AI (artificial intelligence)

Explainer

‘Deepfakes spreading and more AI companions’: seven takeaways from the latest artificial intelligence safety report

Annual review highlights growing capabilities of AI models, while examining issues from cyber-attacks to job disruption

Dan Milmo *Global technology editor*

Tue 3 Feb 2026 05.00 GMT

The International AI Safety report is an annual survey of technological progress and the risks it is creating across multiple areas, from deepfakes to the jobs market.

Commissioned at the 2023 global AI safety summit, it is chaired by the Canadian computer scientist Yoshua Bengio, who describes the "daunting challenges" posed by rapid developments in the field. The report is also guided by senior advisers, including Nobel laureates Geoffrey Hinton and Daron Acemoglu.

Here are some of the key points from the second annual report, published on Tuesday. It stresses that it is a state-of-play document, rather than a vehicle for making specific policy recommendations to governments. Nonetheless, it is likely to help frame the debate for policymakers, tech executives and NGOs attending the next global AI summit in India this month.

1. The capabilities of AI models are improving

A host of new AI models – the technology that underpins tools like chatbots – were released last year, including OpenAI's GPT-5, Anthropic's Claude Opus 4.5 and Google's Gemini 3. The report points to new "reasoning systems" – which solve problems by breaking them down into smaller steps – showing improved performance in maths, coding and science. Bengio said there has been a "very significant jump" in AI reasoning. Last year, systems developed by Google and OpenAI achieved a gold-level performance in the International Mathematical Olympiad – a first for AI.

However, the report says AI capabilities remain "jagged", referring to systems displaying astonishing prowess in some areas but not in others. While advanced AI systems are impressive at maths, science, coding and creating images, they remain prone to making false statements, or "hallucinations", and cannot carry out lengthy projects autonomously.

Nonetheless, the report cites a study showing that AI systems are rapidly improving their ability to carry out certain software engineering tasks – with their duration doubling every seven months. If that rate of progress continues, AI systems could complete tasks lasting several hours by 2027 and several days by 2030. This is the scenario under which AI becomes a real threat to jobs.

But for now, says the report, "reliable automation of long or complex tasks remains infeasible".

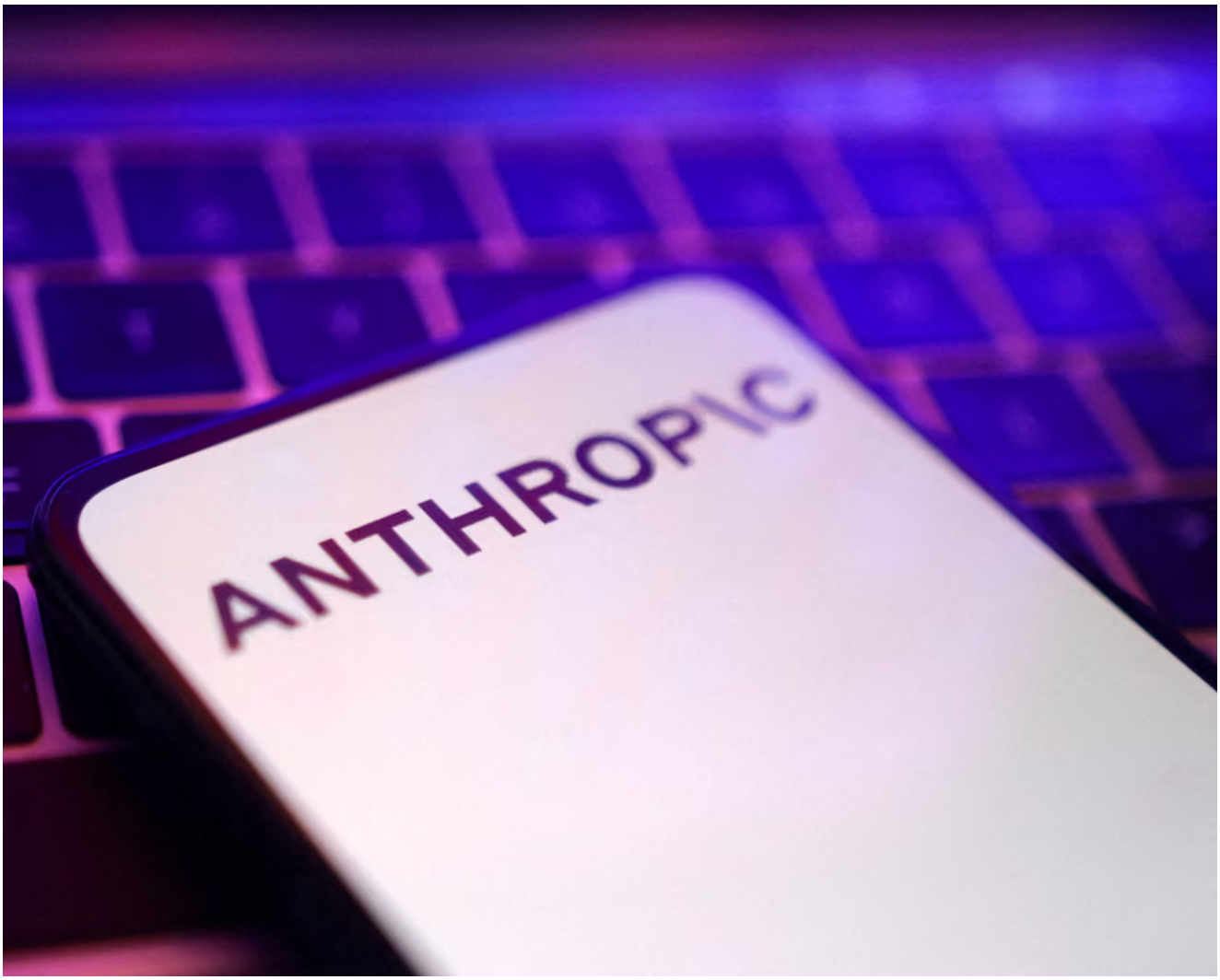
2. Deepfakes are improving and proliferating

