The Bletchley Declaration by Countries Attending the AI Safety Summit 1-2 November 2023

1 November 2023

Artificial Intelligence (AI) presents enormous global opportunities: it has the potential to transform and enhance human wellbeing, peace and prosperity. To realise this, we affirm that, for the good of all, AI should be designed, developed, deployed, and used, in a manner that is safe, in such a way as to be human-centric, trustworthy and responsible. We welcome the international community's efforts so far to cooperate on AI to promote inclusive economic growth, sustainable development and innovation, to protect human rights and fundamental freedoms, and to foster public trust and confidence in AI systems to fully realise their potential.

AI systems are already deployed across many domains of daily life including housing, employment, transport, education, health, accessibility, and justice, and their use is likely to increase. We recognise that this is therefore a unique moment to act and affirm the need for the safe development of AI and for the transformative opportunities of AI to be used for good and for all, in an inclusive manner in our countries and globally. This includes for public services such as health and education, food security, in science, clean energy, biodiversity, and climate, to realise the enjoyment of human rights, and to strengthen efforts towards the achievement of the United Nations Sustainable Development Goals.

Alongside these opportunities, AI also poses significant risks, including in those domains of daily life. To that end, we welcome relevant international efforts to examine and address the potential impact of AI systems in existing fora and other relevant initiatives, and the recognition that the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection needs to be addressed. We also note the potential for unforeseen risks stemming from the capability to manipulate content or generate deceptive content. All of these issues are critically important and we affirm the necessity and urgency of addressing them.

Particular safety risks arise at the 'frontier' of AI, understood as being those highly capable general-purpose AI models, including foundation models, that could perform a wide variety of tasks - as well as relevant specific narrow AI that could exhibit capabilities that cause harm - which match or exceed the capabilities present in today's most advanced models. Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment

with human intent. These issues are in part because those capabilities are not fully understood and are therefore hard to predict. We are especially concerned by such risks in domains such as cybersecurity and biotechnology, as well as where frontier AI systems may amplify risks such as disinformation. There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models. Given the rapid and uncertain rate of change of AI, and in the context of the acceleration of investment in technology, we affirm that deepening our understanding of these potential risks and of actions to address them is especially urgent.

Many risks arising from AI are inherently international in nature, and so are best addressed through international cooperation. We resolve to work together in an inclusive manner to ensure human-centric, trustworthy and responsible AI that is safe, and supports the good of all through existing international fora and other relevant initiatives, to promote cooperation to address the broad range of risks posed by AI. In doing so, we recognise that countries should consider the importance of a pro-innovation and proportionate governance and regulatory approach that maximises the benefits and takes into account the risks associated with AI. This could include making, where appropriate, classifications and categorisations of risk based on national circumstances and applicable legal frameworks. We also note the relevance of cooperation, where appropriate, on approaches such as common principles and codes of conduct. With regard to the specific risks most likely found in relation to frontier AI, we resolve to intensify and sustain our cooperation, and broaden it with further countries, to identify, understand and as appropriate act, through existing international fora and other relevant initiatives, including future international AI Safety Summits.

All actors have a role to play in ensuring the safety of AI: nations, international fora and other initiatives, companies, civil society and academia will need to work together. Noting the importance of inclusive AI and bridging the digital divide, we reaffirm that international collaboration should endeavour to engage and involve a broad range of partners as appropriate, and welcome development-orientated approaches and policies that could help developing countries strengthen AI capacity building and leverage the enabling role of AI to support sustainable growth and address the development gap.

We affirm that, whilst safety must be considered across the AI lifecycle, actors developing frontier AI capabilities, in particular those AI systems which are unusually powerful and potentially harmful, have a particularly strong responsibility for ensuring the safety of these AI systems, including through systems for safety testing, through evaluations, and by other appropriate measures. We encourage all relevant actors to provide context-appropriate transparency and accountability on their plans to measure, monitor and mitigate potentially harmful capabilities and the associated effects that may emerge, in particular to prevent misuse and issues of control, and the amplification of other risks.

In the context of our cooperation, and to inform action at the national and international levels, our agenda for addressing frontier AI risk will focus on:

- identifying AI safety risks of shared concern, building a shared scientific and evidence-based understanding of these risks, and sustaining that understanding as capabilities continue to increase, in the context of a wider global approach to understanding the impact of AI in our societies.
- building respective risk-based policies across our countries to ensure safety in light of such risks, collaborating as appropriate while recognising our approaches may differ based on national circumstances and applicable legal frameworks. This includes, alongside increased transparency by private actors developing frontier AI capabilities, appropriate evaluation metrics, tools for safety testing, and developing relevant public sector capability and scientific research.

In furtherance of this agenda, we resolve to support an internationally inclusive network of scientific research on frontier AI safety that encompasses and complements existing and new multilateral, plurilateral and bilateral collaboration, including through existing international fora and other relevant initiatives, to facilitate the provision of the best science available for policy making and the public good.

In recognition of the transformative positive potential of AI, and as part of ensuring wider international cooperation on AI, we resolve to sustain an inclusive global dialogue that engages existing international fora and other relevant initiatives and contributes in an open manner to broader international discussions, and to continue research on frontier AI safety to ensure that the benefits of the technology can be harnessed responsibly for good and for all. We look forward to meeting again in 2024.

Agreement

The countries represented were:

- Australia
- Brazil
- Canada
- Chile
- China
- European Union
- France

- Germany
- India
- Indonesia
- Ireland
- Israel
- Italy
- Japan
- Kenya
- Kingdom of Saudi Arabia
- Netherlands
- Nigeria
- The Philippines
- Republic of Korea
- Rwanda
- Singapore
- Spain
- Switzerland
- Türkiye
- Ukraine
- United Arab Emirates
- United Kingdom of Great Britain and Northern Ireland
- United States of America

References to ‘governments’ and ‘countries’ include international organisations acting in accordance with their legislative or executive competences.