TechScape: Why Sunak's 'vanity jamboree' on AI safety was actually … a success

Publication Date: 2023-11-07

Author: Dan Milmo

Section: Technology

Tags: Technology, TechScape newsletter, Artificial intelligence (AI), Computing, Consciousness, Rishi Sunak, Sam Bankman-Fried, newsletters

Article URL: https://www.theguardian.com/technology/2023/nov/07/techscape-rishi-sunak-ai-safety-summit-bletchley-park

For Max Tegmark, last week's artificial intelligence summit at Bletchley Park was an emotional moment. The MIT professor and AI researcher was behind a letter this year calling for a pause in development of advanced systems. It didn't happen, but it was a crucial contribution to the political and academic momentum that resulted in the Bletchley gathering. "[The summit] has actually made me more optimistic. It really has superseded my expectations," he told me. "I've been working for about 10 years, hoping that one day there would be an international summit on AI safety. Seeing it happen with my own eyes – and done so surprisingly well – was very moving." Clutching a £50 note with the face of Bletchley codebreaker Alan Turing on it, Tegmark added that the computing genius – a foundational figure in the history of AI – had been proven right. "I agree with Turing – the default outcome if we just rush to build machines that are much smarter than us is that we lose control over our future and we'll probably get wiped out." But Tegmark thinks progress was made at Bletchley. Here is a quick summary of what happened. The Bletchley Declaration The summit began with a communique signed by almost 30 governments including the US and China, along with the EU. Rishi Sunak described the statement as "quite incredible", having succeeded in getting competing superpowers, and countries with varied views on AI development, to agree to a joint message. The declaration starts with a reference to AI providing "enormous global opportunities" with potential to "transform and enhance human wellbeing, peace and prosperity" – but the technology needs to be designed in a way that is "human-centric, trustworthy and responsible". There is also an emphasis on international cooperation, including a reference to an "internationally inclusive network of scientific research" on AI safety. But the most noteworthy section referred to the summit's central purpose: making sure frontier AI – the term for the most advanced systems – does not get horribly out of hand. Referring to AI's potential for causing catastrophe, it said: "There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models." That attention-grabbing sentence was preceded by a reference to more immediate harms like cyber-attacks and disinformation. The debate over whether AI could wipe out humanity is ongoing – there is also a belief that the fears are overplayed – but there did appear to be a consensus that AI-generated disinformation is an immediate concern that needs to be addressed. Sunak's ambition to make the summit a regular event has been met. South Korea will host a virtual event in six months and France will host a full-blown summit in 12 months. So will those carefully assembled words lead to regulatory or legislative change? Charlotte Walker-Osborn, technology partner at the international law firm Morrison Foerster, says the declaration will "likely further drive some level of international legislative and governmental consensus around key tenets for regulating AI”. For example, she cites core tenets such as transparency around when and how AI is being used, information on the data used in training systems and a requirement for trustworthiness (covering everything from biased outcomes to deepfakes). However, Walker-Osborn says a “truly uniform approach is unlikely” because of “varying approaches to regulation and governance in general” between countries. Nonetheless, the declaration is a landmark, if only because it recognises that AI cannot continue to develop without stronger oversight. State of AI report Sunak announced a “state of AI science” report at the summit, with the inaugural one chaired by Yoshua Bengio, one of three so-called “godfathers of AI”, who won the ACM Turing award – the computer science equivalent of the Nobel prize – in 2018 for his work on artificial intelligence. The group writing the report will include leading AI academics and will be supported by an advisory panel drawn from the countries that attended the summit (so the US and China will be on it). Bengio was a signatory of Tegmark’s letter and also signed a statement in May warning that mitigating the risk of extinction from AI should be a global priority alongside pandemics and nuclear war. He takes the subject of AI safety seriously. The UK prime minister said the idea was inspired by the Intergovernmental Panel on Climate Change and was supported by the UN secretary-general, António Guterres, who attended the summit. However, it won’t be a UN-hosted project and the UK government-backed AI safety institute will host Bengio’s office for the report. International safety testing A group of governments attending the summit and major AI firms agreed to collaborate on testing of their AI models before and after their public release. The 11 government signatories included the EU, the US, the UK, Australia, Japan – but not China. The eight companies included Google, ChatGPT developer OpenAI, Microsoft, Amazon and Meta. The UK has already agreed partnerships between its AI safety institute and its US counterpart (which was announced ahead of the summit last week) and also with Singapore, to collaborate on safety testing. This is a voluntary set-up and there is some scepticism about how much impact the Bletchley announcements will have if they are not underpinned by regulation. Sunak told reporters last week that he was not ready to legislate yet and further testing of advanced models is needed first (although he added that “binding requirements” will probably be needed at some point). It means that the White House’s executive order on AI use, issued in the same week as the summit, and the forthcoming European Union’s AI Act are further ahead of the UK in introducing new, binding regulation of the technology. “When it comes to how the model builders behave … the impending EU AI Act and President Biden’s executive order are likely to have a larger impact,” says Martha Bennett, a principal analyst at the company Forrester. Others, nonetheless, are happy with how Bletchley has shaped the debate and brought disparate views together. Prof Dame Muffy Calder, vice-principal and head of the college of science and engineering at the University of Glasgow, was worried the summit would dwell too much on existential risk and not “real and current issues”. That fear, she believes, was assuaged. “The summit and declaration go beyond just the risks of ‘frontier AI’,” she says. “For example, issues like transparency, fairness, accountability, regulation, appropriate human oversight, and legal frameworks are all called out explicitly in the declaration. As is cooperation. This is great.” Read more on this story Five takeaways from the summit. Sister newsletter First Edition runs through what we learned about the dangers of AI. The great powers signed up to Sunak’s meetup – while jostling for position. Zoe Williams is very good on the problem with tech bro philanthropy. It’s not AI but the tech giants that control it that need reining in, writes John Naughton. AI pioneer Fei-Fei Li: “I’m more concerned about the risks that are here and now.” Sunak’s summit trades in Silicon Valley celebrity and lays bare the UK’s Brexit dilemmas, argues Rafael Behr. If you want to read the complete version of the newsletter please subscribe to receive TechScape in your inbox every Tuesday