



TOPICS:

AI

SOURCE:

The Alan Turing Institute

The Alan Turing Institute: Frontier AI Risks - A Practical Way Forward

- The paper prioritises **five risks**:
Cybersecurity: AI may accelerate vulnerability discovery and attacks, while dependence on shared model and cloud providers could allow failures to spread across essential services; **Democratic instability**: The concern is less a proven effect on election results than the gradual erosion of trust and shared facts through false information and increasingly personalised influence; **Misalignment**: Autonomous agents can pursue assigned goals through actions their operators did not intend. The paper treats this as an existing problem involving both technology and organisational oversight; **Loss of meaningful human control**: More capable systems could make decisions faster than people can assess or challenge them. The paper presents superintelligent AI within five years as a possibility to prepare for, not a prediction; **Chemical and biological threats**: AI may lower the specialist knowledge needed for harmful activity, although physical materials and laboratory skills remain barriers.
- The Institute proposes research and collaboration on AI-enabled cyber defence, measures of information integrity, assurance for autonomous agents, and verification of **whole deployed systems**, including their people and processes. It also calls for work on whether shutdown mechanisms can be shown to function reliably, how to preserve visibility into AI reasoning, and how to measure AI's contribution to biological threats. A simple "kill switch," it cautions, is not yet a demonstrated general solution.
- The authors say slowing frontier AI would require **binding regulation and international cooperation**. They recommend developing shared safety standards and verification methods, while leaving decisions about regulation and democratic safeguards to policymakers and the public.

