UK Government

AI SAFETY SUMMIT

AI SAFETY SUMMIT 1-2 NOVEMBER 2023 - HOSTED BY THE UNITED KINGDOM AT BLETCHLEY PARK

STATEMENT BY THE CHAIR ON 2 NOVEMBER 2023

SAFETY TESTING: STATEMENT OF SESSION OUTCOMES

Meeting at Bletchley Park today, 2 November 2023, and building on the Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023, world leaders representing Australia, Canada, the European Union, France, Germany, Italy, Japan, the Republic of Korea, Singapore, the United States of America, and the United Kingdom; and industry leaders of Amazon Web Services, Anthropic, Google, Google DeepMind, Inflection AI, Meta, Microsoft, Mistral AI, Open AI and xAI have recognised the importance of bringing together governments and actors developing AI within their countries,[1] to collaborate on testing the next generation of Artificial Intelligence (AI) models against a range of critical national security, safety and societal risks.

AI in general, and frontier AI in particular, is a powerful technology that brings great potential for the economy and society.

The participants announced a shared objective of supporting public trust in AI safety, initially through increased emphasis on AI safety testing and research. They noted that comprehensive and effective evaluation processes are a complex technical challenge and collaboration will be important to advance the frontier of knowledge and best practice in this important area. The participants acknowledged:

- To enjoy the potential of frontier AI, as described in the Bletchley Declaration, it is critical that frontier AI is developed safely and that the potential risks of new models are rigorously assessed before and after they are deployed, including by evaluating for potentially harmful capabilities.
- Ensuring the safety of frontier AI systems is a shared responsibility across the AI life cycle and in particular between the actors developing and deploying them. Developers both have responsibility to devise and conduct safety testing through evaluations, transparency, and other appropriate measures, and the technical means of mitigating risks and addressing vulnerabilities. Other actors, including deployers and users, also have responsibility for ensuring the safe use of frontier AI systems.
- Governments have a responsibility for the overall framework for AI in their countries, including in relation to standard setting. Governments recognise their increasing role for seeing that external evaluations are undertaken for frontier AI models developed within their countries in accordance with their locally applicable legal frameworks, working in collaboration with other governments with aligned interests and relevant capabilities as appropriate, and taking into account, where possible, any established international standards. These tests should address the

[1] References to 'governments' and 'countries' include international organisations acting in accordance with their legislative or executive competences.

potentially harmful capabilities of AI models, especially capabilities in certain critical domains, such as national security, safety of the population, and those that may result in significant societal harms, and should be conducted before and after those models are deployed. This is without prejudice to the responsibilities and competences of governments where frontier AI models are deployed to protect their citizens against critical national security, safety and societal risks.

- Governments plan, depending on their circumstances, to invest in public sector capability for testing and other safety research, including advancing the science of evaluating frontier AI models, and to work in partnership with the private sector and other relevant sectors, and other governments as appropriate to this end.
- Governments will plan to collaborate with one another and promote consistent approaches in this effort, and to share the outcomes of these evaluations, where sharing can be done safely, securely and appropriately, with other countries where the frontier AI model will be deployed.

Further, the governments will plan to:

- work together to develop best practices in this emerging area;
- explore arrangements for skills exchange and secondments to further build capability in AI safety;
- develop arrangements for research collaboration to support excellence in developing the science and techniques of AI safety; and
- work towards shared methodologies on testing and develop in due course shared standards in this area.

As its contribution to this collaboration, as chair of the AI Safety Summit on 1-2 November 2023 the United Kingdom (UK) has launched the AI Safety Institute to build public sector capability to conduct safety testing and to research AI safety. The first milestone for the UK AI Safety Institute will be to build its evaluations process in time to assess the next generation of models, including those which will be deployed next year. The UK, through the AI Safety Institute, will partner with other countries to develop their own capability and facilitate collaboration between governments.

The participants to today's session welcomed new and existing initiatives on AI safety by those present, including the creation of the AI Safety Institute. They also highlighted the mutual benefit of the international collaboration that these initiatives will facilitate on external evaluation of frontier AI systems against a range of national security, safety and societal risks before they are deployed.

The participants expressed the shared ambition that, going forwards, they would work collaboratively to put into practice the outcomes of the session. Participants will look to expand their cooperation with further countries, and in particular will, as appropriate, look to support the development of safety testing capabilities with developing country partners.

ENDS