HM Government

Safety and Security Risks of Generative Artificial Intelligence to 2025

Summary

Generative AI development has the potential to bring significant global benefits. But it will also increase risks to safety and security by enhancing threat actor capabilities and increasing the effectiveness of attacks.

- The development and adoption of generative AI technologies has the potential to bring substantial benefits if managed appropriately. Productivity and innovation across many sectors including healthcare, finance and information technology will accelerate.
- Generative AI will also significantly increase risks to safety and security. By 2025, generative AI is more likely to amplify existing risks than create wholly new ones, but it will increase sharply the speed and scale of some threats. The difficulty of predicting technological advances creates significant potential for technological surprise; additional threats will almost certainly emerge that have not been anticipated.
- The rapid proliferation and increasing accessibility of these technologies will almost certainly enable less-sophisticated threat actors to conduct previously unattainable attacks.
- Risks in the digital sphere (e.g. cyber-attacks, fraud, scams, impersonation, child sexual abuse images) are most likely to manifest and to have the highest impact to 2025.
- Risks to political systems and societies will increase in likelihood as the technology develops and adoption widens. Proliferation of synthetic media risks eroding democratic engagement and public trust in the institutions of government.
- Physical security risks will likely rise as Generative AI becomes embedded in more physical systems, including critical infrastructure.
- The aggregate risk is significant. The preparedness of countries, industries and society to mitigate these risks varies. Globally regulation is incomplete and highly likely failing to anticipate future developments.

Our definitions and scope

Safety and Security: The protection, wellbeing and autonomy of civil society and the population.

Artificial Intelligence (AI): Machine-driven capability to achieve a goal by performing cognitive tasks.

Frontier AI: Highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models.

Generative AI (GenAI): AI systems that can create new content. The most popular models generate text and images from text prompts, but some use other inputs such as images to create audio, video and images.

Large language model (LLM): Models trained on large volumes of text-based data, typically from the internet.

Risk: A situation involving exposure to detrimental impacts.

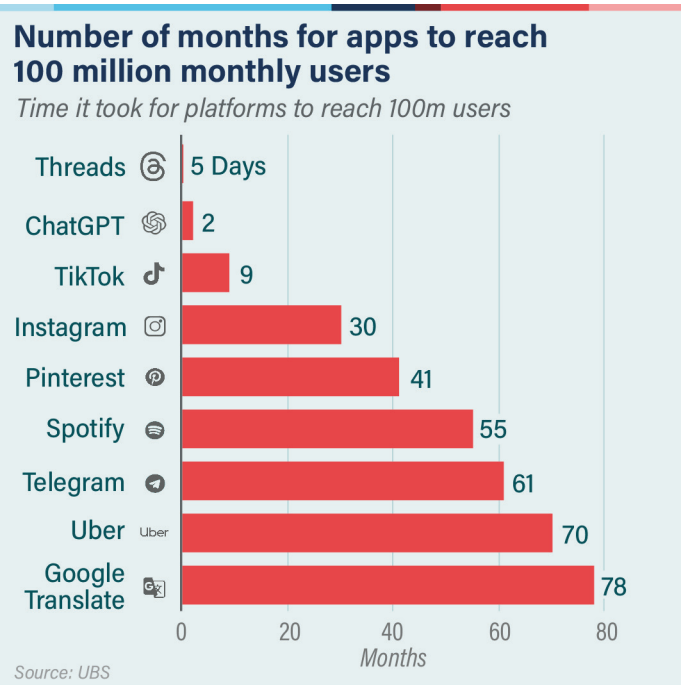
Threat: A malicious risk involving an actor with intent.

This assessment does not consider military risks relating to Generative AI.

This assessment draws on a broad range of sources including existing and novel research, intelligence assessments, expert insights and open source.

