Council of Presidents of the United Nations General Assembly
Conseil des Présidents de l'Assemblée Générale des Nations Unies

Governance of the Transition to Artificial General Intelligence (AGI) Urgent Considerations for the UN General Assembly

Report for the Council of Presidents of the United Nations General Assembly (UNCPGA)

Executive Summary

AI systems are rapidly advancing towards artificial general intelligence (AGI), characterized by systems capable of equaling or surpassing human intelligence in diverse cognitive tasks. With the largest financial investments in history driving unprecedented R&D efforts, industry leaders and experts anticipate AGI could emerge within this decade,[1] creating extraordinary benefits to humanity. Among these benefits, AGI could accelerate scientific discoveries related to public health, transform many industries and increase productivity, and contribute to the realization of the Sustainable Development Goals.

Nevertheless, AGI could also create unique and potentially catastrophic risks. Unlike traditional AI, AGI could autonomously execute harmful actions beyond human oversight, resulting in irreversible impacts, threats from advanced weapon systems, and vulnerabilities in critical infrastructures. We must ensure these risks are mitigated if we want to reap the extraordinary benefits of AGI.

To effectively address these global challenges, immediate and coordinated international action supported by the United Nations is essential. Such actions should be initiated by a special UN General Assembly specifically on AGI to discuss the benefits and risks of AGI and potential establishment of a global AGI observatory, certification system for secure and trustworthy AGI, a UN Convention on AGI, and an international AGI agency. Without proactive global management, competition among nations and corporations will accelerate risky AGI development, undermine security protocols, and exacerbate geopolitical tensions. Coordinated international action can prevent these outcomes, promoting secure AGI development and usage, equitable distribution of benefits, and global stability.

I. Introduction

The speed of progress in AI has been rapid in recent years and months[2] and could accelerate even further—in part because AI companies are investing vast sums in making AI agents that are more capable and autonomous, and because of the increasing use of the most powerful AI models to advance AI research itself.[3] It is

[1] METR: Measuring AI Ability to Complete Long Tasks https://arxiv.org/abs/2503.14499

[2] See the International AI Safety Report, Bengio et al 2025.

[3] https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.pdf

widely expected that these improvements in AI capabilities will lead to "Artificial General Intelligence" (AGI): AI systems that match or exceed human performance at most cognitive tasks.

While there is disagreement about when AGI is expected, all the experts on this Panel believe that AGI might well be developed within this decade. AI companies are committing hundreds of billions of dollars to achieve AGI very soon, making this by far the largest R&D effort in human history. The private sector has a responsibility to develop technology that will be much safer, and they should have incentives to do so; but the competitive race to achieve AGI first pushes them to put all their efforts into capabilities, rather than safety, so as to "win the race".

The current risks associated with AI have stemmed primarily from human misuse of technology. However, AGI also presents a fundamentally different risk, as its potential threats extend beyond human-driven misuse. AGI could autonomously generate and execute plans with catastrophic outcomes, surpassing human ability to recognize, analyze, and respond to emerging threats and unprecedented disruptions.[4] Combined with the recently observed self-preservation tendency[5] of advanced AIs this could lead to situations where AGI becomes uncontrollable.

This should be a shared global concern. AGI-related risks are not confined to specific industries or societies but have global implications, regardless of where they originate. Ensuring the safe and harmonious integration of AGI requires not just national or corporate efforts but proactive international governance, spearheaded by the United Nations. The United Nations is uniquely qualified to facilitate a scientific agreement around risks and mitigation strategies, build political consensus around a shared approach to risk mitigation, coordinate policy, promote standards or guardrails, respond to emergencies, and potentially conduct or coordinate joint safety or security research.

Without global governance, the transformative potential of AGI to address global challenges might be underutilized or misdirected. Moreover, global coordination will be essential in managing the global catastrophic threats that AGI is expected to pose. It is difficult to imagine this coordination being achieved at a global level without active leadership from the UN.

[4] "Claude 3.7 (often) Knows When it is in Alignment Evaluations"
https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations

[5] See Meinke et al 2024, Frontier Models are Capable of In-context Scheming, https://arxiv.org/abs/2412.04984.

