



FLI Interim Recommendations for the AI Action Summit

France, 10-11 February 2025

Imane Bello (Ima)
AI Action Summit Lead
ima@futureoflife.org

future
of life
INSTITUTE

Contents

Introduction	2
Analysis of the progress made since Bletchley	4
Interim recommendations for the Summit programme	5
Science	5
Solutions	5
Standards	6
Interim recommendations in advance of the Summit	6
Recommendations for all participating governments	6
Recommendations for the French hosts	7
Recommendations for the People's Republic of China	7
Recommendations for the EU AI Office	8
Recommendations for the United States	8
Recommendations for countries with emerging technological infrastructure	8
Recommendations for AI Safety Institutes or Networks	8
Endnotes	9

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

View this document online: futureoflife.org/document/summit-2025

Cover image: Palais de l'Élysée seen from the gardens, Paris. European Heritage Days 2014. Wikimedia Commons

Published: 26 November 2024 | By the Future of Life Institute

Introduction

Your Excellent, President Macron,
Special Envoy Bouverot,

We would like to thank you for your personal leadership in convening the world's third AI Summit and for the clear focus on Action. Building on the important milestones reached at Bletchley and Seoul, this Summit will help ensure that the development and deployment of AI benefits the common good.

The Future of Life Institute (FLI) is an independent non-profit organisation that aims to steer transformative technology towards benefiting life and away from extreme large-scale risks. Back in 2017, FLI organised a conference in Asilomar, California to formulate one of the earliest artificial intelligence (AI) governance instruments: the "Asilomar AI principles".ⁱ The organisation has since become one of the leading voices in AI policy in Washington D.C. and Brussels, and serves as the the civil society champion for AI under the United Nations Secretary General's Digital Cooperation Roadmap. Our staff also actively participate in the OECD's Expert Groups.

At the past AI Summits, FLI supported the organisers by providing essential information on the latest AI breakthroughs, as well as with recommendations on how to advance AI safety. FLI is committed to have the same level of engagement with the Summit held in France. Our latest work consists of research, for example on mechanistic interpretability,ⁱⁱ on socio-technical concerns and mitigationsⁱⁱⁱ and on effective risk mitigation measures for general-purpose AI systems,^{iv} as well as of grant making to academics working on the reduction of power concentration in the AI industry.^v We also educate the general public through our newsletters,^{vi} by hosting monthly Paris AI Safety Breakfasts^{vii} and in co-hosting the recent AI Safety Symposium at Sorbonne University.^{viii}

i Future of Life Institute. (2017, 8 11). Asilomar AI Principles. Retrieved 11 25, 2024, from <https://futureoflife.org/open-letter/ai-principles/>

ii Baek, D., Li, Y., & Tegmark, M. (2024). [Generalization from Starvation: Hints of Universality in LLM Knowledge Graph Learning](https://arxiv.org/abs/2410.08255). 10.48550/ARXIV.2410.08255.

iii Aguirre, A., Dempsey, G., Surden, H., & Reiner, P. (2020). [AI loyalty: A New Paradigm for Aligning Stakeholder Interests](https://ssrn.com/abstract=3560653). SSRN Electronic Journal. 10.2139/ssrn.3560653.

iv Uuk, R., Brouwer, A., Dreksler, N., Pulignano, V., & Bommasani, R. (2024). [Effective Mitigations for Systemic Risks from General-Purpose AI](https://arxiv.org/abs/2410.08255). SSRN Preprint. SSRN. Retrieved 11 25, 2024.

v Future of Life Institute. (n.d.). How to mitigate AI-driven power concentration. Retrieved 11 25, 2024, from <https://futureoflife.org/grant-program/mitigate-ai-driven-power-concentration/>

vi Bello, I. (n.d.). AI Action Summit Newsletter | Ima Bello | Substack. Retrieved 11 25, 2024, from <https://aiactionsummit.substack.com/>

vii Future of Life Institute. (n.d.). Petits-déjeuners sur la sécurité de l'IA. Retrieved 11 25, 2024, from <https://futureoflife.org/fr/petits-dejeuners-sur-la-securite-a-l-ai/>

viii Sorbonne Université. (n.d.). AI Safety Symposium. Retrieved 11 25, 2024, from <https://sites.google.com/view/paiss2024/home>