Detail

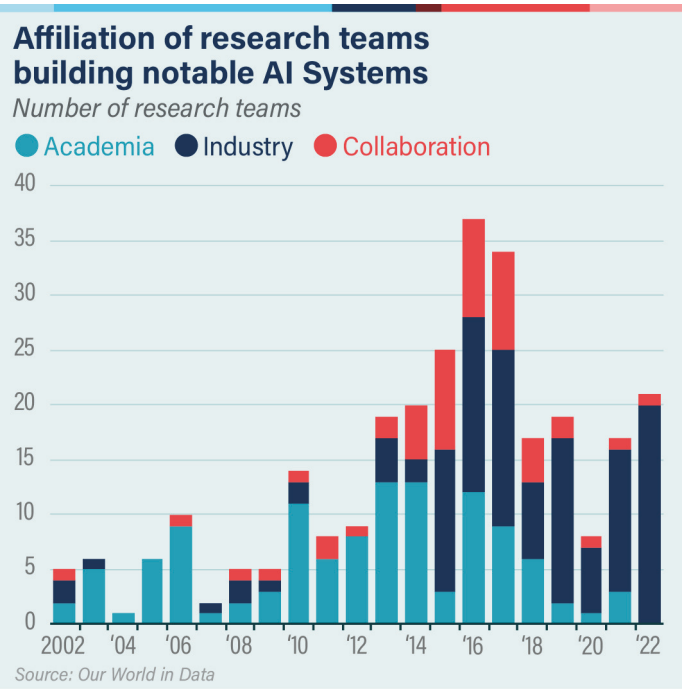
1. The development and application of generative AI intersects with many other technologies. Its development and use will have broad impacts - positive and negative - internationally. The rapid pace of technological progress, lack of consensus on how to measure and compare performance of AI models, and the broad capabilities of the technology means that the safety and security implications are challenging to assess. We have therefore limited our analysis to the key risks and imposed a limited time horizon to 2025. We exclude consideration of the risks resulting from military applications of generative AI.



2. The perceived advantages from first-mover status and widespread media attention have accelerated global interest in generative AI. Since 2020, progress in generative AI has greatly outpaced expert expectations, with models outperforming humans in a small number of specific tasks. Progress continues to be rapid and to 2025, it is unlikely that the pace of technological development will slow. Higher performing, larger LLMs will almost certainly be released, but it is unclear how far this will translate into significantly improved practical applications by 2025. Global regulation is incomplete, falling behind current technical advances and highly likely failing to anticipate future developments.

