
Advancing Global Governance for Frontier AI: A Proposal for an AISI-Led Working Group under the AI Safety Summit Series

Bengüsu Özcan
AI Safety Türkiye

With
Johns Hopkins University & Apart Research

Abstract

The rapid development of frontier AI models, capable of transformative societal impacts, has been acknowledged as an urgent governance challenge since the first AI Safety Summit at Bletchley Park in 2023 [1]. The successor summit in Seoul in 2024 marked significant progress, with sixteen leading companies committing to publish safety frameworks by the upcoming AI Action Summit [2]. Despite this progress, existing efforts, such as the EU AI Act [3] and voluntary industry commitments, remain either regional in scope or insufficiently coordinated, lacking the international standards necessary to ensure the universal safety of frontier AI systems.

This policy recommendation addresses these gaps by proposing that the AI Safety Summit series host a working group led by the AI Safety Institutes (AISIs). AISIs provide the technical expertise and resources essential for this endeavor, ensuring that the working group can develop international standard responsible scaling policies for frontier AI models [4]. The group would establish risk thresholds, deployment protocols, and monitoring mechanisms, enabling iterative updates based on advancements in AI safety research and stakeholder feedback.

The Summit series, with its recurring cadence and global participation, is uniquely positioned to foster a truly international governance effort. By inviting all countries to participate, this initiative would ensure equitable representation and broad adoption of harmonized global standards. These efforts would mitigate risks

such as societal disruption, security vulnerabilities, and misuse, while supporting responsible innovation. Implementing this proposal at the 2025 AI Action Summit in Paris would establish a pivotal precedent for globally coordinated AI governance.

Keywords: AI governance, frontier AI, responsible scaling, risk management, international coordination, AI Safety Institutes, global standards

1. Introduction

Frontier AI models present transformative opportunities but also risks that cross national boundaries, such as misuse, unintended societal harms, and exacerbated inequality. Existing regulatory efforts, such as the EU AI Act and voluntary industry commitments, have made progress but lack the global coherence required to govern frontier AI effectively.

The AI Safety Summit series has proven effective in convening diverse stakeholders, advancing voluntary agreements, and emphasizing risk management frameworks. However, the absence of a structured mechanism to sustain and evolve these efforts creates a critical gap. Based on the official statements, The AI Action Summit strives to lead more concrete action for international AI governance [5]. In line with this theme, this proposal invites the summit to address this gap by initiating an AISI-led working group for an effort to be led by the ongoing Summit series to establish international scaling policies for frontier AI. The working group would leverage technical expertise and regulatory insight to create actionable, globally relevant guidelines.

2. Methods

A qualitative review of governance initiatives was conducted, covering a broad range of resources such as AI regulations, risk management frameworks, governance models and responsible scaling policies, which can be mainly grouped as:

1. Analysis of Governance Initiatives: Existing frameworks like the Seoul Ministerial Statement [6] and the Frontier AI Safety Commitments were examined to identify gaps in coordinated global governance.
2. Research on Risk Assessment Frameworks: A thorough analysis was conducted on established AI risk management frameworks, including those outlined by NIST [7], the EU AI Act Code of Conduct for General Purpose AI [8], and the MIT AI Risk database [9]. An analysis of existing responsible scaling policies by the leading AI labs [10][11] is also conducted to better understand the industry practices on this front. This ensured the technical feasibility and practical applicability of the proposed AISI-led working group.

These resources highlighted that there are many existing efforts for developing an efficient safety framework for frontier AI development that are strong in the technical foundation. However, this effort is quite fragmented and lacks a structure to make these universally adaptable. This highlighted the need for a recurring, globally representative mechanism like an AISI-led working group to coordinate responsible scaling policies for frontier AI.

3. Results

The proposed AISI-led working group, hosted under the AI Safety Summit series, would focus on:

1. **Developing Responsible Scaling Guidelines:** Establish clear and universally adaptable responsible scaling policies and safety frameworks for frontier AI, addressing deployment risks and ensuring scalability.
2. **Iterative Policy Updates:** Enable periodic updates to guidelines based on advancements in AI research and risk assessments.
3. **Stakeholder Inclusion:** Ensure participation from diverse stakeholders, including governments, industry leaders, civil society, and academia.
4. **Global Policy Alignment:** Reduce regulatory fragmentation and enhancing cross-border trust in AI governance.

4. Discussion and Conclusion

The establishment of an AISI-led working group under the AI Safety Summit series represents a concrete and vital step for the international coordination of frontier AI governance. Unlike existing efforts, which are either prepared by one single AI developer or regionally constrained, this working group would iteratively update its policies to reflect the latest safety research and would invite all countries to participate under the coordination of the AI Summit series.

To support equitable participation, the working group would implement capacity-building programs and foster strategic partnerships. Capacity-building initiatives could empower nations with limited AI expertise by providing technical support and resources, enabling them to contribute meaningfully. Partnerships with international organizations, such as the UN or OECD, could facilitate the participation of countries without dedicated efforts or institutes by providing them the chance to assign expert representatives to collaborate with.

By leveraging AISIs' technical expertise and the Summit's ability to convene key stakeholders, the working group would develop scaling policies that can be adapted by all frontier AI developers, ensuring universal safety standards. This effort would harmonize global standards, reducing regulatory fragmentation and building international trust, while encouraging responsible innovation by creating predictability for developers and policymakers. This effort can later be supported by global monitoring and verification mechanisms.

Implementing this proposal at the 2025 AI Action Summit in Paris would establish a foundational precedent for collaborative and adaptive AI governance, in line with the key objective of the summit, which is to take concrete action towards international AI governance.

5. References

1. UK Government. (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023. Retrieved 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. UK Government. (2024). Frontier AI Safety Commitments: AI Seoul Summit 2024. Retrieved from <https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024>.
3. Artificial Intelligence Act. (n.d.). *Regulating AI for safety and accountability*. Retrieved from <https://artificialintelligenceact.eu/>.
4. International Center for Future Governance. (2024). *The AI Safety Institute Network: Who, What, and How*. Retrieved from <https://icfg.eu/the-ai-safety-institute-network-who-what-and-how/>.
5. Élysée Palace. (n.d.). *AI Action Summit: Presentation*. Retrieved from <https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia/presentation>.
6. UK Government. (2024). *Seoul Ministerial Statement for Advancing AI Safety, Innovation, and Inclusivity: AI Seoul Summit 2024*. Retrieved from <https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024>.
7. National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework*. Retrieved from <https://www.nist.gov/itl/ai-risk-management-framework>.
8. European Commission. (n.d.). *AI Code of Practice*. Retrieved from <https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice>.
9. MIT AI Risk. (n.d.). *AI Risk Database*. Retrieved from <https://airisk.mit.edu/>.
10. Institute for AI Policy and Strategy (IAPS). (n.d.). *Responsible Scaling of Frontier AI Models*. Retrieved from <https://www.iaps.ai/research/responsible-scaling>.
11. Safer AI Initiative. (2024). *Responsible Scaling Policies Are Risk Management Done Wrong*. Retrieved from <https://www.safer-ai.org/post/rsps-are-risk-management-done-wrong>.

Disclaimer

This document represents a personal effort by the author and reflects their individual thoughts and opinions. It is not affiliated with or representative of the professional work or views of the author’s employer or collaborators.