Chair's Summary of the AI Safety Summit 2023, Bletchley Park

Introduction

A statement by the United Kingdom of Great Britain and Northern Ireland in their capacity as Chair of the inaugural AI Safety Summit which met at Bletchley Park on 1-2 November 2023.

- This Summit, the first of its kind, was convened by the UK to identify next steps for the safe development of frontier AI.
- On 1 November, countries attending agreed to the Bletchley Declaration on AI safety, a landmark agreement recognising a shared consensus on the opportunities and risks of AI, and the need for collaborative action on frontier AI safety.
- They participated in a broad and inclusive discussion, involving representatives from across sectors and, reflecting on the urgent need for a shared international understanding, on 2 November agreed to support the development of an independent and inclusive 'State of the Science' Report, led by the Turing Award-winning scientist Yoshua Bengio.
- A number of countries, together with the companies developing frontier AI, further recognised the importance of bringing together governments and AI developers, and on 2 November agreed to state-led testing of the next generation of models before they are released, including through partnerships with AI Safety Institutes.
- Participants raised a number of more ambitious policies around AI safety and agreed to return to discuss these issues in subsequent discussions in forthcoming AI Safety Summits by the Republic of Korea and France.
- The UK will act to progress the conclusions reached at the Summit.

The AI Safety Summit 2023

Over two days the Summit brought together approximately 150 representatives from across the globe including government leaders and ministers, and industry, academia and civil society leaders. Summit sessions on 1 November allowed for an open, interdisciplinary conversation, which considered the types of risks arising from frontier AI and the role of different actors in responding to them, as well as the significant opportunities of AI across different domains. On 2 November, further discussions focused on the impacts of AI, options for effective collaboration, and how to further the mission of global AI safety. The sessions convened at the Summit, and

the range of conversations held with a broader group of participants beforehand, including those hosted by the Royal Society, TechUK, British Academy, and Alan Turing Institute, allowed for substantive, practical discussion across domestic and international participants.

This Chair's Summary seeks to reflect the discussions held, as well as setting out key considerations which were noted for further action, including priorities for the UK.

A time for action

In his speech on 26 October the UK Prime Minister, Rishi Sunak, framed the Summit discussion. He set out that the world stands at the inflection point of a generational technological revolution, and if we are to seize the benefits of AI we must address the risks. The Prime Minister noted that AI is developing at unprecedented speed, driven by greater access to better chips and more computing power. The capabilities of powerful AI systems will only increase, with profound economic and societal consequences, bringing unprecedented opportunities and risks.

Summit participants agreed that we need to proactively address these impacts if we are to harness this technology's full potential, and that doing so requires collaborative international action. Next year, the next generation of considerably more powerful models will be released and participants identified a narrow window for clear, decisive, and committed action, to engage constructively, globally, and inclusively. They noted that the challenges posed by frontier AI could not be resolved at a single Summit, but that such discussions would set the foundation for realising the ambitions of the Bletchley Declaration and into the next Summits, hosted by the Republic of Korea and France.

A space for discussion

To support a broad and open discussion, the UK published the five objectives of the Summit (4 September[1]), which were addressed across both days.

Objective 1. a shared understanding of the risks posed by frontier AI and the need for action; Objective 2. a forward process for international collaboration on frontier AI safety, including how best to support national and international frameworks.

The Bletchley Declaration agreed an initial mutual understanding of frontier AI, and the risks associated with it, and set out that countries will work in an inclusive manner to ensure human-centric, trustworthy and responsible AI that is safe. It

[1] www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions

committed countries to further collaborate on establishing a shared scientific and evidence-based understanding of the relevant risks.

As an initial step, the UK published a discussion paper on the capabilities of, and risks arising from, frontier AI[2] (25 October), building upon the existing understanding of this matter. Across the Summit, participants exchanged views on the most significant risks and opportunities arising from frontier AI. They recognised that potential harms from misuse, loss of control, and the potential for leaps in capability were particularly pressing.

There was also substantive discussion of the impact of AI upon wider societal issues, and suggestions that such risks may themselves pose an urgent threat to democracy, human rights, and equality. Participants expressed a range of views as to which risks should be prioritised, noting that addressing frontier risks is not mutually exclusive from addressing existing AI risks and harms.

Participants affirmed the importance of continued collaboration and agreed on the urgency of establishing a shared international consensus on the capabilities and risks of frontier AI, which will evolve as the technology develops. Participants noted that, to maintain public trust, future decisions on AI safety must be underpinned by appropriate evidence, and recognised the necessity of fast, flexible and collaborative action by all actors, in particular governments and frontier AI developers, to further understand those risks and ensure effective oversight. To support discussion, the key themes from day one were collated and published on 1 November[3].

