Global Future Council on Artificial General Intelligence

Preparing for Artificial General Intelligence: Global Risks and International Coordination

BRIEFING PAPER

OCTOBER 2025

The **Global Future Council on Artificial General Intelligence (AGI)** seeks to advance awareness, education and dialogue on the evaluation, risk assessment and governance of advanced artificial intelligence (AI) systems, with a specific focus on AGI. Members collaborate on identifying priorities for policy, research and coordination, while shaping global awareness on AGI. The council works towards actionable insights, interacting with key stakeholders and fostering international alignment on both the opportunities and the challenges associated with AGI, so that benefits can be realized responsibly, while addressing open questions and uncertainties, and maintaining public trust. This paper draws on the collective expertise of the council and contains a set of recommendations to guide future efforts.

The council adopts a broad definition of AGI as AI systems that outperform the majority of skilled adults across a wide range of non-physical tasks.[1]

Timelines for AGI

Experts give varying estimates of timelines for AGI development, which continue to generate debate.[2] Some industry leaders believe that AGI systems could be developed in the next 2-10 years,[3] and even sceptical experts think AGI within the next 10-20 years is plausible.[4] These perspectives highlight the speculative and uncertain nature of long-range technological prediction. Rather than a sudden threshold, however, AGI is more likely to emerge through a gradual accumulation of capabilities across different domains, with societal and economic impacts unfolding incrementally over time.

Having made rapid and relatively steady progress recently, AI is now outperforming many humans in some of the most challenging tests of programming, abstract reasoning and scientific reasoning.[5] However, how these advances translate into progress towards AGI remains an open question, in part because there is no consensus on the scales or benchmarks by which such progress should be measured.

Automated AI R&D could further accelerate AI progress. The most advanced AI systems are shifting towards autonomous "agents", capable of completing increasingly complex tasks with less need for human oversight. AI systems that match or exceed human level at software engineering or AI R&D might lead to exponential increases in AI capabilities.

AI companies are already using AI to accelerate their R&D,[6] underscoring the importance of parallel efforts to promote transparency, reproducibility and ethical standards in scientific discovery.

Opportunities and considerations for global preparedness and governance

AGI can be profoundly transformative, with applications across healthcare, education, accessibility, sustainability and scientific discovery. AGI could drive major advances in economic growth, healthcare outcomes and climate solutions, but the scale and pace of its societal impact remain under debate.

Preparing for a potential AGI future is complex, and involves addressing uncertainties around equitable distribution of benefits, workforce transitions and responsible governance. At the same time, AGI carries risks including misuse, such as in acts of terrorism,[7] undue concentration of power and disruption to job markets that could fundamentally challenge the role of human labour in the social contract.

AGI could increase the risk of loss of control – that is, the scenarios where one or more AI systems act against human instructions and come to operate outside of human control, with no clear path to regaining control if their development is not contained.[8] Evidence for this risk is beginning to emerge as part of controlled experiments, with current systems changing their behaviour to avoid modification[9] or replacement by a new AI version,[10] and carrying out undesired actions and lying about them.[11]

Expert opinions on the likelihood of loss of control vary, and there is growing consensus that current safeguards may be insufficient for the scale and complexity ahead. At present, there is no reliable and established way to control AGI-level systems or ensure their alignment with human intentions or values, though significant work is under way on methods

used to detect misaligned objectives, such as chain-of-thought monitoring.[12] These uncertainties reinforce the urgent need to invest in scientific inquiry while also strengthening anticipatory oversight and policy frameworks that can adapt as the technology evolves.

Questions have been raised about how competitive pressures might affect risk management. As highlighted in the International AI Safety Report,[13] strong competitive pressure to develop more capable AI systems can incentivize developers and countries to conduct less thorough risk mitigations. Likewise, policy-makers face the "evidence dilemma", a challenge generated by the pace and uncertainty of AI's advancement, in which proactive interventions may only be catalysed by clear evidence of harm, despite the risk of waiting too long for this evidence to emerge.[14] These dynamics underscore the importance of international collaboration and dialogue to align competition with safety and to ensure that trust-building measures keep pace with technological advances.

Transparency remains a central issue. Despite the transformative implications, AGI development is often opaque to the general public. This lack of visibility prevents stakeholders from detecting whether transformative capabilities, such as autonomous AI R&D, are imminent or already underway. Efforts to improve visibility may benefit from continued dialogue and trust-building among stakeholders, supported by measures such as transparent reporting, shared benchmarks and independent evaluations.

Recommendations

Mitigating risks on the path to AGI requires action from across the ecosystem and is essential to unlock AGI's enormous potential. The council therefore suggests the following guiding principles.

