# UK’s AI Safety Institute ‘needs to set standards rather than do testing’

Publication Date: 2024-02-11

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, Chatbots, news

Article URL: https://www.theguardian.com/technology/2024/feb/11/ai-safety-institute-needs-to-set-standards-rather-than-do-testing

The UK should concentrate on setting global standards for artificial intelligence testing instead of trying to carry out all the vetting itself, according to a company assisting the government’s AI Safety Institute. Marc Warner, the chief executive of Faculty AI, said the newly established institute could end up “on the hook” for scrutinising an array of AI models – the technology that underpins chatbots like ChatGPT – owing to the government’s world-leading work in AI safety. Rishi Sunak announced the formation of the AI Safety Institute (AISI) last year ahead of the global AI safety summit, which secured a commitment from big tech companies to cooperate with the EU and 10 countries, including the UK, US, France and Japan, on testing advanced AI models before and after their deployment. The UK has a prominent role in the agreement because of its advanced work on AI safety, underlined by the establishment of the institute. Warner, whose London-based company has contracts with the UK institute that include helping it test AI models on whether they can be prompted to breach their own safety guidelines, said the institute should be a world leader in setting test standards. “I think it’s important that it sets standards for the wider world, rather than trying to do everything itself,” he said. Warner, whose company also carries out work for the NHS on Covid and the Home Office on combating extremism, said the institute had made a “really great start” and that, “I don’t think I’ve ever seen anything in government move as fast as this.” He added, however, that “the technology is moving fast as well”. He said the institute should put in place standards that other governments and companies can follow, such as “red teaming”, where specialists simulate misuse of an AI model, rather than take on all the work itself. Warner said the government could find itself in a situation where it was “red teaming everything” and that a backlog could build up “where they don’t have the bandwidth to get to all the models fast enough”. Referring to the institute’s potential as an international standard setter, he said: “They can set really brilliant standards such that other governments, other companies … can red team to those standards. So it’s a much more scalable, long-term vision for how to keep these things safe.” Warner spoke to the Guardian shortly before AISI released an update on its testing programme last week and acknowledged that it did not have the capacity to test “all released models” and will focus on the most advanced systems only. Last week, the Financial Times reported that big AI companies are pushing the UK government to speed up its safety tests for AI systems. Signatories to the voluntary testing agreement include Google, the ChatGPT developer OpenAI, Microsoft and Mark Zuckerberg’s Meta. The US has also announced an AI safety institute which will take part in the testing programme announced at the summit in Bletchley Park. Last week, the Biden administration announced a consortium to assist the White House in meeting the goals set out in its October executive order on AI safety, which include developing guidelines for watermarking AI-generated content. Members of the consortium, which will be housed under the US institute, include Meta, Google, Apple and OpenAI. The UK's department for science, innovation and technology said governments around the world "need to a play a key role" in testing AI models. "The UK is driving forward that effort through the world's first AI Safety Institute, who are conducting evaluations, research and information sharing, and raising the collective understanding of AI safety around the world," a spokesperson said. "The institute's work will continue to help inform policymakers across the globe on AI safety."