All countries in attendance welcomed the UK's initiative to deliver a first-of-its-kind State of the Science Report on frontier AI. Building on the commitment for scientific and evidence-based collaboration as set out in the Bletchley Declaration, the Report will facilitate a shared science-based understanding of the risks and capabilities associated with frontier AI. The UK's Department for Science, Innovation, and Technology has commissioned Yoshua Bengio, a Turing Award-winning AI scientist and member of the United Nations' (UN) Scientific Advisory Body, to Chair the Report's writing group. He will be supported by a diverse group of leading AI academics, advised by an inclusive, international Expert Advisory Panel, with representatives from participating countries.

Objective 3. appropriate measures which individual organisations should take to increase frontier AI safety; Objective 4. areas for potential collaboration on AI safety research, including evaluating model capabilities and the development of new standards to support governance.

[2] www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper
[3] www.gov.uk/government/publications/ai-safety-summit-1-november-roundtable-chairs-summaries

The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023[4], sets out that no single part of society can address the impacts of frontier AI alone and that delivering on the potential of AI requires the sustained attention of governments, businesses, academia, and civil society, with a particularly strong responsibility for actors developing frontier AI capabilities.

During the second set of roundtables on 1 November, participants debated the role of different actors, and affirmed the importance of collaboration and information-sharing. Participants welcomed the UK's Emerging Processes for AI Safety[5] (27 October) and the detailed precedent it established. Participants also discussed the publication by leading frontier AI developers (Amazon, Anthropic, Google DeepMind, Inflection, Meta, Microsoft, OpenAI) of their AI Safety Policies (27 October[6]). They pushed all frontier AI developers to consider how they can build trust through the further development and publication of such policies.

On 2 November, world leaders and their deputies, leading AI developers, and representatives from civil society met and affirmed the need for deeper cooperation. Countries and companies participating agreed on the importance of bringing together the respective responsibilities of governments and frontier AI developers and, in recognition of their existing close partnership, agreed to a plan for safety testing at the frontier, set out in the Safety Testing: Statement of Session Outcomes (2 November 2023).

Participating countries committed, depending on their circumstances, to the development of appropriate state-led evaluation and safety research while participating companies agreed that they would support the next iteration of their models to undergo appropriate independent evaluation and testing.

Across the Summit, including the discussions on 2 November, participants discussed a set of more ambitious policies to be returned to in future sessions:

1. Multiple participants suggested that existing voluntary commitments would need to be put on a legal or regulatory footing in due course. There was agreement about the need to set common international standards for safety, which should be scientifically measurable.
2. It was suggested that there might be certain circumstances in which governments should apply the principle that models must be proven to be

[4] www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration
[5] www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety
[6] www.gov.uk/government/news/leading-frontier-ai-companies-publish-safety-policies

safe before they are deployed, with a presumption that they are otherwise dangerous. This principle could be applied to the current generation of models, or applied when certain capability thresholds were met. This would create certain ‘gates’ that a model had to pass through before it could be deployed.

3. It was suggested that governments should have a role in testing models not just pre- and post-deployment, but earlier in the lifecycle of the model, including early in training runs. There was a discussion about the ability of governments and companies to develop new tools to forecast the capabilities of models before they are trained.
4. The approach to safety should also consider the propensity for accidents and mistakes; governments could set standards relating to how often the machine could be allowed to fail or surprise, measured in an observable and reproducible way.
5. There was a discussion about the need for safety testing not just in the development of models, but in their deployment, since some risks would be contextual. For example, any AI used in critical infrastructure, or equivalent use cases, should have an infallible off-switch.
6. There was a debate about open-source models; these might pose particular risks for safety but might also promote innovation and transparency, including with respect to safety techniques.
7. Several attendees raised the prospect of models being used to interfere with elections in the near future and the need to take action to reduce this risk.
8. Finally, the participants also discussed the question of equity, and the need to make sure that the broadest spectrum was able to benefit from AI and was shielded from its harms.

As an initial contribution to this new collaboration, the UK detailed its launch of the world’s first AI Safety Institute, which will build public sector capability to conduct safety testing and research into AI safety. In exploring all the risks, from social harms including bias and misinformation, through to the most extreme risks of all, including the potential for loss of control, the UK will seek to make the work of the Safety Institute widely available. The UK welcomed commitments from companies in attendance to work with the Institute to allow for pre-deployment testing of their frontier AI models and commitments to work in partnership with other countries’ Institutes including the US.

Objective 5. Showcase how ensuring the safe development of AI will enable AI to be used for good globally.