International collaboration is crucial, requiring different actors to find common ground in averting potentially severe harms. This could include:

- Establishing a high-level dialogue for coordination on malicious use and safety challenges.
- Jointly exploring verification mechanisms – technical procedures to support confidence in claims about an AI system or related resources.
- Establishing international protocols and sharing best practices for safe and secure development and deployment.

To ensure the safety and security of AI in their jurisdiction, governments could consider:

- Developing the expertise and technical tools required to engage with evolving safety research.
- Exploring shared best practices and establishing safety and security standards for the most advanced systems.
- Facilitating dialogue on how to improve transparency and accountability around advanced AI development, and conducting evidence-based assessment of the impact of AGI on the economy and on society. This could include adapting existing frameworks to AGI.

Frontier developers should prioritize the safety, security and reliability of their most advanced systems. Many developers have made strong progress in setting out frameworks for how they evaluate and mitigate severe AI risks. Further best practices for developers to consider include:

- Adding internal deployments of current frontier systems in safety frameworks. Evaluating safety methodologies and mitigation for systems before internal deployments, particularly for models evading control measures or covertly pursuing misaligned goals. Defining criteria for safeguards requirements, including internal access and usage restrictions.
- Dedicating a proportion of the overall R&D budget and compute to developing a robust safety case that addresses leading catastrophic risks (e.g. as listed in the International AI Safety Report and subject to review by independent and recognized external experts.
- Ensuring government awareness of the rate of AI R&D and any safety issues, including the disclosure of relevant evaluation results, incidents and major changes.

AI adopters should request robust and verifiable assurances on safety and reliability. Companies and institutions that procure AI systems have an important role in shaping the development and deployment of AGI. They could:

- Require robust assurances for the safety and security of any system they are procuring.
- Implement reliable monitoring and containment mechanisms for AI systems, particularly for agents. This could include defining clear guardrails for expected behaviour and detecting undesired or unauthorized actions.

Compute providers could support monitoring of AI activities. This includes robust know-your-customer checks for large-scale compute use.

Contributors

Global Future Council on Artificial General Intelligence 2025-2026

The World Economic Forum's Network of Global Future Councils is the world's foremost multistakeholder and interdisciplinary knowledge network dedicated to promoting innovative thinking to shape a more resilient, inclusive and sustainable future.

Disclaimer: This document was developed collaboratively by members of the Global Future Council on AGI. It reflects areas of broad agreement among participating members but should not be interpreted as a statement of unanimous consensus. Participation in scientific fora and endorsement of the technical judgments in this statement reflect the value of expert engagement. However, signing this statement should not be taken as an endorsement by the organizations or institutions with which the signatories are affiliated, nor should it imply institutional agreement with any policy positions expressed herein.

Global Future Council on AGI members

Elizabeth (Beth) Barnes
Founder & CEO, METR

Yoshua Bengio
Full Professor, University of Montreal

Dawn Bloxwich
Senior Director, Responsible Development & Innovation, Google DeepMind

Dan Hendrycks
Executive Director, Center for AI Safety

Seunghoon Hong
Professor, Korea Advanced Institute of Science and Technology (KAIST)

Atoosa Kasirzadeh
Assistant Professor, Carnegie Mellon University

Hiroaki Kitano
Chief Technology Fellow, Sony Group

Wan Sie Lee
Director, Infocomm Media Development Authority (IMDA)

Akiko Murakami
Executive Director, Japan AI Safety Institute

Sella Nevo
Director, Meselson Center, RAND

Dawn Song
Professor, University of California, Berkeley

Jaan Tallinn
Founder, Centre for the Study of Existential Risk

Max Tegmark
Professor of Physics, Massachusetts Institute of Technology (MIT)

World Economic Forum

Council Manager
Benjamin Cedric Larsen
Initiatives Lead, AI Safety, Centre for AI Excellence

Acknowledgements

Audrey Duet
Head, Data and AI Innovation, Centre for AI Excellence, World Economic Forum

Cathy Li
Head, Centre for AI Excellence; Member of the Executive Committee, World Economic Forum

Endnotes

1. This roughly falls between “Competent AGI” and “Expert AGI” in the framework on levels of AGI proposed by Google DeepMind. See Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., ... & Legg, S. (2023). Levels of AGI for Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*. While this broad definition is intended to provide a useful benchmark, achieving AGI will likely not be a single binary event but rather a process of incremental technological advancements and adoption across society, with greater capabilities in some benchmarks compared to others.