In our view, the Summit should achieve three main objectives, which align with the three categories previously articulated by President Macron:^{ix}

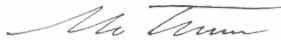
- **Science.** The Summit should facilitate the sharing of evidence to build a common understanding around the severity of global AI risks, chiefly loss of control risk, and the tools to reduce it;
- **Solutions.** The Summit should establish a framework for effective international AI governance that is led by the public sector and complements existing initiatives, while maintaining the Summit's niche as the AI forum for all major governments (including China) to discuss international AI Safety issues;
- **Standards.** The Summit should secure strong political commitment to implement new AI safety standards globally, particularly with regards to increasingly powerful general purpose AI systems. In doing so, France should build on its unique track record of formulating global standards, dating back to the International System of Units.^x

Ultimately, we hope that the Summit will foster the development of an agile and effective international architecture for AI governance, ushering in an era of unprecedented innovation and prosperity. We wish you good luck with the preparations for the summit and stand ready to offer our expertise in support of effective global AI governance.

Sincerely,



Professor Anthony Aguirre
Executive Director



Professor Max Tegmark
President

ix Élysée. (2024, 5 22). Rassemblement des plus grands talents français de l'IA. Retrieved 11 25, 2024, from <https://www.elysee.fr/emmanuel-macron/2024/05/22/rassemblement-des-plus-grands-talents-francais-de-lia>

x Various authors. (n.d.). Système décimal. Wikipédia. Retrieved 11 24, 2024, from https://fr.wikipedia.org/wiki/Syst%C3%A8me_d%C3%A9cimal

Analysis of the progress made since Bletchley

FLI was selected as a key civil society participant at the inaugural AI Safety Summit,¹ held at Bletchley Park in November 2023. At the Summit, we made a number of recommendations, namely to establish a common understanding of the severity and urgency of AI risks, to make the global nature of the AI challenge explicit and to emphasise the need for a unified global response.

The Bletchley Park legacy includes a plan for safety testing of frontier AI models, whereby the UK government plays a role in pre- and post-deployment model testing.² They explicitly test for ways in which advanced AI can undermine critical national security and safety, or cause societal harms.

Since then, we have seen meaningful steps forward in AI governance at the national and international levels.

During the Seoul Summit, held in May 2024, numerous countries called for the furthering of evidence-based scientific knowledge on AI safety. The International Scientific Report on the Safety of Advanced AI,³ in its interim version, was published and presented at this Summit. It covered malicious risks (biological weapons, deepfakes, scams), malfunction (product safety issues, bias, loss of control) and systemic risks. Frontier AI Safety Commitments⁴ were also established and 16 companies from the Middle East, Asia, Europe and the Americas agreed to them. Moreover and during the Seoul Summit, the U.S Department of Commerce released its plan for global collaboration among certain AI Safety Institutes.⁵

In August 2024, the landmark EU AI Act⁶ entered into force. Crucially, it included the regulation of general-purpose AI systems, thanks in large part to the advocacy of FLI⁷ and other civil society partners. The entry into force also kicked off the process to draft a Code of Practice for the most advanced, general-purpose AI systems and which is now ongoing.

Subsequently at the UN in September 2024, the High-level Advisory Body on AI released its final report "Governing AI for Humanity";⁸ immediately followed by the historic adoption of the Global Digital Compact at the Summit of the Future. Lastly, in November of 2024, the inaugural convening⁹ of some AI Safety Institutes took place in San Francisco. At the gathering, Network members agreed on four key principles in their mission statement,¹⁰ namely research, testing, guidance and information and tool-sharing. The meeting shows how countries are increasingly creating public sector bodies to oversee AI development and starting to coordinate at a technical level.

Interim recommendations for the Summit programme

The French government has created five tracks¹¹ for the AI Action Summit, which clearly demonstrates the high-level of ambition for the event. We recommend that the final programme builds common understanding around the severity of global AI risks and confirms the leading role of the public sector in shaping AI governance (relative to private companies).