Through the Bletchley Declaration, participants recognised a shared ambition to unlock the significant potential of frontier AI, which has the ability to transform

economies and societies for the better. They agreed the need for AI to be designed, developed, deployed, and used in a manner which is inclusive, and discussed its potential impact across sectors including healthcare, education and climate change.

Participants welcomed the exchange of ideas and evidence on current and upcoming initiatives, including individual countries' efforts to utilise AI in public service delivery and elsewhere to improve human wellbeing. They also affirmed the need for the benefits of AI to be made widely available, and so welcomed discussion on activity undertaken through international initiatives and fora, such as the UN's AI for Good platform. Many participants set out that for AI to be inclusive, it must also be accessible. Participants discussed a range of measures to that effect and the UK, with Canada, the United States of America, the Bill and Melinda Gates Foundation, and other partners, announced an £80 million collaboration on a new AI for Development collaboration, working with innovators and institutions across Africa to support responsible AI.

A need to go further

The UK is grateful to participants for identifying suggested actions during the Summit, including in the course of multidisciplinary roundtables on day one. There was consensus that more proactive, risk-based, internationally collaborative action is required to build safe frontier AI. A non-exhaustive overview of themes raised is set out below, including a number of points that are a priority for the UK itself, beyond its role as Chair.

The necessity of immediate action to build a shared understanding of frontier AI

One key challenge identified before and during the Summit was the issue of a fragmented and incomplete understanding of frontier AI. Participants agreed that to fully realise the opportunities presented by AI, governments need to take the lead in building public trust, which requires clarity about the technology itself. Digital Ministers were therefore pleased to commit to working together on the State of the Science Report and intend for that work to also provide a shared basis of understanding beyond those present at the Summit. To that end, they noted the potential for complementarity with other processes including the United Nations AI Advisory Body, the Global Partnership on AI (GPAI), and the Organisation for Economic Co-operation and Development (OECD).

Across the Summit, participants welcomed the potential role of the UN AI Advisory Body, in particular its diverse and expert membership, its consideration of analysis and recommendations for the international governance of AI, and its call for contributors to support its work. In addition, the UK and other participants

commended the work of organisations involved in applied research and practical activities such as GPAI, for which they welcomed India's upcoming Chairmanship and hosting of the December Summit. Further discussion also covered contributions of the OECD, including its AI Futures expert group and the OECD AI Policy Observatory, and participants welcomed those activities in contributing to the development of a robust evidence base. Participants shared additional suggestions regarding other relevant initiatives and encouraged discussion between organisations and institutions to ensure complementarity.

The need for an inclusive approach to address frontier AI and other risks

Participants affirmed the importance of inclusivity so that AI may be developed equitably and help bridge the digital and development divides, narrowing rather than widening entrenched inequalities. A number of participants referenced the importance of continued multi-stakeholder collaboration, including governments, businesses, civil society, and academia. Informed by the range of voices across the Summit, many participants agreed that the risks of frontier AI necessitate that we look beyond traditional groupings and work cooperatively with those who may have different views and interests. To that end, participants encouraged discussion in a range of fora, including future AI Safety Summits. Examples of other initiatives mentioned included the G20, the UN and its bodies including the United Nations Educational, Scientific and Cultural Organizations (UNESCO), as well as country or regional-led initiatives including the Republic of Korea's New Digital Order proposal, China's Global AI Governance Initiative, the recently agreed Santiago Declaration, and the African Union's development of a continental AI strategy.

Across both days of the Summit, participants also supported the view that inclusivity should consider the equitable realisation of the benefits of AI, including the importance of breaking down barriers to entry and specific challenges faced by particular groups, such as women and minority groups. Participants raised the challenges faced by developing countries, who may have limited access to the technology stack required to design and develop AI but will still be significantly impacted by its deployment and use. They also discussed the unequal impact where AI is trained on biased or discriminatory data sets which may perpetuate harms. With a focus on development, many participants advised further activity to unlock the potential of 'AI for Good', including via the UN programme of the same name, initiatives such as the AI for Development programme announced by the UK and partners, and the work of philanthropic organisations. Building from these discussions, participants encouraged countries and international initiatives to consider what additional steps may be taken in collaboration, with one another and across sectors, to utilise AI to realise the UN Sustainable Development Goals.

The importance of addressing current AI risks alongside those at the frontier

As the UK Prime Minister set out in his speech on 26 October[7], and as discussed both before and during the Summit, whilst a focus on the frontier is vital, it should be taken forwards alongside action to address immediate AI risks and harms. Participants across both days noted a range of current AI risks and harmful impacts, and reiterated the need for them to be tackled with the same energy, cross-disciplinary expertise, and urgency as risks at the frontier. Concern was raised about the risks of AI spreading false narratives and harming the credibility of individuals, especially where they may threaten electoral processes, as well as AI-enabled misuse in relation to crime, and the dangers of AI increasing inequality and amplifying biases and discrimination.