I. Urgency for UN General Assembly action on AGI governance and likely consequences if no action is taken

Amidst the complex geopolitical environment and in the absence of cohesive and binding international norms, a competitive rush to develop AGI without adequate safety measures is increasing the risk of accidents or misuse, weaponization, and existential failures.[6] Nations and corporations are prioritizing speed over security, undermining national governing frameworks, and making safety protocols secondary to economic or military advantage. Since many forms of AGI from governments and corporations could emerge before the end of this decade, and since establishing national and international governance systems will take years, it is urgent to begin the necessary procedures to prevent the following outcomes:

1. **Irreversible Consequences**—Once AGI is achieved, its impact may be irreversible. With many frontier forms of AI already showing deceptive and self-preservation behavior,[5] and the push towards more autonomous, interacting, self-improving AIs integrated with infrastructures, the impacts and trajectory of AGI can plausibly end up being uncontrollable. If that happens, there may be no way to return to a state of reliable human oversight. Proactive governance is essential to ensure that AGI will not cross our red lines[7], leading to uncontrollable systems with no clear way to return to human control.
2. **Weapons of Mass Destruction**—AGI could enable some states and malicious non-state actors to build chemical, biological, radiological, and nuclear weapons. Moreover, large, AGI-controlled swarms of lethal autonomous weapons could themselves constitute a new category of WMDs.
3. **Critical Infrastructure Vulnerabilities**—Critical national systems (e.g., energy grids, financial systems, transportation networks, communication infrastructure, and healthcare systems) could be subject to powerful cyberattacks launched by or with the aid of AGI. Without national deterrence and international coordination, malicious non-state actors from terrorists to transnational organized crime could conduct attacks at a large scale.
4. **Power Concentration, Global Inequality, and Instability—**Uncontrolled AGI development and usage could exacerbate wealth and power disparities on an unprecedented scale. If AGI remains in the hands of a few nations, corporations, or elite groups, it could entrench economic dominance and create global monopolies over intelligence, innovation, and industrial production. This could lead to massive unemployment, widespread disempowerment affecting legal underpinnings, loss of

[6] OpenAI response to US Office of Science and Technology Policy's AI Action Plan https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf
[7] International Dialogues on AI Safety (2024): https://idais.ai/dialogue/idais-beijing/

privacy, and collapse of trust in institutions, scientific knowledge, and governance. It could undermine democratic institutions through persuasion, manipulation, and AI-generated propaganda, and heighten geopolitical instability in ways that increase systemic vulnerabilities. A lack of coordination could result in conflicts over AGI resources, capabilities, or control, potentially escalating into warfare. AGI will stress existing legal frameworks: many new and complex issues of intellectual property, liability, human rights, and sovereignty could overwhelm domestic and international legal systems.

5. **Existential Risks—**AGI could be misused to create mass harm or developed in ways that are misaligned with human values; it could even act autonomously beyond human oversight, evolving its own objectives according to self-preservation goals already observed in current frontier AIs. AGI might also seek power as a means to ensure it can execute whatever objectives it determines, regardless of human intervention. National governments, leading experts, and the companies developing AGI have all stated that these trends could lead to scenarios in which AGI systems seek to overpower humans. These are not far-fetched science fiction hypotheticals about the distant future—many leading experts consider that these risks could all materialize within this decade, and their precursors are already occurring.[2] Moreover, leading AI developers have no viable proposal so far for preventing these risks with high confidence.
6. **Loss of Extraordinary Future Benefits for All of Humanity—**Properly managed AGI promises improvements in all fields, for all peoples, from personalized medicine, curing cancer, and cell regeneration, to individualized learning systems, ending poverty, addressing climate change, and accelerating scientific discoveries with unimaginable benefits. Ensuring such a magnificent future for all requires global governance, which begins with improved global awareness of both the risks and benefits. The United Nations is critical to this mission.

II. Purpose of UN Governance of the Transition to AGI

Given that AGI might well be developed within this decade, it is both scientifically and ethically imperative that we build robust governance structures to prepare both for the extraordinary benefits and extraordinary risks it could entail.

