AI dangers must be faced ‘head on’, Rishi Sunak to warn ahead of tech summit

Publication Date: 2023-10-25

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Rishi Sunak, news

Article URL: https://www.theguardian.com/technology/2023/oct/25/ai-dangers-must-be-faced-head-on-rishi-sunak-to-tell-tech-summit

Artificial intelligence brings new dangers to society that must be addressed “head on”, the prime minister will warn on Thursday, as the government admitted it could not rule out the technology posing an existential threat. Rishi Sunak will refer to the “new opportunities” for economic growth offered by powerful AI systems but will also acknowledge they bring “new dangers” including risks of cybercrime, designing of bioweapons, disinformation and upheaval to jobs. In a speech delivered as the UK government prepares to host global politicians, tech executives and experts at an AI safety summit in Bletchley Park next week, Sunak is expected to call for honesty about the risks posed by the technology. “The responsible thing for me to do is to address those fears head on, giving you the peace of mind that we will keep you safe, while making sure you and your children have all the opportunities for a better future that AI can bring,” Sunak will say. “Doing the right thing, not the easy thing, means being honest with people about the risks from these technologies.” The risks from AI were outlined in government documents published on Wednesday. One paper on future risks of frontier AI – the term for advanced AI systems that will be the subject of debate at the summit – states that existential risks from the technology cannot be ruled out. “Given the significant uncertainty in predicting AI developments, there is insufficient evidence to rule out that highly capable Frontier AI systems, if misaligned or inadequately controlled, could pose an existential threat.” The document adds, however, that many experts consider the risk to be very low. Such a system would need to be given or gain control over weapons or financial systems and then be able to manipulate them while rendering safeguards ineffective. The document also outlines a number of alarming scenarios for the development of AI. One warns of a public backlash against the technology led by workers whose jobs have been affected by AI systems taking their work. “AI systems are deemed technically safe by many users … but they are nevertheless causing impacts like increased unemployment and poverty,” says the paper, creating a “fierce public debate about the future of education and work”. In another scenario, dubbed the “wild west”, misuse of AI to perpetrate scams and fraud causes social unrest as many people fall victim to organised crime, businesses have trade secrets stolen on a large scale and the internet becomes increasingly polluted with AI-generated content. One other scenario depicts the creation of a human-level artificial general intelligence that passes agreed checks but triggers fears it could bypass safety systems. The documents also refer to experts warning of the risk that the existential question draws attention “away from more immediate and certain risks”. A discussion paper to be circulated among the 100 attendees at the summit outlines a number of these risks. It states the current wave of innovation in AI will “fundamentally alter the way we live” and could also produce breakthroughs in fields including treating cancer, discovering new drugs and making transport greener.

However, it outlines areas of concern to be discussed at the meeting including the possibility for AI tools to produce “hyper-targeted” disinformation at an unprecedented scale and level of sophistication. “This could lead to ‘personalised’ disinformation, where bespoke messages are targeted at individuals rather than larger groups and are therefore more persuasive,” says the discussion document, which warns of the potential for a reduction in public trust in true information and in civic processes such as elections. “Frontier AI can be misused to deliberately spread false information to create disruption, persuade people on political issues, or cause other forms of harm or damage,” it says. Other risks raised by the paper include the ability of advanced models to perform cyber-attacks and design biological weapons. The paper states there are no established standards or engineering best practices for safety testing of advanced models. It adds that systems are often developed in one country and deployed in another, underlining the need for global coordination. “Frontier AI may help bad actors to perform cyber-attacks, run disinformation campaigns and design biological or chemical weapons,” the document states. “Frontier AI will almost certainly continue to lower the barriers to entry for less sophisticated threat actors.” The technology could “significantly exacerbate” cyber risks, for instance by creating tailored phishing attacks – where someone is tricked, often via email, into downloading malware or revealing sensitive information like passwords. Other AI systems have helped create computer viruses that change over time in order to avoid detection, the document says. It also warns of a “race to the bottom” by developers where the priority is rapid development of systems while under-investing in safety systems. The discussion document also flags job disruption, with the IT, legal and financial industries most exposed to upheaval from AI automating certain tasks. It warns that systems can also reproduce biases contained in the data they are trained on. The document states: “Frontier AI systems have been found to not only replicate but also to perpetuate the biases ingrained in their training data.”