Dangerous AI systems need a ‘smoke alarm’, warn UK ministers

Publication Date: 2023-09-25

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, news

Article URL: https://www.theguardian.com/technology/2023/sep/25/dangerous-ai-systems-need-a-smoke-alarm-warn-ministers

A “smoke alarm” for dangerous artificial intelligence systems is needed in order to head off a range of serious threats such as mass loss of life, cyber-attacks and AI technology spiralling out of control, UK ministers have warned. The technology secretary, Michelle Donelan, said she hoped a forthcoming safety summit hosted in the UK would help to establish an early warning system whereby tech companies look for risks in the AI products they are building and know how to respond to them. “We need [something] almost like a smoke alarm established so that not only are companies searching for the risks, but they have a response to the risk,” she said. “That’s the type of system that we need to be seeing across the board.” Speaking at Bletchley Park, where the two-day summit will be staged in November, Donelan said there were “incredible opportunities” around AI, but “we can only really seize those opportunities if we’re gripping the risks”. The government said on Monday the gathering would focus on two areas in particular: misuse of AI systems to create bioweapons or cyber-attacks; and an inability to control the most advanced systems. It is understood that Rishi Sunak believes time is running out to forge a global consensus on what represent the most serious AI risks and how to deal with them, as tech firms use greater computing power, technological breakthroughs and increased investment to build ever more powerful models. The summit is not expected to produce a nuclear weapons-style international treaty on AI development but is instead expected to outline the range of serious risks that AI systems could pose and measures to mitigate them. It will focus on “frontier” AI models, or cutting-edge systems whose power matches or exceeds the most advanced models currently operating – and could represent a threat to human life. The government said the summit was hoping to identify where the frontier of AI development is now and where it may head. The gathering at Bletchley Park, Buckinghamshire, the home of codebreakers including Alan Turing during the second world war, will be attended by global leaders, AI companies, academics and civil society groups. The government said in a statement on Monday that models “many times” more powerful than those currently operating, such as the GPT-4 model that powers OpenAI’s ChatGPT, could be released soon. “The capabilities of these models are very difficult to predict – sometimes even to those building them – and by default they could be made available to a wide range of actors, including those who might wish us harm.” A concern among AI experts is that advanced systems could evade human control. These fears centre on the possibility of breakthroughs in artificial general intelligence, the term for an AI with human or above-human levels of intelligence, which could theoretically surmount any guardrails put around it. The government’s statement previewing the summit on Monday said that systems evading would be a subject of focus. It referred to “loss of control risks” that could emerge from “advanced systems that we would seek to be aligned with our values and intentions”.