The Generative AI Ecosystem

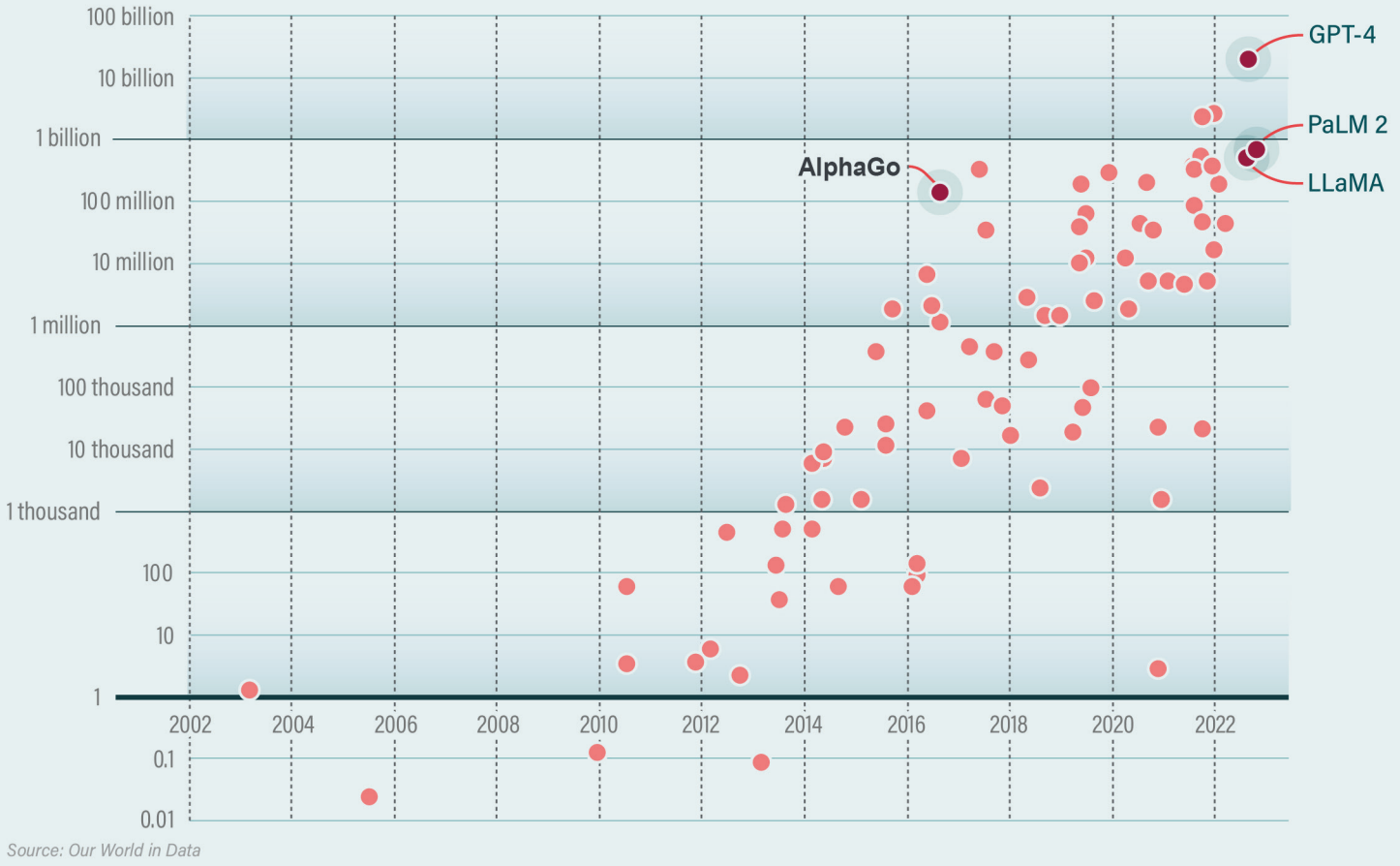
3. Private sector AI firms will remain key actors in cutting-edge generative AI research and frontier models to 2025. The researchers, funding, hardware, compute and data will continue to be concentrated in these commercial organisations, enabling them to undertake the most advanced developments.



4. Open-source generative AI is facilitating rapid proliferation and increasing democratisation of generative AI by reducing the barriers to entry for developing models. To date their performance has mostly lagged behind that of the frontier models; open source models will almost certainly improve, but they are highly unlikely to be more capable than leading commercial frontier models by 2025. The proliferation of open-source models increases accessibility and therefore brings global safety and security implications, especially for models which have the potential to allow malicious use through lack of effective safeguards.

Computation used to train artificial intelligence systems

Computation is measured in total petaFLOP, which is 1015 floating-point operations



Source: Our World in Data

Threat actors

5. The increasing performance, availability and accessibility of generative AI tools allows potentially anyone to pose a threat through malicious use, misuse or mishap. Generative AI will almost certainly continue to lower the barriers to entry for less sophisticated threat actors seeking to conduct previously unattainable attacks. As well as organised groups, political activists and lone actors will likely use generative AI for ideological, political and personal purposes.
6. Criminals are highly likely to adopt generative AI technology at the same rate and pace as the general population, but some innovative groups and individuals will be early adopters. Use of the technology by criminals will highly likely accelerate the frequency and sophistication of scams, fraud, impersonation, ransomware, currency theft, data harvesting, child sexual abuse images and voice cloning. But to 2025, criminals will be less likely to successfully exploit generative AI to create novel malware.
7. To 2025, generative AI has the potential to enhance terrorist capabilities in propaganda, radicalisation, recruitment, funding streams, weapons development and attack planning. But dependence on physical supply chains will almost certainly remain an impediment to the use of generative AI for sophisticated physical attacks.