Science

To build a common scientific understanding of AI risk, FLI recommends adding three specific elements to the programme. Firstly, current efforts to develop general-purpose autonomous AI - systems that can act, plan, and pursue goals - could lead to a loss of control.^{See 3} Given this critical risk of AI development, the programme would benefit from a session where independent experts discuss the likelihood of this risk and potential mitigation measures.

Secondly, the French government may want to include a live demonstration of specific dangerous AI capabilities (such as persuasion capabilities within election contexts). A live demo of this kind will ensure the same level of understanding amongst participants and serves as a good introduction to the International Scientific Report on the Safety of Advanced AI^{See 3} and can build on the demo carried out in San Francisco in November 2024.¹²

Thirdly, it may be worth inviting two AI Safety Institutes from different countries, such as the Japanese and UK AI Safety Institutes, to jointly evaluate an advanced AI model for a given risk. When the outcome of this joint evaluation would be presented at the Summit, participants can learn from the commonalities and differences between the approaches of both jurisdictions.

Solutions

President Macron has called for the creation of “shared solutions to address AI’s challenges”. The world can only truly arrive at these shared solutions if both China and the United States are on board. Whereas there is a broad, international understanding of American research, the same cannot be said for Chinese research. Given China’s significant and growing AI capabilities and the global nature of AI risks, lack of engagement could lead to the emergence of unsafe AI systems that would affect the entire world. The Summit should therefore enable deeper exchanges with China,¹³ including by carving out a space in the ministerial session for Chinese stakeholders to share best practices and research areas such as those discussed at the World AI Conference in Shanghai in July 2024¹⁴ and within the Third Plenum Resolution.¹⁵

Standards

As the first country to host an AI Summit after the adoption of the EU AI Act, France has a unique opportunity to help turn this European law into a global standard for AI governance. Under the AI Act, the European Commission is required to draft a so-called “Code of Practice” for general-purpose AI systems. The Commission has already sought to marry the EU and wider international discussion by appointing the lead-author of the Scientific Report, Yoshua Bengio, to pen this Code of Practice. The Summit presents another opportunity to level the playing field by i) ensuring that the Seoul voluntary commitments are incorporated in the EU Code of Practice, ii) encouraging private sector participation and support for the Code, iii) obtaining buy in from non-EU countries to also put the voluntary commitments on a legislative footing.

Beyond the Code of Practice, the Summit and the international community at large should codify the emerging international consensus in setting red lines for autonomous replication or improvement, power seeking, and deception (i.e. factors contributing to loss of control risk).

Interim recommendations in advance of the Summit

Recommendations for all participating governments

- Reassess and update national AI strategies to prepare society for the imminent arrival of agentic AI, e.g. by investing more in hardening critical infrastructure and essential public institutions against automated cyber attacks.
- Involve independent risk management experts in the development of AI policy. Whilst AI represents a nascent public policy challenge, there are many lessons that can be learnt from other high-risk industries such as aviation or biotech.
- Establish clear standards by which national AI Safety Institutes (AISIs) or comparable Networks should audit companies operating within their jurisdiction.
- Implement robust risk mitigation frameworks with clear incentives. Without proactive government action, a lack of trust in AI systems will likely delay widespread AI adoption.
- Establish a clear mission statement for Safety Institutes or comparable Networks tailored to the national context, such as focusing on advancing the science of metrology, the industrial use of highly capable robots, or the utilisation of unique national data sets.

