centre for future generations

Policy Recommendations for the AI Action Summit

10-11 February 2025, Paris, France

The upcoming AI Action Summit in Paris in February 2025 presents an opportunity to advance international AI governance. Building on lessons from other safety-critical industries, these policy recommendations emphasize the need for standardised global risk thresholds for advanced AI systems. We outline four key recommendations for policymakers: establish interim AI risk thresholds, agree on standardised terminology, agree on standardised AI terminology, and commit to creating national AI regulatory bodies to implement international safety standards. By addressing these challenges, countries can ensure safer innovation while mitigating cross-border risks posed by advanced AI technologies.

Table of contents

Context

The French organisers of the third international AI Summit which will take place in February 2025 in Paris aim for this event to initiate more concrete, actionable steps in international AI governance than its predecessors. Fittingly, it is named the AI Action Summit. Given this goal, it is a unique opportunity for attending countries to agree on concrete next steps for addressing large-scale risks from AI by drawing key lessons from other safety-critical industries like civil nuclear technology and aviation.

Since the last international summit on AI in Seoul in May 2024, AI technology has continued to advance rapidly, with the latest models reportedly achieving performance comparable to PhD-level expertise in natural sciences and being deployed as sophisticated coding agents. AI governance has also progressed since Seoul, with the EU setting up its AI Office, which is tasked with enforcing the EU AI Act's provisions on general-purpose AIs and several other countries have founded AI Safety Institutes, or analogous institutions, such as Singapore and France. But many urgent governance challenges remain unaddressed or lack coordination, despite general consensus on their importance.

For instance, in the Seoul Ministerial Statement, signatory countries recognise that they must play a role in establishing "frameworks for managing risks posed by the design, development, deployment and use of commercially or publicly available frontier AI models or systems," and in "identifying thresholds at which the risks posed by the design, development, deployment, and use of frontier AI models or systems would be severe without appropriate mitigations."

At this summit, 16 organisations involved in the development of advanced AI voluntarily agreed to the Frontier AI Safety Commitments, which involve developing and publishing rigorous risk assessments and safety frameworks by the upcoming AI Action Summit. As CFG identified through our comprehensive research, international risk thresholds play an important role in assessing risk levels in other safety-critical industries — and they could do the same for advanced AI without undermining innovation.

Anne Bouverot, the French Special Envoy for AI, emphasised the Summit's top goal is to focus on deliverables and to "move to concrete actions" as opposed to further voluntary commitments. In this spirit, signatories of the Seoul Ministerial Statement and other participants should use the AI Action Summit to agree on concrete actions they can take in the following months to establish internationally compatible AI risk management frameworks, beginning with risk thresholds for advanced AI systems.

Overview

In accordance with the Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity in which signatory countries recognised their role in identifying appropriate AI risk thresholds and AI developers voluntarily committed to adopting safety frameworks, we propose countries attending the AI Action Summit in Paris jointly take first steps towards establishing international, standardised advanced AI risk thresholds.

To work towards this goal, we suggest that governments pursue four points:

- Agree on the need for international, standardised advanced AI risk thresholds.
- Establish interim qualitative risk thresholds, and improve them iteratively over time.
- Agree on standardised AI terminology.
- Agree to create national AI regulatory bodies which implement international AI safety standards.

Policy recommendations

The effects of advanced AI systems, positive or negative, can cross borders; therefore, countries should develop an international approach to mitigating AI risks. Internationally recognised thresholds and safety standards, stipulated via international agreements and guidelines, could lay the groundwork for global AI safety standards. This is a key best practice observed in other safety-critical industries like civilian nuclear technology and aviation.

1. Agree on the need for international, standard advanced AI risk thresholds

The foundation for universal and effective safety assurance of advanced AI is the consensus on its risks and acceptable risk levels. In many other high-risk industries, including civilian nuclear, food, pharmaceutical, and aviation, risk thresholds are defined internationally and adopted either by national regulations or international treaties. The 2025 AI Action Summit in Paris is the ideal venue to kick off the establishment of risk thresholds, something that should involve leading countries in AI development, academia, industry, and civil society participants. The summit agenda should include the establishment of standard risk thresholds for advanced AI, formalised through a post-summit commitment by each participating country to actively contribute to their development and implementation.

Establishing risk thresholds requires niche technical expertise in advanced AI and risk management. Motivated by the earlier AI Safety Summits, many participating countries have AI Safety Institutes (AISIs) collaborating internationally and working on advanced AI safety frameworks. Moreover, 16 advanced AI developers committed to publishing their own voluntary safety frameworks by the AI Action Summit, which serves as a starting point. The summit can coordinate these efforts, establishing a shared research agenda and trusted information-sharing regime across the AISIs and the industry to develop

advanced AI risk thresholds. The summit also brings together experts from academia, CSOs, and organisations like the OECD, which can contribute expertise or chair the collaboration. The previous AI Safety Summits have already demonstrated effective international coordination, highlighted by the International Scientific Report on the Safety of Advanced AI, underscoring the urgency of addressing risks of advanced AI and setting the stage for standardized risk thresholds as a vital next step.

