As the AI world gathers in Seoul, can an accelerating industry balance progress against safety?

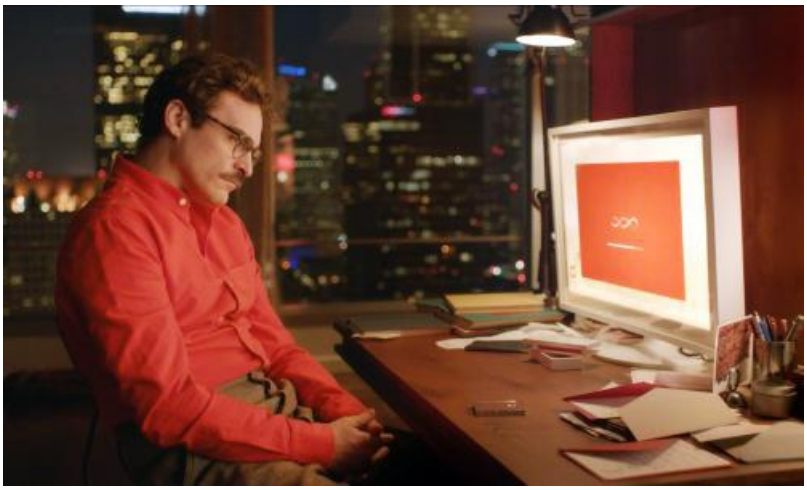
Publication Date: 2024-05-18

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, OpenAI, Alphabet, Google, Technology sector, features

Article URL: https://www.theguardian.com/technology/article/2024/may/18/ai-seoul-global-summit-safety-openai-meta



This week, artificial intelligence caught up with the future – or at least Hollywood's idea of it from a decade ago. "It feels like AI from the movies," wrote the OpenAI chief executive, Sam Altman, of his latest system, an impressive virtual assistant. To underline his point he posted a single word on X – "her" – referring to the 2013 film starring Joaquin Phoenix as a man who falls in love with a futuristic version of Siri or Alexa, voiced by Scarlett Johansson. For some experts, that new AI, GPT-4o, will be an unsettling reminder of their concerns about the technology's rapid advances, with a key OpenAI safety researcher leaving this week following a disagreement over the company's direction. For others the GPT-4o release will be confirmation that innovation continues in a field promising benefits for all. Next week's global AI summit in Seoul, attended by ministers, experts and tech executives, will hear both perspectives, as underlined by a safety report released before the meeting that referred to potential positives as well as numerous risks. The inaugural AI Safety Summit at Bletchley Park in the UK last year announced an international testing framework for AI models, after calls from some concerned experts and industry professionals for a six-month pause in development of powerful systems. There has been no pause. The Bletchley declaration signed by UK, US, EU, China and others hailed the "enormous global opportunities" from AI but also warned of its potential for causing "catastrophic" harm. It also secured a commitment from big tech firms including OpenAI, Google and Mark Zuckerberg's Meta to cooperate with governments on testing their models before they are released. While the UK and US have established national AI safety institutes, the industry's development of AI has continued. Big tech firms and others have all announced new AI products recently: OpenAI released GPT-4o (the o stands for "omni") for free online; a day later, Google previewed a new AI assistant called Project Astra as well as updates to its Gemini model. Last month, Meta released new versions of its own AI model, Llama, and continues to offer them "open source", meaning they are freely available to use and adapt; and in March, the AI startup Anthropic, formed by former OpenAI staff who disagreed with Altman's approach, updated its Claude model and took the lead in capability on offer. Dan Ives, an analyst at the US stockbroker Wedbush Securities, estimates that the spending boom on generative AI – the general term for this latest method of building smart systems – will reach $100bn (£79bn) this year, part of a $1tn expenditure over the next decade. More landmark developments are looming: OpenAI is working on its next model, GPT-5, as well as a search engine; Google is preparing to release Astra and is rolling out AI-generated search queries outside the US; Microsoft is reportedly working on its own AI model and has hired the British entrepreneur Mustafa Suleyman to oversee a new AI division; Apple is reportedly in talks with OpenAI to put ChatGPT in its smartphones; and billions of dollars of AI investment is being poured into tech firms of all sizes. Hardware startups such as Humane and Rabbit are racing to build the AI-powered