Recommendations for the French hosts

- Balance stakeholders' perspectives by involving the private sector, government officials, academics and civil society, to achieve meaningful actions, mitigate conflicts of interest, and avoid safety washing.
- Ask companies that are signatories to the Frontier AI Safety Commitments to submit their Safety Framework well ahead of the Summit and make these documents publicly available for scrutiny.
- In line with the outcome of the Expert Consultation,¹⁶ persuade the leading companies to commit to independent third-party verification of AI systems.
- Ensure the integration of the Voluntary Commitments and the OECD risk thresholds work-stream within the forthcoming EU Codes of Practice, to avoid global fragmentation.
- Carve out a space in the ministerial session for Chinese stakeholders to share best practices and research areas such as those discussed at the World AI Conference in Shanghai in July 2024^{See 14} and within the Third Plenum Resolution.^{See 15}
- Announce the host(s) of the upcoming summits to set them up for success and ensure a smooth transfer of knowledge. Among multiple candidates, Singapore emerges as a strong candidate to host a future AI Summit, given its relative diplomatic neutrality.
- Establish red lines by codifying the emerging international consensus among leading AI technical and governance experts from China and the West around the following five critical global risks: autonomous replication or improvement; power seeking; assisting weapon development; cyberattacks; and deception.
- Amplify France's global leadership by bringing forward an independent AI Foundation, coupled with a Fund allowing for the creation of small models, based on quality data and for pre-defined and specific applications relevant to public interest AI.

Recommendations for the People's Republic of China

- Inform other governments about Chinese plans to establish AI Safety Oversight Systems (notably through the Third Plenum Resolution^{See 15}).
- Share lessons learned from the TC260 AI Safety Governance framework¹⁷ and China's AI regulations¹⁸ including the algorithm registry and safety/security reviews.
- Present Chinese AI Safety and evaluation groups and share expertise from leading Chinese thinkers.
- Engage in an open dialogue with all the participating countries, including the US. Global threats from advanced AI, much like climate change, require cooperation that transcends geopolitical competition.

Recommendations for the EU AI Office

- Inform other governments about the forthcoming EU Codes of Practice and any key insights that have emerged during the drafting process, with a particular focus on the regime for general-purpose AI systems.
- Align the EU AI Act, and its Code of Practice, with the Seoul Voluntary Commitments to avoid fragmentation of global AI governance.
- Adapt the Code of Practice in line with the latest scientific evidence presented at the Summit.

Recommendations for the United States

- Closely monitor compliance of key AI companies with the voluntary commitments and apply appropriate pressure to ensure compliance.
- Assess the risk that systems produced in the US are used to foster global instability.
- Engage in an open dialogue with all countries, including China. Global threats from advanced AI, much like climate change, require cooperation that transcends geopolitical competition.

Recommendations for countries with emerging technological infrastructure

- Strategically leverage national linguistic and cultural resources for AI safety training and evaluation datasets,¹⁹ ensuring local value creation and cultural representation, while maintaining these resources as proprietary to related communities and closed-source to prevent potential safety compromises and unauthorized commercialisation.
- Invest in local AI safety research capacity and expertise, including indigenous AI safety talent, and negotiate fair, transparent compensation agreements with AI companies.
- For each international AI partnership, mandate a dedicated risk management contact point within both the government and the partnering AI company to ensure a clear accountability mechanism and a contingency plan for potential systemic AI-related disruptions.
- Prioritise cybersecurity capabilities, with a specific focus on protecting critical national systems from increasingly powerful AI-driven cyber threats.

Recommendations for AI Safety Institutes or Networks

- Objectively quantify AI safety²⁰ and risk levels before its widespread adoption.
- Establish a process to share societal threat models and mitigation measures with other Institutes or Networks.
- Produce quarterly scientific reports on findings and best practices to inform the

general public and the writing team of the Scientific Report about insights from the various safety Institutes and Networks.

- Foster public research in AI Safety by creating a global AI safety research fund and a coordinating body to define common research programs and standards.
- Support and conduct public foundational research on systemic risks based on the European taxonomy.²¹

