International Center for Future Generations ICFG

Policy Recommendations for the South Korea AI Safety Summit

21-22 May 2024, Republic of Korea

Eva Behrens, David Janků, Bengüsu Özcan, Max Reddel

FOR MORE INFORMATION PLEASE CONTACT:

Eva Behrens
Policy Researcher – Advanced Artificial Intelligence

e.behrens@icfg.eu

The Context

The upcoming AI Safety Summit in South Korea on 21-22 May 2024 presents a rare opportunity to galvanize international efforts on addressing catastrophic risks from advanced AI, a continuation and expansion of the promising steps witnessed at the 2023 UK AI Safety Summit in Bletchley Park.

Recent breakthroughs in AI innovation have only heightened the urgency for closer international cooperation and the swift implementation of safety measures, particularly those addressing the most capable, most high-risk advanced AI systems.

Since the 2023 UK AI Safety Summit, developers have unlocked several AI capabilities deemed impossible before, including advanced video generation by Sora, math olympiad-winning-level problem-solving by AlphaGeometry, and autonomous software engineering by Devin.

In light of this unprecedented proliferation of AI technology, experts have called for stronger policy measures and international cooperation to reduce the heightened catastrophic risks from unchecked, rapidly improving AI. Some experts are particularly concerned with preventing the creation of highly capable AI systems that can devise and execute long-term plans.

Signatories of the 2023 Summit's Bletchley Declaration recognized that advanced AI has the potential of causing "serious, even catastrophic, harm, either deliberate or unintentional" and that "many risks arising from AI are inherently international in nature, and so are best addressed through international cooperation". Furthermore, signatories committed to continuing and deepening international cooperation to "identify, understand and as appropriate act" on these threats. Signatories should use the upcoming AI Safety Summit in South Korea to act on their commitments and agree on urgent next steps towards mitigating catastrophic risks posed by advanced AI through international cooperation.

Overview

In accordance with the shared intent communicated through the Bletchley Declaration to deepen international cooperation where necessary and mitigate catastrophic risks from advanced AI, we urge countries attending the Summit in South Korea to jointly recognise that:

- The **development of so-called long-term planning agents (LTPAs) should be prohibited** until proven safe,
- Advanced **AI models trained on 10^{25} Floating Point Operations (FLOP) of compute capacity or more should be considered high-risk** and need to be regulated accordingly, and
- The **open-sourcing of advanced AI models** trained on 10^{25} FLOP or more should be prohibited.

To build a strong foundation for international cooperation on the governance of high-risk advanced AI, we urge that Summit participants jointly agree to:

- Hold **biannual international AI Safety Summits**, and **pick a host country to follow after France** and
- Keep the focus of the Summits on international collaboration for **mitigating catastrophic risks from advanced AI**.

Steps Towards Safety-First International AI Governance

1. The development of "Long-Term Planning Agents" (LTPAs) should be prohibited until proven safe.

The term "long-term planning agents" (LTPAs) refers to AI systems capable of devising and executing complex and multi-step strategies based on human-assigned reward mechanisms and self-generated internal goals that are not transparent to their human creators.

LTPAs bear significant control risks as they can diverge from human intent and develop their own opaque, potentially harmful goals, a concern mentioned in the Bletchley Declaration. Current testing and evaluation methods for advanced AI cannot clearly document or predict such emergent internal goals that LTPAs might develop.

We recommend that the signatories of the Summit declaration commit to prohibiting the development of LTPAs until and if developers know how to build understandable, transparent, fully controllable LTPAs. This measure would not undermine the potential future benefits of LTPAs but relies on a precautionary framework under which safety measures precede the development of high-risk technology.

2. AI models trained on 10^{25} FLOP or more should be considered high-risk and regulated accordingly.

The Bletchley Declaration underscored a global commitment to mitigate the potentially catastrophic harms of high-risk AI; however, signatories did not yet agree on criteria to distinguish high-risk AI systems, systems that pose catastrophic risks, from other types of AI.

