Elon Musk launches AI startup and warns of a ‘Terminator future’

Publication Date: 2023-07-13

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Elon Musk, Technology sector, Technology startups, Computing, Consciousness, Chatbots, news

Article URL: https://www.theguardian.com/technology/2023/jul/13/elon-musk-launches-xai-startup-pro-humanity-terminator-future

Elon Musk has launched an artificial intelligence startup that will be “pro-humanity”, as he said the world needed to worry about the prospect of a “Terminator future” in order to avoid the most apocalyptic AI scenarios. Musk said xAI would seek to build a system that would be safe because it was “maximally curious” about humanity rather than having moral guidelines programmed into it. “From an AI safety standpoint … a maximally curious AI, one that is trying to understand the universe, is I think going to be pro-humanity,” he said on a Spaces discussion on Twitter announcing xAI. The world’s wealthiest person was one of the signatories to a letter this year that called for a pause in building large AI models such as ChatGPT, the chatbot built by the US firm OpenAI. There are growing fears that development of AI technology will race beyond human control. Musk said a pause no longer seemed realistic and he hoped xAI would provide an alternative path. “If I could press pause on AI or really advanced AI digital superintelligence I would. It doesn’t seem like that is realistic so xAI is essentially going to build an AI … in a good way, sort of hopefully,” he said. Musk, who owns Twitter, said there was a benign scenario in which the emergence of artificial general intelligence – a system capable of human-level intelligence – led to an “age of plenty” where there was no shortage of goods and services. However, there was also the possibility of a darker future, he added. Referring to the Terminator films and their vision of a future destroyed by AI-powered robots, Musk said: “It’s actually important for us to worry about a Terminator future in order to avoid a Terminator future.” He said superintelligence – AI more intelligent and gifted than humans – could be five or six years away, which is faster than many experts’ estimates. Musk, who is also chief executive of electric car firm Tesla, said it would be a “while” before xAI reaches the level of OpenAI or Google, which has released its own chatbot called Bard and owns the world-leading UK artificial intelligence firm DeepMind. On Thursday, Google announced that Bard, which has already been launched in the US and UK, would be rolled out across rest of Europe and Brazil with new features including the option of the chatbot speaking its answers back to a user, alongside using images when prompting it. The team at xAI includes Igor Babuschkin, a former engineer at DeepMind; Tony Wu, who worked at Google; Christian Szegedy, who was also a research scientist at Google; and Greg Yang, who was previously at Microsoft. In March, Musk registered a firm named X.AI Corp, incorporated in Nevada, according to a state filing. The firm lists Musk as the sole director and Jared Birchall, the managing director of the multibillionaire’s family office, as a secretary. Dan Hendrycks, who will advise the xAI team, is director of the Center for AI Safety (Cais) and his work revolves around the risks of AI. The Cais issued a statement in May, carrying multiple signatures from AI professionals

and experts, saying that dealing with the risk of extinction from artificial intelligence should be a global priority on a par with mitigating the risk of pandemics and nuclear war. Musk said in the Spaces event that he had doubts over the safety implications of programming a moral stance into generative AI tools. This week, Anthropic, a leading US AI firm, launched a chatbot that operates from a list of safety principles drawn from sources such as the Universal Declaration of Human Rights. “If you programme a certain reality [into an AI] you have to say what morality you are programming. Whose decision is that?” he asked, adding that once an AI is programmed with a specific moral standpoint it would be easier to prompt it into reversing it. This is known as the “Waluigi effect”, named after Luigi’s mischievous arch-rival in the Super Mario video game franchise. Musk’s new company is separate from X Corp, but will work closely with Twitter, Tesla and other companies, according to its website, and is recruiting experienced engineers and researchers in the San Francisco Bay area.