Downing Street trying to agree statement about AI risks with world leaders

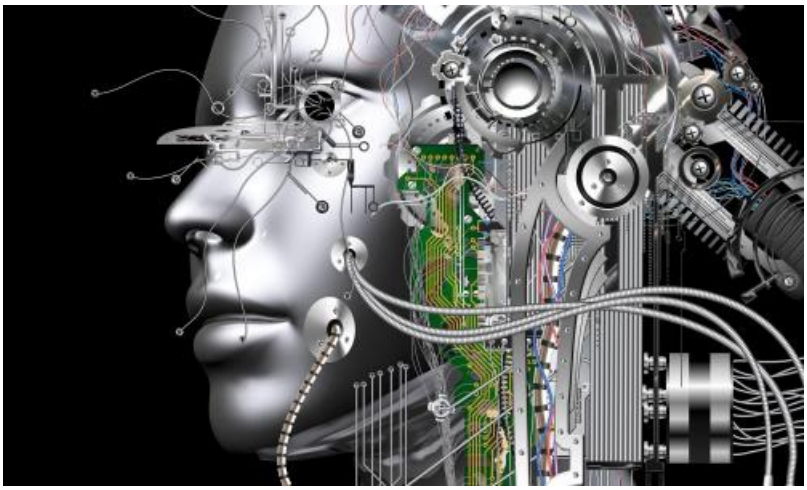
Publication Date: 2023-10-10

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), news

Article URL: https://www.theguardian.com/technology/2023/oct/10/downing-street-trying-to-agree-statement-about-ai-risks-with-world-leaders



Rishi Sunak's advisers are trying to thrash out an agreement among world leaders on a statement warning about the risks of artificial intelligence as they finalise the agenda for the AI safety summit next month. Downing Street officials have been touring the world talking to their counterparts from China to the EU and the US as they work to agree on words to be used in a communique at the two-day conference. But they are unlikely to agree a new international organisation to scrutinise cutting-edge AI, despite interest from the UK in giving the government's AI taskforce a global role. Sunak's AI summit will produce a communique on the risks of AI models, provide an update on White House-brokered safety guidelines and end with "like-minded" countries debating how national security agencies can scrutinise the most dangerous versions of the technology. The possibility of some form of international cooperation on cutting-edge AI that can pose a threat to human life will also be discussed on the final day of the summit on 1 and 2 November at Bletchley Park, according to a draft agenda seen by the Guardian. The draft refers to establishing an "AI Safety Institute" to enable the national security-related scrutiny of frontier AI models – the term for the most advanced versions of the technology. However, last week the prime minister's summit representative downplayed the establishment of such an organisation, although he emphasised that "collaboration is key" in managing frontier AI risks. In a post last week on X, formerly known as Twitter, Matt Clifford wrote: "It's not about setting up a single new international institution. Our view is that most countries will want to develop their own capabilities in this space, particularly to evaluate frontier models." The UK is leading the way in the frontier AI process so far, having established a frontier AI taskforce under the tech entrepreneur Ian Hogarth. The deputy prime minister, Oliver Dowden, said last month he hoped the taskforce "can evolve to become a permanent institutional structure, with an international offer on AI safety". Clifford announced last week that about 100 people would attend the summit, drawn from cabinet ministers around the world, company chief executives, academics and representatives from international civil society. According to the draft agenda, the summit includes a three-track discussion on day one based on a discussion of risks associated with frontier models, a discussion of mitigating those risks, and discussing opportunities from those models. This would be followed by short communique to be signed by country delegations that expresses a consensus on the risks and opportunities of frontier models. Companies participating in the summit, which are expected to include ChatGPT developer OpenAI, Google and Microsoft, will then publish details about how they are adhering to AI safety commitments agreed with the White House in July. Those commitments include external security testing of AI models before they are released and ongoing scrutiny of those systems once they are operating. According to a report in Politico last week, the White House is updating the

voluntary commitments – with reference to safety, cybersecurity and how AI systems could be used for national security purposes – and could make an announcement this month. The second day will feature a smaller gathering of about 20 people including “like-minded” countries, according to the draft agenda, with a conversation about where AI could be in five years’ time and positive AI opportunities linked to sustainable development goals. This included a discussion about a safety institute. In his thread on X, Clifford said the UK remained keen on collaborating with other countries on AI safety. “Collaboration is key to ensuring we can manage risks from Frontier AI – with civil society, academics, technical experts and other countries,” he wrote. A government spokesperson said: “We have been very clear that these discussions will involve exploring areas for potential collaboration on AI safety research, including evaluation and standards. “International discussions on this work are already under way and are making good progress, including discussing how we can collaborate across countries and firms and with technical experts to evaluate frontier models. There are many different ways to do this and we look forward to convening this conversation in November at the summit.”