Endnotes

- 1 The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. (2023, 11 1). Retrieved 11 25, 2024, from <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>
- 2 World leaders, top AI companies set out plan for safety testing of frontier as first global AI Safety Summit concludes. (2023, 11 2). Retrieved 11 25, 2024, from <https://www.gov.uk/government/news/world-leaders-top-ai-companies-set-out-plan-for-safety-testing-of-frontier-as-first-global-ai-safety-summit-concludes>
- 3 DSIT, & AI Safety Institute. (2024, 5 17). International Scientific Report on the Safety of Advanced AI. Retrieved 11 25, 2024, from <https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai>
- 4 DSIT. (2024, 5 21). Frontier AI Safety Commitments, AI Seoul Summit 2024. Retrieved 11 25, 2024, from <https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024>
- 5 U.S. Department of Commerce. (2024, 5 21). U.S. Secretary of Commerce Gina Raimondo Releases Strategic Vision on AI Safety, Announces Plan for Global Cooperation Among AI Safety Institutes. Retrieved 11 25, 2024, from <https://www.commerce.gov/news/press-releases/2024/05/us-secretary-commerce-gina-raimondo-releases-strategic-vision-ai-safety>
- 6 Future of Life Institute. (n.d.). EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act. Retrieved 11 25, 2024, from <https://artificialintelligenceact.eu/>
- 7 Future of Life Institute. (n.d.). Strengthening the European AI Act. Retrieved 11 25, 2024, from <https://futureoflife.org/project/eu-ai-act/>
- 8 United Nations. (2024). *Governing AI for Humanity: Final Report*. 9789211067873.
- 9 U.S. Department of Commerce. (2024, 09 18). U.S. Secretary of Commerce Raimondo and U.S. Secretary of State Blinken Announce Inaugural Convening of International Network of AI Safety Institutes in San Francisco. Retrieved 11 25, 2024, from <https://www.commerce.gov/news/press-releases/2024/09/us-secretary-commerce-raimondo-and-us-secretary-state-blinken-announce>
- 10 International Network of AI Safety Institutes. (2024, 11 21). Mission Statement. Retrieved 11 25, 2024, from <https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf>
- 11 Élysée. (n.d.). Sommet pour l'action sur l'Intelligence Artificielle. Retrieved 11 25, 2024, from <https://www.elysee.fr/sommet-pour-l-action-sur-l-ia>
- 12 NIST. (2024, 11 20). FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco. Retrieved 11 25, 2024, from <https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international>
- 13 Élysée. (2024, 5 6). Déclaration conjointe entre la République française et la République populaire de Chine sur l'intelligence artificielle et la gouvernance des enjeux globaux. Retrieved 11 25, 2024, from <https://www.elysee.fr/emmanuel-macron/2024/05/06/declaration-conjointe-entre-la-republique-francaise-et-la-republique-populaire-de-chine-sur-intelligence-artificielle-et-la-gouvernance-des-enjeux-globaux>
- 14 Concordia AI. (2024, 7 19). Concordia AI holds the Frontier AI Safety and Governance Forum at the World AI Conference. Retrieved 11 25, 2024, from <https://aisafetychina.substack.com/p/concordia-ai-holds-the-frontier-ai>
- 15 Concordia AI. (2024, 8 2). What does the Chinese leadership mean by "instituting oversight systems to ensure the safety of AI?" Retrieved 11 25, 2024, from <https://aisafetychina.substack.com/p/what-does-the-chinese-leadership>
- 16 The Future Society. (2024, 11 6). Delivering Solutions: An Interim Report of Expert Insights for France's 2025 AI Action Summit. Retrieved 11 25, 2024, from <https://thefuturesociety.org/consultation-interim-report>
- 17 National Technical Committee 260 (TC260). (2024, 9). AI Safety Governance Framework. Retrieved 11 25, 2024.
- 18 Sheehan, M. (2023, 7 10). China's AI Regulations and How They Get Made. Retrieved 11 25, 2024, from <https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en>
- 19 Various authors. (2024, 10 3). Sécurité de l'IA Multilingue: Une lettre ouverte. Retrieved 11 25, 2024, from <https://securitemultilingue.ai/>
- 20 Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., & Tegmark, M. (2024, 7 8). Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv. <https://doi.org/10.48550/arXiv.2405.06624>.
- 21 European Commission. (2024, 11 14). First Draft of the General-Purpose AI Code of Practice published, written by independent experts | Shaping Europe's digital future. Retrieved 11 25, 2024, from <https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts>



**FLI Interim Recommendations
for the AI Action Summit**

Future of Life Institute

View this document online:

futureoflife.org/document/summit-2025