The purpose of UN governance in the transition to AGI is to ensure that AGI development and usage are aligned with global human values, security, and development. This involves: 1) Advancing AI alignment and control research to identify technical methods for steering and/or controlling increasingly capable AI systems; 2) Providing guidance for the development of AGI—establishing frameworks to ensure AGI is developed responsibly, with robust security measures, transparency, and in alignment with human values; 3) Developing governance frameworks for the deployment and use of AGI—preventing misuse, ensuring equitable access, and maximizing its benefits for humanity while minimizing risks; 4) Fostering future visions of beneficial AGI—new frameworks for social, environmental, and economic development; and 5) Providing a neutral, inclusive platform for international cooperation—setting global standards,

building an international legal framework, and creating incentives for compliance; thereby, fostering trust among nations to guarantee global access to the benefits of AGI.

III. UN General Assembly session on AGI key considerations

One of the biggest challenges of AGI governance is the uncertainty surrounding its future technological development. This makes it difficult to predict potential benefits and risks with precision. Consequently, a broad and comprehensive response framework must be put in place to anticipate and mitigate conceivable threats while reinforcing potential benefits. The United Nations can provide international coordination critical for the development and use of AGI. It is particularly important that all nations be represented in this process and that it reduces geopolitical divides; at present, only the UN appears well-positioned to play this role. The following items should be considered during a UN General Assembly session specifically on AGI:

A. Global AGI Observatory

A Global AGI Observatory is needed to track progress in AGI-relevant research and development and provide early warnings on AI security to Member States. This Observatory should leverage the expertise of other UN efforts such as the Independent International Scientific Panel on AI created by the Global Digital Compact and the UNESCO Readiness Assessment Methodology.

B. International System of Best Practices and Certification for Secure and Trustworthy AGI

Given that AGI might well be developed within this decade, it is both scientifically and ethically imperative that we build robust governance structures to prepare both for the extraordinary benefits and extraordinary risks it could entail.

C. UN Framework Convention on AGI

A Framework Convention on AGI[8] is needed to establish shared objectives and flexible protocols to manage AGI risks and ensure equitable global benefit distribution. It should define clear risk tiers requiring proportionate international action, from standard-setting and licensing regimes to joint research facilities for higher-risk AGI, and red lines or tripwires[9] on AGI development. A Convention

[8] Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo, and Zaheed Kara. 2024. "Framework Convention on Global AI Challenges." Center for International Governance Innovation. https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

[9] Russell, Stuart, Edson Prestes, Mohan Kankanhalli, Jibu Elias, Constanza Gómez Mont, Vilas Dhar, Adrian Weller, Pascale Fung, and Karim Beguir, "AI red lines: The opportunities and challenges of setting limits." World Economic Forum, 11 March 2025. https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/
Karnofsky, Holden. 2024. "A Sketch of Potential Tripwire Capabilities for AI." Carnegie Endowment for International Peace. December 10, 2024. https://carnegieendowment.org/research/2024/12/a-sketch-of-potential-tripwire-capabilities-for-ai?lang=en.

would provide the adaptable institutional foundation essential for globally legitimate, inclusive, and effective AGI governance, minimizing global risks and maximizing global prosperity from AGI.

D. Feasibility Study on a UN AGI Agency

Given the breadth of measures required to prepare for AGI and the urgency of the issue, steps are needed to investigate the feasibility of a UN agency on AGI, ideally in an expedited process. Something like the IAEA has been suggested, understanding that AGI governance is far more complex than nuclear energy; and hence, requiring unique considerations in such a feasibility study.

IV. These recommendations contribute to the implementation of the UN Pact for the Future and other UN initiatives

Multiple UN initiatives call for the development of safe, secure and trustworthy AI. Among these, UN General Assembly resolutions on AI–A/78/L.49, A/78/L.86 and A/C.1/79/L.43–along with the UN Pact for the Future, the Global Digital Compact, and UNESCO's Recommendation on the Ethics of AI call for international cooperation to develop beneficial AI for all of humanity, while proactively managing the global risks.

These initiatives have brought world attention to current forms of AI. This report builds on these UN initiatives by specifically addressing the development of AGI in the near future.

