The latest artificial intelligence safety report

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".

The Guardian, Dan Milmo Global technology editor - Tue 3 Feb 2026

The latest artificial intelligence safety report

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.

The Guardian, Dan Milmo Global technology editor - Tue 3 Feb 2026