2. Establish interim risk thresholds for advanced AI and commit to iterative updates

The field of AI safety, while advancing, still bears many uncertainties and lacks the maturity of other industries where risk thresholds are established on scientific and experimental evidence. Both The International Scientific Report on the Safety of Advanced AI and the industry-led voluntary safety commitment from the South Korea AI Safety Summit recognize substantial risks that advanced AI systems may pose, especially as these models quickly scale up in capabilities[1]. Given this rapid pace, adopting reasonable, even if imperfect, risk thresholds would protect the public interest and support trustworthy AI development more than waiting for more certainty and proceeding without any safety standards in the meantime.

Existing regulations targeting advanced AI (such as the EU AI Act) and safety frameworks published by advanced AI developers have identified computing power and certain capability levels as risk indicators. These include a model's ability to facilitate bio-synthesis, its ability to replicate itself, and to alter and improve its own code. These capabilities, combined with an upper threshold of computing power used for training, should serve as interim qualitative risk thresholds that signal an advanced AI model may pose particularly significant risks. Where possible, these qualitative thresholds can then be refined into quantitative benchmarks through dedicated expert analysis.

In areas with limited experimental data, leveraging expert consensus to set interim thresholds is a common and pragmatic approach, as demonstrated in fields like radiology and food safety. This approach ensures a precautionary stance until more precise measurements can be established.

It is essential that, as our understanding of advanced AI and the field of AI safety progresses, these risk thresholds are updated to reflect the latest scientific insights and expert consensus and continue to be recognized as international standards. The AI Summit series should adopt a proactive role in this process, making the establishment and periodic revision of risk thresholds a core, ongoing responsibility.

[1] Some existing governance tools include compute thresholds as a proxy for the level of risk an advanced AI model may pose. The EU AI Act uses a threshold of 10^{25} FLOP, while the US Executive Order on AI uses a threshold of 10^{26} FLOP. These precedents can inform the discussion on where a compute threshold may be set.

It is paramount that the AI summit series continue with its commitment to supporting trustworthy AI development and broad participation to achieve this. Over time, this groundwork could transition to a more formal body, potentially evolving from the existing AI Safety Institutes network or an international organization.

3. Agree on standardised AI terminology

As a first step towards internationally compatible standards and governance regimes, the countries attending the AI Action Summit should agree on definitions for key AI governance-relevant concepts, such as evaluation, AI safety, frontier AI, general-purpose AI, alignment, deception, agency, control, containment, and others. Summit participants should task a working group, similar to the group that authored the International Scientific Report on the Safety of Advanced AI, with creating a guidance document listing definitions of relevant concepts. This would provide a valuable shared point of reference, making it possible for the international community to have constructive conversations based on a shared understanding of AI concepts.

4. Agree to create national AI regulatory and oversight bodies which implement international AI safety standards

At the 2025 AI Action Summit, participating countries should commit to creating their own national AI regulatory bodies that are tasked with implementing international guidelines on AI governance — such as the interim risk management regime proposed above — by the end of 2025, and, crucially, ensure compliance through oversight. Beyond agreeing on basic competencies and tasks of national AI regulatory bodies, it should be left to each country to decide what its national body looks like.

In other technologies with potentially harmful impacts that can cross national borders, for example the civilian nuclear sector, international agencies such as the IAEA outline what regulatory institutions should exist at the national level, and which high-level functions they should perform. This can include granting or withdrawing licences, and performing inspections to ensure compliance. However, how countries implement these guidelines and adopt these standards is left open.

Likewise, in AI, international compatibility is important to address cross-border impacts and risks; but countries should also be able to adapt governance systems to local circumstances. Countries can reduce the regulatory burden for AI developers down the line by introducing adequacy decisions, which would allow models that have been declared low-risk in other countries with adequate AI legislation to enter a market without having to undergo further procedures and bi- or multilateral agreements.

Authors

Eva Behrens | Advanced AI Researcher – Policy

Eva focuses on international AI policy and regulation, with a special interest in international coordination for governing the most powerful AI systems. Her research interests include the international governance of advanced AI systems, hardware, and AI risk management policies and regulation.

Bengüsu Özcan | Advanced AI Researcher

Bengüsu works on a broad range of AI governance topics, aiming to bridge technical best practices with effective AI policy. Her research includes scenario planning and international coordination on AI governance.

About CFG

The Centre for Future Generations is an independent think-and-do tank created to help decision-makers anticipate and govern rapid technological change. We are here to make sure that emerging technologies are used in the best interests of humanity. Find out more on www.cfg.eu.