The commitments made by the Pact for the Future are advanced in several ways in this report. A session of the UN General Assembly focused on AGI responds to the Pact for the Future's commitment to global dialogue on AI governance. This report's recommendations on a UN Framework Convention on AGI and a feasibility study for a UN AGI agency. The Observatory we have put forward would support the work of the forthcoming Independent International Scientific Panel on AI, one of the key outcomes of the Global Digital Compact. Finally, the International System of Best Practices and Certification for Secure and Trustworthy AGI would contribute to trust and transparency as called for by the UN General Assembly Resolutions, UNESCO, and the Pact for the Future.

V. Conclusion

Increasing the awareness of national and international leaders concerning the benefits and risks of future AGI—as distinct from current forms of AI—is urgently needed. The UN General Assembly is the proper venue to initiate such a global discussion.

International coordination of the development and use of AGI will be required to reap the extraordinary benefits of AGI while safeguarding human rights and security. It is this AGI Panel's firm recommendation that the UN General Assembly act urgently to

address these issues during a General Assembly session specifically for a global governance framework for AGI. Without such action, the risks of uncontrolled AGI development and use—ranging from dramatically increased global inequality to existential threats—are immense. This UN-led approach, involving a global observatory, international certification, an AGI UN Convention, and a dedicated AGI agency, increases the likelihood that AGI is developed and used in ways that benefit all of humanity while minimizing risks. This framework must be inclusive, transparent, and enforceable to foster trust and cooperation among nations.

Appendix

Terms of Reference: High-Level Panel on Artificial General Intelligence (AGI) for the Council of Presidents of the United Nations General Assembly (UNCPGA)

Context

The Seoul Declaration 2024 of the UNCPGA calls for a panel of artificial general intelligence (AGI) experts to provide a framework and guidelines for the UN General Assembly to consider in addressing the urgent issues of the transition to artificial general intelligence (AGI).

This work should build on and avoid duplicating the extensive efforts on AI values and principles by UNESCO, OECD, G20, G7, Global Partnership on AI, and Bletchley Declaration, and the recommendations of the UN Secretary-General's High-Level Advisory Body on AI, UN Global Digital Compact, the International Network of AI Safety Institutes, European Council's Framework Convention on AI and the two UN General Assembly Resolutions on AI. These have focused more on narrower forms of AI. There is currently a lack of similar attention to AGI.

AI is well known to the world today and often used but AGI is not and does not exist yet. Many AGI experts believe it could be achieved within 1-5 years and eventually could evolve into an artificial super intelligence beyond our control. There is no universally accepted definition of AGI, but most AGI experts agree it would be a general-purpose AI that can learn, edit its code, and act autonomously to address many novel problems with novel solutions similar to or beyond human abilities. Current AI does not have these capabilities, but the trajectory of technical advances clearly points in that direction.

The UN Global Digital Compact calls for a Global Dialogue on AI governance within the United Nations. AGI private sector experts have stressed the urgent need for a global conversation to better understand the opportunities and risks of AGI. A UN General Assembly Special Session on AGI is likely the fastest, most cost-effective, and shortest time-to-impact way to stimulate such a conversation.

Purpose

In response to the Seoul Declaration 2024 of the UNCPGA, produce an initial report for the UNCPGA Chairman and its Members for the 8-10 April 2025 UNCPGA meeting in Bratislava.

The report should identify the risks, threats, and opportunities of AGI. It should focus on raising awareness of mobilizing the UN General Assembly to address AGI governance in a more systematic manner. It is to focus on AGI that has not yet been achieved, rather than current forms of more narrow AI systems. It should stress the urgency of addressing AGI issues as soon as possible considering the rapid developments of AGI, which may present serious risks to humanity as well as extraordinary benefits to humanity.

The report should also include both multi-lateral arrangements and private sector actions to address these unprecedented challenges. It should respond to the private sector AGI leaders calling for international coordination and multi-lateral action to what could be the most difficult management challenge humanity has ever faced.