replacement for the smartphone, while others are experimenting with how much of one person's life can be used to teach an AI. The US-based startup Rewind is marketing a product that records every action you ever take on your computer screen, training an AI system to know your life in intricate detail. Coming soon, it is a lapel-worn mic and camera so that it can even learn from what goes on when you're offline. Niamh Burns, a senior analyst at Enders Analysis, says there will be a stream of new products as companies, backed by multibillion-dollar investments, try to win over consumers. "We're going to keep seeing these flashy releases, because the tech is new and exciting, and because the actual consumer use case hasn't been landed on. New models and even just new interfaces – simply put, things to do with the models – need to be released until something sticks from a user perspective," she says. Rowan Curran, an analyst at the research firm Forrester, says the six months since Bletchley have already seen significant changes such as the emergence of so-called "multi-modal" models like GPT-4 and Gemini, meaning they can handle a variety of formats such as text, image and audio. The GPT model that went public in 2022, for instance, could only handle text. "It has really opened up possibilities for AI," says Curran. "Although we have seen a few of these models already I expect many more to emerge." Other recent breakthroughs cited by Curran include the emergence of video generating models such as OpenAI's Sora, which has not been released publicly but whose demos were enough to persuade film and TV mogul Tyler Perry to halt an $800m studio expansion. Then there is retrieval-augmented generation, or RAG, a technique for giving a generalist AI a specialism – turning a video generator such as Sora, for instance, into an anime impresario, or teaching the image generator StableDiffusion how to paint like Picasso, or teaching a chatbot to specialise in scientific papers. Some already see a market that will be dominated by a handful of wealthy companies who can afford the vast energy and data crunching costs that come with building AI models and operating them. Would-be competitors are also being brought under their wings, to the concern of competition authorities in the UK, the US and EU. Microsoft, for instance, is a backer of OpenAI and France's Mistral, while Amazon has invested heavily in Anthropic. "The market for GenAI is febrile," says Andrew Rogoyski, a director at the Institute for People-Centred AI at the University of Surrey. "It is so costly to develop large language models that only the very largest companies, or companies with extraordinarily generous investors, can play." Meanwhile, some experts feel safety is not the priority it should be, because of the rush. "Governments and safety institutes say they plan to regulate and the companies say they are concerned too," says Dame Wendy Hall, a professor of computer science at the University of Southampton and a member of the UN's advisory body on AI. "But progress is slow because companies have to react to market forces." Google and OpenAI point to statements about safety alongside this week's announcements, with Google referring to making its models "more accurate, reliable and safer" and OpenAI detailing how GPT-4o has safety "built-in by design". However, on Friday a key OpenAI safety researcher, Jan Leike, who had resigned earlier in the week, warned that "safety culture and processes have taken a backseat to shiny products" at the company. In response Altman wrote on X that OpenAI was "committed" to doing more on safety. The UK government will not confirm which models are being tested by its newly established AI Safety Institute, but the Department for Science, Innovation and Technology said it was continuing to "work closely with companies to deliver on the agreements reached in the Bletchley declaration". The biggest changes are yet to come. "The last 12 months of AI progress were the slowest they'll be for the foreseeable future," the economist Samuel Hammond wrote in early May. Until now, "frontier" AI systems, the most powerful on the market, have largely been confined to simply handling text. Microsoft and Google have incorporated their offerings into their office products, and given them the authority to carry out simple administrative functions upon request. But the next step of development is "agentic" AI: systems that can truly act to influence the world around them, from surfing the web, to writing and executing code. Smaller AI labs have experimented with such approaches, with mixed successes, putting commercial pressure on the larger companies to give their own AI models the same power. By the end of the year, expect the top AI systems to not only offer to plan a holiday for you, but book the flights, hotels and restaurants, arrange your visa, and prepare and lead a walking tour of your destination. But an AI that can do anything the internet offers is also an AI with a much greater capability for harm than anything before. The meeting at Seoul might be the last chance to discuss what that means for the world before it arrives.