Tech firms to allow vetting of AI tools, as Musk warns all human jobs threatened

Publication Date: 2023-11-03

Author: Dan Milmo

Section: Technology

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

Article URL: https://www.theguardian.com/technology/2023/nov/02/top-tech-firms-to-let-governments-vet-ai-tools-sunak-says-at-safety-summit

The most advanced technology companies will allow governments to vet their artificial intelligence tools for the first time, Rishi Sunak has announced, as Elon Musk warned the technology could eventually replace all human jobs. Companies including Meta, Google DeepMind and OpenAI have agreed to allow regulators to test their latest AI products before releasing them to the public, in a move that officials say will slow the race to develop systems that can compete with humans. Sunak made the announcement on Thursday after a two-day summit at Bletchley Park at which a diverse group including the world's richest man, the vice-president of the US and a senior Chinese government official agreed that AI poses a grave risk to humanity. Speaking to reporters at the end of the summit, Sunak said: "I believe the achievements of this summit will tip the balance in favour of humanity." The prime minister also announced international support for an expert body inspired by the Intergovernmental Panel on Climate Change, to be chaired by one of the "godfathers" of modern AI. The moves were welcomed afterwards by the technology billionaire Elon Musk in a conversation between the pair in central London, during which Musk described what he sees as a dramatically different future for humanity. "We are seeing the most disruptive force in history here," he said. "There will come a point where no job is needed. You can have a job if you want a job … but the AI will be able to do everything." Musk said he thought the summit had achieved a meaningful shift in the development of advanced AI. "Simply having an insight and being able to highlight concerns to the public will be very powerful," he said. Speaking at the close of the summit, Sunak said agreements reached with multiple countries and AI companies had significantly reduced the threat posed by the technology. However, he was forced to defend the voluntary nature of the testing agreement, with his government declining to introduce legislation to rein in AI development. Explaining that the UK had to move faster than a legislative timetable would allow, he said: "Technology is developing at such a pace that governments have to make sure that we can keep up." He said that ultimately "binding requirements" would probably be necessary for AI firms. Under the agreement announced at Bletchley Park, "like-minded" governments and AI companies have agreed to work together on testing the safety of new AI models before and after they are released. Large language models, which underpin tools like the ChatGPT chatbot, will be tested in collaboration with governments against a range of dangers including national security, safety and societal harms. The move follows the issuance of an executive order by the White House this week requiring major tech firms to submit test results for their models to the US government before they go public. Sunak said the testing would be led by new AI safety institutes in the US and UK, with the British body positioning itself as a "global hub" for the multilateral initiative – which does not include the participation of China. The testing agreement is backed by the EU and 10 countries including the

US, the UK, Japan, France and Germany. The leading AI companies that have agreed to testing include Google, OpenAI, Amazon, Microsoft and Meta. Sunak said the UN secretary general, António Guterres, had helped to secure international community backing for an expert panel to publish a "state of AI science" report. "This idea is inspired by the way the Intergovernmental Panel on Climate Change was set up to reach international science consensus," Sunak said. "With the support of the UN secretary general, every country has committed to nominate experts." Yoshua Bengio, known as one of the godfathers of modern AI, will chair the production of the first safety report. Bengio, a winner of the ACM Turing award – the computer science equivalent of the Nobel prize – has been a prominent voice of caution in the debate over AI development. He was the lead signatory of a letter published in March calling for a six-month hiatus in "giant" AI experiments, and he backed a statement in May warning that the risk of extinction from AI should be treated with the same priority as mitigating the societal risk from pandemics and nuclear war. Speaking to reporters earlier on Thursday, Sunak raised the nuclear and pandemic comparison. "People developing this technology themselves have raised the risk that AI may pose and it's important to not be alarmist about this," he said. "There's debate about this topic. People in the industry themselves don't agree and we can't be certain. "But there is a case to believe that it may pose a risk on a scale like pandemics and nuclear war, and that's why, as leaders, we have a responsibility to act to take the steps to protect people, and that's exactly what we're doing." Concerns around AI – the term for computer systems that can perform tasks typically associated with intelligent beings – range from the threat of mass-produced disinformation in elections to advanced systems evading control and threatening humanity. The prime minister's announcements were the culmination of a second day of intense diplomatic activity at the international AI safety summit. British officials were delighted to be able to issue a communique at the beginning of the summit, signed by 28 governments including the UK, the US, EU and China. The "Bletchley declaration" promised that the signatories would work together on shared safety standards in a process officials likened to the Cop summits on the climate crisis. Sunak's attempts to position the UK as the world leader in developing new AI rules were partly overshadowed on Wednesday by an announcement by the US commerce secretary, Gina Raimondo, of a new AI safety institute in Washington. British officials said they expected to work closely with the new US institute and others to create a network of similar organisations that could do testing around the world, with the multilateral testing agreement forming a framework for that collaboration. Speaking at a press conference at the close of the summit, Sunak said he recognised concerns about the potential for AI to eliminate jobs, but workers should view the technology as a "co-pilot". "I know this is an anxiety that people have," he said. "We should look at AI much more as a co-pilot than something that necessarily is going to replace someone's job. "AI is a tool that can help almost everybody do their jobs better, faster, quicker, and that's how we're already seeing it being deployed."