Sunak announces UK AI safety institute but declines to support moratorium

Publication Date: 2023-10-26

Author: Dan Milmo

Section: Technology

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

Article URL: https://www.theguardian.com/technology/2023/oct/26/sunak-announces-uk-ai-safety-institute-but-declines-to-support-moratorium

Rishi Sunak has announced the establishment of a UK AI safety institute but has declined to support a moratorium on advanced development of the technology. The prime minister said the institute would be a world first and would test new types of AI for a range of risks from generating misinformation to posing an existential threat. Announcing the move before next week's global summit on AI safety at Bletchley Park, Sunak said the institute would "advance the world's knowledge of AI safety". "It will carefully examine, evaluate and test new types of AI so that we understand what each new model is capable of," he said in a speech at the Royal Society, an association of leading scientists. He said it would explore "all the risks, from social harms like bias and misinformation through to the most extreme risks of all". A prototype of the safety institute, which the government hopes will become a vehicle for international collaboration on AI safety, already exists in shape of the UK's frontier AI taskforce, which is scrutinising the safety of cutting-edge AI models and was established this year. Sunak said a pause in developing powerful models was not feasible. Asked after the speech if he would support a moratorium or ban on developing a highly capable form of AI known as artificial general intelligence, he said: "I don't think it's practical or enforceable. As a matter of principle, the UK has rightly been an economy and society that has encouraged innovation for all the good that it can bring. And I think that is the right approach." The debate over AI safety reached a new peak in March when an open letter signed by thousands of prominent tech figures including Elon Musk called for an immediate pause in the creation of "giant" AIs for at least six months. Sunak said it was still unclear whether China would attend the summit, despite Beijing receiving an invite to attend along with technology executives, experts and other global leaders. The prime minister said he could not say with "100% certainty" if Chinese officials would join. Liz Truss, Sunak's predecessor, added to the pressure over Chinese attendance on Thursday by asking the prime minister to rescind the invitation, warning that Beijing has used technology to attack "freedom and democracy". The White House confirmed Kamala Harris, the vice president, will attend the summit. She will deliver a speech on the US approach to AI on 1 November before attending the event on 2 November, when Sunak will convene a smaller group of international partners, companies and experts to discuss what concrete steps can be taken to address AI risks. UK officials said they did not see the planned speech by Harris as overshadowing the summit. One potential development in AI that alarms some experts is AGI, the term for a system that can carry out an array of tasks at a human level of intelligence or beyond. Sunak was speaking after the government released its assessment of AI safety risks including the admission that an existential threat from the technology could not 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," said a government document published on Wednesday. Other threats detailed in the government risk papers included the ability of systems to design bioweapons, mass-produce "hyper-targeted" disinformation and cause substantial disruption to the jobs market. Sunak said the worst-case scenario of an existential threat from a "superintelligent" system that evades human control was a scenario that divided opinion among experts and might not happen at all. He added, nonetheless, that major AI developers had voiced concerns about existential risks. "However uncertain and unlikely these risks are, if they did manifest themselves, the consequences would be incredibly serious," he said. "And when so many of the biggest developers of this technology themselves warn of these risks, leaders have a responsibility to take them seriously and to act." Sunak added that he would the use the two-day summit to call for the formation of an expert AI monitoring group similar to the Intergovernmental Panel on Climate Change. "Next week, I will propose that we establish a truly global expert panel nominated by the countries and organisations attending [the summit] to publish a state of AI science report," he said.