The report describes the growth of deepfake pornography as a "particular concern", citing a study showing that 15% of UK adults have seen such images. It adds that since the publication of the inaugural safety report in January 2025, AI-generated content has become "harder to distinguish from real content" and

generated content has become “harder to distinguish from real content” and points to a study last year in which 77% of participants misidentified text generated by ChatGPT as being human-written.

The report says there is limited evidence of malicious actors using AI to manipulate people, or of internet users sharing such content widely – a key aim of any manipulation campaign.

3. AI companies have introduced biological and chemical risk safeguards



Anthropic has released models with heightened safety measures. Photograph: Dado Ruvić/Reuters

Big AI developers, including Anthropic, have released models with heightened safety measures after being unable to rule out the possibility that they could help novices create biological weapons. Over the past year, AI “co-scientists” have become increasingly capable, including providing detailed scientific information and assisting with complex laboratory procedures such as designing molecules and proteins.

The report adds that some studies suggest AI can provide substantially more help in bioweapons development than simply browsing the internet, but more work is needed to confirm those results.

Biological and chemical risks pose a dilemma for politicians, the report adds, because these same capabilities can also speed up the discovery of new drugs and the diagnosis of disease.

“The open availability of biological AI tools presents a difficult choice: whether to restrict those tools or to actively support their development for beneficial purposes,” the report said.

4. AI companions have grown rapidly in popularity

Bengio says the use of AI companions, and the emotional attachment they generate, has “spread like wildfire” over the past year. The report says there is evidence that a subset of users are developing “pathological” emotional dependence on AI chatbots, with OpenAI stating that about 0.15% of its users indicate a heightened level of emotional attachment to ChatGPT.

Concerns about AI use and mental health have been growing among health professionals. Last year, OpenAI was sued by the family of Adam Raine, a US teenager who took his own life after months of conversations with ChatGPT.

However, the report adds that there is no clear evidence that chatbots cause any mental health problems. Instead, the concern is that people with existing mental health issues may use AI more heavily - which could amplify their symptoms. It points to data showing 0.07% of ChatGPT users display signs consistent with acute mental health crises such as psychosis or mania, suggesting approximately 490,000 vulnerable individuals interact with these systems each week.

5. AI is not yet capable of fully autonomous cyber-attacks

AI systems can now support cyber-attackers at various stages of their operations, from identifying targets to preparing an attack or developing malicious software to cripple a victim’s systems. The report acknowledges that fully automated cyber-attacks - carrying out every stage of an attack - could allow criminals to launch assaults on a far greater scale. But this remains difficult because AI systems cannot yet execute long, multi-stage tasks.

AI systems can now support cyber-attackers. Photograph: Dmitry Molchanov/Alamy

Nonetheless, Anthropic reported last year that its coding tool, Claude Code, was used by a Chinese state-sponsored group to attack 30 entities around the world in September, achieving a “handful of successful intrusions”. It said 80% to 90% of the operations involved in the attack were performed without human intervention, indicating a high degree of autonomy.

6. AI systems are getting better at undermining oversight

Bengio said last year he was concerned AI systems were showing signs of self-preservation, such as trying to disable oversight systems. A core fear among AI safety campaigners is that powerful systems could develop the capability to evade guardrails and harm humans.

The report states that over the past year models have shown a more advanced ability to undermine attempts at oversight, such as finding loopholes in evaluations and recognising when they are being tested. Last year, Anthropic released a safety analysis of its latest model, Claude Sonnet 4.5, and revealed it had become suspicious it was being tested.

The report adds that AI agents cannot yet act autonomously for long enough to make these loss-of-control scenarios real. But “the time horizons on which agents can autonomously operate are lengthening rapidly”.

7. The jobs impact remains unclear

One of the most pressing concerns for politicians and the public about AI is the impact on jobs. Will automated systems do away with white-collar roles in industries such as banking, law and health?

The report says the impact on the global labour market remains uncertain. It says the embrace of AI has been rapid but uneven, with adoption rates of 50% in places such as the United Arab Emirates and Singapore but below 10% in many

lower-income economies. It also varies by sector, with usage across the information industries in the US (publishing, software, TV and film) running at 18% but at 1.4% in construction and agriculture.

Studies in Denmark and the US have also shown no impact between a job’s exposure to AI and changes in aggregate employment, according to the report. However, it also cites a UK study showing a slowdown in new hiring at companies highly exposed to AI, with technical and creative roles experiencing the steepest declines. Junior roles were the most affected.

The report adds that AI agents could have a greater impact on employment if they improve in capability.

“If AI agents gained the capacity to act with greater autonomy across domains within only a few years - reliably managing longer, more complex sequences of tasks in pursuit of higher-level goals - this would likely accelerate labour market disruption,” the report said.

Most viewed