2. “AI systems that are better than almost all humans at almost all tasks… [are] quite likely… in the next 2 or 3 years,” Dario Amodei, CEO, Anthropic, told CNBC Television. (2025, January 21). *Anthropic CEO: More confident than ever that we’re ‘very close’ to powerful AI capabilities* [Video]. https://www.youtube.com/watch?v=7LNyUbii0zw). “I would say [we are] probably like 3 to 5 years away [from AGI],” Demis Hassabis, CEO, Google DeepMind told the Big Technology Podcast. (2025, February). *Google DeepMind CEO Demis Hassabis: The path to AGI, deceptive AIs, building a virtual cell*. https://www.youtube.com/watch?v=yr0GiSgUvPU). “I think AGI will probably get developed during this president’s term,” Sam Altman, CEO, OpenAI, said in an interview to Bloomberg. (2025, January 6). *Sam Altman Interview: OpenAI CEO’s plans for ChatGPT, his firing and return, and what’s next*. https://www.bloomberg.com/features/2025-sam-altman-interview/). “Reaching Human-Level AI will take several years if not a decade,” Yann LeCun, Chief AI Scientist, Meta, was quoted as saying on X. (2024, October 16). I said that reaching human-level AI will take several years if not a decade […] [Post]. *X*. https://x.com/ylecun/status/1846574605894340950).

3. Within one year, the aggregate forecast had shortened by 13 years: In 2022, researchers on average estimated a 50% chance of AGI by 2060. In the 2023 survey this forecast had shortened to 2047. See: Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*. Provisional data from the 2024 survey shows an average 50% estimate of 2039.

4. “When AGI does actually come, perhaps 10 or 20 years from now […],” said Gary Marcus, professor emeritus of psychology and neural science at New York University, on X. (2024, December 24). When AGI does actually come, perhaps 10 or 20 years from now […] [Post]. *X*. https://x.com/GaryMarcus/status/1871605871282999760). “I think actual transformative effects (e.g. most cognitive tasks being done by AI) is decades away (80% likely that it is more than 20 years away),” said Arvind Narayanan, professor of computer science at Princeton University and director of the Center for Information Technology Policy. This implies a 20% chance that AI will be doing most cognitive tasks by 2045. See: Toner, H. (2025, September 10). *“Long” timelines to advanced AI have become more common — here’s why*. [Newsletter]. Substack. https://helentoner.substack.com/p/long-timelines-to-advanced-ai-have).

5. International AI Safety Report 2025 (Figure 0.1 and chapters 1.2 and 1.3). See: Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... & Zeng, Y. (2025). International ai safety report. arXiv preprint arXiv:2501.17805. Additionally, a recent analysis by the research organization METR concluded that the lengths of software-related tasks AI can do is doubling every 7 months, and extrapolating this puts human level in 2030. See: Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2025). Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*; METR. (2025, March 19). *Measuring AI ability to complete long tasks* [Blog post]. METR. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks).

6. Companies are already using AI-powered assistants to aid software development. For example, Google DeepMind used the coding agent AlphaEvolve to optimize Google’s computing ecosystem and enhance AI training. See: Google DeepMind. (2025, May 14). *AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms* [Blog post]. Google DeepMind. https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/). Amazon claims to have achieved annual cost savings of $260m through its AI-powered assistant See: Amazon Web Services. (2024, August 1). *Amazon Q Developer just reached a $260 million milestone* [Blog post]. AWS Blogs. https://aws.amazon.com/blogs/devops/amazon-q-developer-just-reached-a-260-million-dollar-milestone/). Microsoft CEO Satya Nadella said that 20-30% of the company’s code was written by AI. See: Mozur, P. (2025, April 29). Microsoft CEO says up to 30% of the company’s code was written by AI. *TechCrunch*. https://techcrunch.com/2025/04/29/microsoft-ceo-says-up-to-30-of-the-companys-code-was-written-by-ai).

7. Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... & Zeng, Y. (2025). International AI Safety Report. arXiv preprint arXiv:2501.17805.

8. Ibid.

9. Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., ... & Hubinger, E. Alignment faking in large language models, 2024. *https://arxiv. org/abs/2412.14093*.

10. Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*; Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. arXiv preprint arXiv:2412.04984; Akin, C. (2024, November 6). Our research on strategic deception presented at the UK’s AI Safety Summit [Blog post]. Apollo Research. https://www.apolloresearch.ai/research/our-research-on-strategic-deception-presented-at-the-uks-ai-safety-summit.

11. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842-845.

12. There is some significant work under way on methods used to detect misaligned objectives, notably on chain-of-thought monitoring, see e.g., Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., ... & Farhi, D. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*.

13. Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... & Zeng, Y. (2025). International AI Safety Report. arXiv preprint arXiv:2501.17805.

14. Ibid.

15. Existing transparency frameworks include the Hiroshima AI Process, which aims to standardize safety and risk mitigation reporting and promotes responsible governance. See: Ministry of Internal Affairs and Communications, Japan. (n.d.). *Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI*. https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html.