Safety and Security Risks
Overview

8. Over the next 18 months, generative AI is more likely to amplify existing risks than create new ones. But it will increase sharply the speed and scale of some threats, and introduce some vulnerabilities. The risks fall into at least three overlapping domains:

   - **Digital risks** are assessed to be the most likely and have the highest impact to 2025. Threats include cybercrime and hacking. Generative AI will also improve digital defences to these threats.
   - **Risks to political systems and societies** will increase in likelihood to 2025, becoming as significant as digital risks as generative AI develops and adoption widens. Threats include manipulation and deception of populations.
   - **Physical risks** will likely rise as generative AI becomes embedded into more physical systems, including critical infrastructure and the built environment. If implemented without adequate safety and security controls, AI may introduce new risks of failure and vulnerabilities to attack.

9. These risks will not occur in isolation; they are likely to compound and influence other risks. There will also almost certainly be unanticipated risks, including risks that result from lack of predictability of AI systems.

Safety and Security Risks
Types

10. The most significant risks that could manifest by 2025 include:

Cyber-attacks: Generative AI can be used to create faster paced, more effective and larger scale cyber intrusion via tailored phishing methods or replicating malware. But experiments in vulnerability discovery and evading detection are significantly less mature at this stage. We assess that generative AI is unlikely to fully automate computer hacking by 2025.

Increased digital vulnerabilities: Generative AI integration into critical functions and infrastructure presents a new attack surface through corrupting training data ('data poisoning'), hijacking model output ('prompt injection'), extracting sensitive training data ('model inversion'), misclassifying information ('perturbation') and targeting computing power.

Erosion of trust in information: Generative AI could lead to a pollution of the public information ecosystem with hyper-realistic bots and synthetic media ('deepfakes') influencing societal debate and reflecting pre-existing social biases. This risk includes creating fake news, personalised disinformation, manipulating financial markets and undermining the criminal justice system. By 2026 synthetic media could comprise a large proportion of online content, and risks eroding public trust in government, while increasing polarisation and extremism. Authentication solutions (e.g. 'watermarking') are under development but are currently unreliable, requiring updates as generative AI evolves.

Political and societal influence: Generative AI tools have already been shown capable of persuading humans on political issues and can be used to increase the scale, persuasiveness and frequency of disinformation and misinformation. More generally, generative AI can generate hyper-targeted content with unprecedented scale and sophistication.

Insecure use and misuse: Generative AI integration into critical systems and infrastructure risks data leaks, biased and discriminatory systems or compromised human decision-making through poor information security and opaque algorithm processes (e.g. 'hallucinations'). Inappropriate use by any large-scale organisation could have unintended consequences and result in cascading failures. Generative AI integration into critical functions may also result in over-reliance on supply chains that are opaque, potentially fragile and controlled by a small number of firms.

Weapon instruction: Generative AI can be used to assemble knowledge on physical attacks by non-state violent actors, including for chemical, biological and radiological weapons. Leading generative AI firms are building safeguards against dangerous outputs, but the effectiveness of these safeguards vary. Other barriers to entry will persist (e.g. acquiring components, manufacturing equipment, tacit knowledge), but these barriers have been falling and generative AI could accelerate this trend.

Conclusions

Generative AI has the potential to bring substantial benefits if managed appropriately, accelerating productivity and innovation across many sectors including healthcare, finance and information technology. But there is a risk that inadequate understanding of the technology resulting in disproportionate public anxiety could result in failure to adopt generative AI and put some benefits out of reach.

- Generative AI will also almost certainly act as a force multiplier for safety and security risks by proliferating and enhancing threat actor capabilities and increasing the speed, scale and sophistication of attacks. The aggregate risk is significant.
- Governments will highly likely not have full insight into private sector progress, limiting their ability to mitigate all of the safety and security risks. Additionally, monitoring adoption of AI-based technologies by the broad range of potential threat actors will prove challenging. There is significant potential for technological surprise; there will almost certainly be unanticipated risks.
- The race to develop the best performing generative AI models will almost certainly intensify: experts disagree on whether generative AI is a stepping stone to progress in Artificial General Intelligence. But it will unlock progress in a broad range of domains. To 2025, there is a realistic possibility that generative AI will accelerate development of some of the technologies it converges with, including quantum computing, novel materials, telecommunications and bio-technologies. But increases in risk as a result of these convergences will likely be felt beyond 2025.

Professional Head of Intelligence Assessment

Probability Yardstick