Views were shared about the most critical of these issues and many participants suggested that each must be tackled simultaneously. Participants also pointed towards work in other international initiatives and fora, recognising the importance of more targeted and bespoke action where appropriate. The UK and other nations welcomed the Council of Europe's work to negotiate the first intergovernmental treaty on AI, with respect to human rights, democracy, and the rule of law, recognising that both the technology and their shared values are global in nature. G7 member countries also noted the project-based work committed to in the G7 Hiroshima AI Process, which includes specific action on disinformation and election integrity, and the cooperative Global Challenge to Build Trust in the Age of Generative AI. Whilst it was not the main subject of discussion on either day, one further risk noted was the use of AI for military purposes. To that end, the UK and other nations welcomed the Summit on Responsible AI in the Military Domain (REAIM) of February 2023, co-hosted by the Netherlands and the Republic of Korea.

The value of appropriate standardisation and interoperability in AI

Participants discussed the benefits of establishing interoperable approaches, often supported by appropriate standardisation and, where suitable, shared principles, codes, or similar frameworks. A number of participants encouraged the development of interoperable frameworks to enable effective risk-based mitigation of frontier AI risks, as well as facilitating the broad and inclusive realisation of the benefits of AI. Building upon this, many participants affirmed that such interoperability does not require complete uniformity of domestic approaches, given that there may be a need for targeted approaches based on national circumstances and applicable legal frameworks. Many participants set forth views on their own domestic frameworks including the UK's anticipated response to the AI Regulation White Paper, the EU AI

[7] www.gov.uk/government/speeches/prime-ministers-speech-on-ai-26-october-2023

Act, the US Voluntary Measures and Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, and China's AI governance framework.

Several participants therefore welcomed efforts to identify the right balance between domestic and international action. The UK and other parties affirmed the vital role of multistakeholder organisations such as the OECD and GPAI, in providing the detailed evidence and policy guidance to enable better, more interoperable AI development, application, and governance. Participants also discussed the important role of global technical AI standards in promoting safe and secure development and adoption of AI. The UK and others recognised the importance of a global digital standards ecosystem which is open, transparent, multi-stakeholder and consensus-based and many standards bodies were noted, including the International Standards Organisation (ISO), International Electrotechnical Commission (IEC), Institute of Electrical and Electronics Engineers (IEEE) and relevant study groups of the International Telecommunication Union (ITU).

One key area of discussion was on the value of common principles and codes. In this regard, several countries welcomed the forthcoming review of the 2019 OECD Recommendation on Artificial Intelligence, which informed the principles agreed by the G20. Countries also recognised the important role of UN bodies and activities, such as the UNESCO Recommendation on the Ethics of AI, which benefits from having the broadest current international applicability. The UK and other countries welcomed action undertaken by the G7 Hiroshima AI Process under the presidency of Japan, and in particular the publication of the Hiroshima Process International Guiding Principles and Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. Participants set out an expectation that the Principles and Code would establish a baseline for developers at the forefront of AI and looked forward to building upon them through further multistakeholder engagement.

The need to develop the broader AI ecosystem including skills and talent

Participants noted that delivering AI safety will require the convergence of multiple branches of activity, including skills, talent, and physical infrastructure. There were a range of views on key priorities, which included ensuring people can access the necessary skills and knowledge both to design, develop, deploy and use AI, and to benefit from the newly created jobs. In addition, there was a discussion on the infrastructure needed, including access to resources such as data and compute. Some participants prioritised such access as amongst the most critical enablers for AI safety and others noted the need to think about how AI could best gain widespread use in an environmentally sustainable way. Participants also discussed

the risks of market concentration and the potential challenges of monopolisation, as well as encouraging further discussion on how best practices could be shared.

A shared challenge

The UK encourages all parties to further consider how they may build upon the dialogue of the Summit, the Bletchley Declaration, this Chair's Summary and the further details shared of the discussions held.

With the frontier of AI constantly moving, the ambitions of the Bletchley Declaration and the Summit discussions cannot be rooted in a single moment. Recognising the need for continued international collaboration, participants committed to meet again in future and welcomed news that the Republic of Korea has agreed to co-host a mini virtual summit on AI in the next six months, with France to then host the next in-person Summit a year from now.

The UK is pleased to have chaired the first AI Safety Summit and thanks all who have contributed to the discussions. The UK is confident that the work undertaken at this Summit will underpin the international response to frontier AI risks and looks forward to progressing this work with partners, and with the next Chair, to unlock AI's transformative potential.