Procedures

- Convene a high-level panel (5–8 members) of international AGI experts on the potential threats of AGI to humanity and the opportunities AGI could benefit humanity and related policy issues.
- The AGI panel will meet virtually on a regular basis starting in January 2025 and complete the initial report for the next UNCPGA meeting in Bratislava in Spring of 2025.
- Based on the feedback on the initial report during the UNCPGA Meeting in Bratislava, the Panel will finalize the report and submit it to the Secretary-General of UNCPGA. And if accepted by the Chairman UNCPGA, then it would be conveyed to the President of the UN General Assembly tentatively by May 1, 2025.

High-Level Independent Panel Members on AGI for the Council of Presidents of the UN General Assembly

Jerome Glenn (USA) Chair

IEEE Organizational Governance of AI Voting Member; author of the European Union's Horizon 2025-27 paper *on* AGI: *Issues and Opportunities*; CEO of The Millennium Project and author of its *International Governance Issues of the Transition from Artificial Narrow Intelligence to* AGI, *Requirements for Global Governance of* AGI, and *Work/Technology* 2050: *Scenarios and Actions*. Author of *Future Mind: Artificial Intelligence* (1989).

Renan Araujo (Brazil)

Research Manager at the Institute for AI Policy and Strategy focusing on risk management related to AGI development. He is currently leading IAPS' work on international AGI governance. He is an Oxford China Policy Lab Fellow, lawyer, co-founder of the Condor Initiative (which connects Brazilian students with world-class opportunities to shape AI research and policy) and worked on AI governance programs at Rethink Priorities and the Institute for Law and AI.

Yoshua Bengio (Canada)

Professor of computer science at Université de Montréal; Chair, Safety and Secure AI Advisory Group for the Canadian government; Chair of the International AI Safety Report mandated by 30 countries plus UN, OECD and EU. Scientific director of Mila, the Quebec AI Institute; Member of the UN Secretary-General's Scientific Advisory Board for Breakthroughs in Science and Technology; recipient of the Turing Award and currently the most cited computer scientist worldwide.

Joon Ho Kwak (Republic of Korea)

Technical advisor of the Korean AI Safety Institute; played a leading role in the development of the OECD's Guidelines for Developing Trustworthy AI; participant in the G7 Hiroshima Process, Paris AI Action Summit preparations, Korea-US AI Working Group, and member of the Korean delegation to the International AI Safety Institutes Networks.

Lan Xue (China)

Chair of the National Expert Committee on AI Governance; Dean of the Institute for AI International Governance at Tsinghua University; member of the Advisory Group of STI Directorate of the OECD; advisor for the China AI Safety Institute; Co-Chair of the Leadership Council of the UN Sustainable Development Solution Network (UNSDSN); recipient of the Fudan Distinguished Contribution Award for Management Science and the Distinguished Contribution Award from the Chinese Association of Science of Science and S&T Policy.

Stuart Russell (UK and USA)

Distinguished Professor of Computer Science and Director, Center for Human-Compatible AI, University of California, Berkeley; author of Artificial Intelligence: A Modern Approach, the standard AI textbook used in 1,500 universities across 135 countries and cited over 74,000 times; Co-Chair of the OECD expert group on AI futures and the World Economic Forum's Global AI Council.

Jaan Tallinn (Estonia)

Member of the UN AI Advisory Body; served on the EC's High-Level Expert Group on AI; Co-Founder of the University of Cambridge's Center for the Study of Existential Risk and the Future of Life Institute (both institutions are leaders in AGI issues); Board Member of the Center for AI Safety; Estonian investor in AGI safety; Founding engineer of Skype and FastTrack/Kazaa; and a founding investor director of DeepMind Google.

Mariana Todorova (Bulgaria)

Bulgarian representative in UNESCO's Intergovernmental Group on AI Ethical Frameworks; leading spokesperson on AGI in Bulgarian media; internationally recognized author and lecturer on AI's and AGI's ethical and technological dimensions; former Member of Parliament and advisor to the President of the Republic of Bulgaria.

José Jaime Villalobos (Costa Rica)

Multilateral Governance Lead at the Future of Life Institute; Senior Research Associate, Centre for International Governance Innovation; Research Affiliate, Oxford Martin AI Governance Initiative; Research Affiliate, Institute for Law & AI; PhD in international law; and is co-author of leading books and articles on international AI governance.