Training highly capable, high-risk AI models requires substantially more compute resources than training other types of AI models. Therefore, compute makes for a quantifiable, measurable lever that can be used for governance tools directed only at the most capable AI models.

We urge summit participants to officially recognize a compute threshold of 10^{25} Floating Point Operations (FLOP) of compute capacity used in training as a classification criterion for high-risk AI models. This threshold builds on the US Executive Act and the EU AI Act, which outline some reporting requirements for training runs above 10^{26} FLOP and 10^{25} FLOP, respectively.

Agreeing on a specific compute threshold ensures a consistent, transparent baseline for distinguishing the small number of high-risk AI models from all other types of AI. This distinction builds a shared international foundation for future regulatory frameworks targeting high-risk AI models without impacting other types of AI.

As algorithms and hardware components improve over time, this threshold must be revised and adjusted downwards periodically to remain effective.

3. The open-sourcing of high-risk AI systems should be prohibited.

Advanced AI development requires significant investment and expertise. Yet, once trained and deployed, the distribution and utilization of such models for harmful purposes can be relatively low-cost if the source code and model weights are openly accessible. Despite the promising beneficial applications, high-risk AI systems can be used for persuasion, generating powerful bioweapons or executing cyberattacks, which makes the distribution of these models a matter of national and global security.

Therefore, we recommend that summit participants commit to prohibiting the open-sourcing of high-risk models, meaning models trained on 10^{25} FLOP or more. A robust and adequate open-sourcing security protocol tailored for the distributed nature of this technology should be a priority of future Summits. Such a limit on open-sourcing would not impede most innovation and scientific research reliant on open-source AI models, as the size and nature of AI models in these contexts typically do not meet the high-risk classification outlined above.

Towards the Summit in France and Beyond

4. Participants should commit to holding biannual AI Safety Summits.

The AI Safety Summit is a valuable forum providing participating governments with the opportunity to exchange knowledge and reach a consensus on international cooperation to mitigate catastrophic risks from advanced AI. Last year's AI Safety Summit already made large strides in galvanizing international attendance and support for international coordination on AI governance. It also initiated the creation of a shared knowledge base on risks from AI development by commissioning the International Scientific Report on Advanced AI Safety.

The community of countries organizing and attending these Summits should build on these successes by committing to holding biannual AI Safety Summits. Such a commitment will create predictability and continuity, reinforcing the forum's visibility and authority. Regularly recurring Summits will also enable more systematic AI progress monitoring and improve international coordination for mitigating catastrophic risks from advanced AI.

Furthermore, the Summits' frequency will create agility and adaptability to changing circumstances in this rapidly evolving field. The current group of participants should also invite other nations to join the Summit series in order to better mitigate unintended distributional effects of advanced AI development.

Lastly, participating countries of the AI Safety Summit in South Korea should agree on and announce the host country of the next Summit, following the upcoming Summit in France.

5. Catastrophic AI Risks should remain the central focus of the biannual AI Safety Summits.

We urge Summit participants to agree that mitigating catastrophic AI risk will remain the central focus of the biannual AI Safety Summits.

Catastrophic AI risks range from large-scale malicious misuse of advanced AI systems or an international AI arms race to unintentional accidents caused by out-of-control advanced AI, all of which experts worry could cause human extinction. These extreme risks are inherently global in nature, and therefore urgently require an international response. A dedicated forum that convenes at regular intervals will make it easier for the international community to discuss and devise such a response.

The UK government created such a forum when it held the first AI Safety Summit in 2023 to discuss the risks posed by the development of frontier AI, and their mitigation through international coordination. Focusing the Summits explicitly on catastrophic AI risks, which are most severe yet only emanate from advanced AI models ensures that the AI Safety Summits continue to fulfill their original purpose, while minimally impacting countries' ability to nationally govern the vast majority of the AI sector.

FOR MORE INFORMATION PLEASE CONTACT:

Eva Behrens
Policy Researcher – Advanced Artificial Intelligence

e.behrens@icfg.eu