the Pope wearing a “swagged-out jacket”, to fakes with potentially more dire consequences, like an AI-generated video of the Ukrainian president, Volodymyr Zelenskiy, announcing a surrender in April 2022. “Misinformation is the individual [AI] harm that has the most potential and highest risk in terms of larger-scale potential harms,” said Rebecca Finlay, of the Partnership on AI. “The question emerging is: how do we create an ecosystem where we are able to understand what is true? How do we authenticate what we see online?” ‘Like a friend, not a tool’: malicious use and manipulation of users While most experts say misinformation has been the most immediate and widespread concern, there is debate over the extent to which the technology could negatively influence its users’ thoughts or behavior. Those concerns are already playing out in tragic ways, after a man in Belgium died by suicide after a chatbot allegedly encouraged him to kill himself. Other alarming incidents have been reported – including a chatbot telling one user to leave his partner, and another reportedly telling users with eating disorders to lose weight. Chatbots are, by design, likely to engender more trust because they speak to their users in a conversational manner, said Newman. “Large language models are particularly capable of persuading or manipulating people to slightly change their beliefs or behaviors,” she said. “We need to look at the cognitive impact that has on a world that’s already so polarized and isolated, where loneliness and mental health are massive issues.” The fear, then, is not that AI chatbots will gain sentience and overtake their users, but that their programmed language can manipulate people into causing harms they may not have otherwise. This is particularly concerning with language systems that work on an advertising profit model, said Newman, as they seek to manipulate user behavior and keep them using the platform as long as possible. “There are a lot of cases where a user caused harm not because they wanted to, but because it was an unintentional consequence of the system failing to follow safety protocols,” she said. Newman added that the human-like nature of chatbots makes users particularly susceptible to manipulation. “If you’re talking to something that’s using first-person pronouns, and talking about its own feeling and background, even though it is not real, it still is more likely to elicit a kind of human response that makes people more susceptible to wanting to believe it,” she said. “It makes people want to trust it and treat it more like a friend than a tool.” The impending labor crisis: ‘There’s no framework for how to survive’ A longstanding concern is that digital automation will take huge numbers of human jobs. Research varies, with some studies concluding AI could replace the equivalent of 85m jobs worldwide by 2025 and more than 300m in the long term. The industries affected by AI are wide-ranging, from screenwriters to data scientists. AI was able to pass the bar exam with similar scores to actual lawyers and answer health questions better than actual doctors. Experts are sounding the alarm about mass job loss and accompanying political instability that could take place with the unabated rise of artificial intelligence. Wang warns that mass layoffs lie in the very near future, with a “number of jobs at risk” and little plan for how to handle the fallout. “There’s no framework in America about how to survive when you don’t have a job,” he said. “This will lead to a lot of disruption and a lot of political unrest. For me, that is the most concrete and realistic unintended consequence that emerges from this.” What next? Despite growing concerns about the negative impact of technology and social media, very little has been done in the US to regulate it. Experts fear that artificial intelligence will be no different. “One of the reasons many of us do have concerns about the rollout of AI is because over the last 40 years as a society we’ve basically given up on actually regulating technology,” Wang said. Still, positive efforts have been made by legislators in recent months, with Congress calling the Open AI CEO, Sam Altman, to testify about safeguards that should be implemented. Finlay said she was “heartened” by such moves but said more needed to be done to create shared protocols on AI technology and its release. “Just as hard as it is to predict doomsday scenarios, it is hard to predict the capacity for legislative and regulatory responses,” she said. “We need real scrutiny for this level of technology.” Although the harms of AI are top of mind for most people in the artificial intelligence industry, not all experts in the space are “doomsdayers”. Many are excited about potential applications for the technology. “I actually think that this generation of AI technology we’ve just stumbled into could really unlock a great deal of potential for humanity to thrive at a much better scale than we’ve seen over the last 100 years or 200 years,” Wang said. “I’m actually very, very optimistic about its positive impact. But at the same time I’m looking to what social media did to society and culture, and I’m extremely cognizant of the fact that there are a lot of potential downsides.”