AI Safety Summit 2023

Summary Note

Background

The Summit The AI Safety Summit took place at Bletchley Park on 1-2 November 2023 and brought together international stakeholders from governments, leading AI companies, civil society and academia.

What is AI? AI serves as a simulation of human intelligence via machine, encompassing the capacity of computers and systems to learn, reason, and make decisions by processing data. There are two main types of AI with different associated risks.

- **Generative AI** 'generates' new content through algorithms that can be used to create audio, code, images, text, simulations and videos. ChatGPT is an example.
- **Analytical AI** 'analyses' mass amounts of data and produces analysis of the information collected. Examples include AI used in diagnostic med-tech or self-driving cars.

Why all the fuss? The impact of AI on different sectors of the economy is still unfolding but a number of sectors are likely to be drastically changed, notably transportation, manufacturing, healthcare, education, legal and customer services. We can also expect wider impacts on society including, for example, job displacement (structural labour market change) and more opportunities for data breaches as the involvement of personal data increases as AI evolves.

AI & Midlands Engine

Sarah Windrum, Chair of the Midlands Engine Digital Board: "It is crucial that we understand the differences in risk between the types of AI technology and their application as from there we can ensure we have the appropriate naming conventions and the best methods for testing and regulation to deliver confidence and maximise economic opportunity. It is important that both technology developers and technology users have a greater understanding of AI capabilities and risks. For example, generative AI that influences behaviours and outcomes is not the same level of risk as the safety critical AI that controls the braking system in an autonomous vehicle. We are well positioned as a region to lead on the full spectrum of cross-sector AI technology applications with our domain and technology expertise."

AI is rapidly evolving and impacting various key Midlands sectors such as transportation, manufacturing, healthcare, education, financial and business services. The distribution of the impact of AI depends on characteristics such as education, occupation and industry. As such, some parts of the region face a slightly higher risk of job displacement than others. But it is not all bad news: AI offers opportunities for increased productivity and innovation.

The Midlands can benefit from AI, but consideration must be taken to ensure no groups are left behind. Policy needs to focus on workforce training, planning for AI disruption, fostering collaboration, and increasing public awareness.

Midlands AI Cluster*

- **300+ businesses** (11% UK total)
- **122% business growth** since 2013
- **11,000+ jobs** (8% UK total)
- **31 high growth** companies (7% UK)
- **57 companies** with £100m+ turnover
- **4/25 top UK universities** for Computer Science

**This data came from utilising the DataCity platform and should be taken as an approximation. Moreover, due to the rudimentary nature of data in this sector, the findings should be taken with some caution.*

AI Safety Summit Overview

SoS Michelle Donnellan (DSIT) at Summit: "Our task is as simple as it is profound: to develop AI as a force for good."

The Summit had three key aims:

1. Agree on the risks of AI to inform how we manage them.
2. Discuss how we can collaborate better internationally.
3. Look at how safe AI can be used for good globally.

Below is a summary of the key risks discussed at the Summit alongside mitigations that are being considered.

Key Risks	Mitigations to Consider
Availability to bad actors – as AI systems become more prevalent and accessible, bad actors might use AI to carry out cyberattacks and design chemical/biological weapons.	**Greater collaboration between government, industry and experts**, particularly on testing, to improve regulation and safeguards.
Unpredictable 'leaps' in capability – AI's current abilities are far beyond what experts only recently predicted. As investment increases, we will likely be surprised by the unpredicted and unintended future capabilities of AI.	**More secure and rigorous testing** of new 'frontier' AI models. Promise of potential benefits should not be reason to skip or rush safety testing. More discussion required around whether benefits of 'open-access' AI models like ChatGPT around innovation warrants the risks of such transparency going forward.
Losing Control of AI – although current AI systems are relatively easily controlled and require human prompting, future models will likely gain greater autonomy and could consider unanticipated/unintended actions.	**Greater restrictions/pauses on AI development** in order to enjoy existing AI benefits whilst continuing to understand safety. **Regulate AI decision-making** – what subjects should AI systems NOT be handed over to an AI system.
Negative impact of AI integration into society such as on crime, online safety, election disruption, and exacerbating global inequalities.	**Make use of tools we already have to address these risks** including regulation around privacy, liability and IP as well as incorporating societal metrics into technical evaluations of AI. **However, we must not miss out on opportunities** to use AI to solve global problems like strengthening democracy, climate change etc.
Ensuring AI capability scales responsibly to ensure we don't create unnecessary risks.	**AI companies as a baseline must make progress on AI Safety policies,** including around responsible, risk-informed, capability scaling policies. This is urgent and must be done in months, not years.

What should we do in relation to risks and opportunities of AI?

- National policymakers must address the existing and emerging risks of AI through cross-border collaboration, building capacity across governments, and building rapid, agile and innovative governance.
- Priorities for international collaboration with respect to frontier AI over next 12 months:
    - Develop a shared understanding of frontier AI capabilities and risks to global safety and wellbeing
    - Develop a coordinated approach to safety research and model evaluations of AI systems
    - Develop international collaborations aimed at ensuring the benefits of AI are shared by all
- We need better AI model. We need new architectures, which are engineered to be safe by design. We must learn from safety engineering. We need to add non-removable off switches. Caution is crucial as we have lots of uncertainty. Speed is of the essence. Burden of proof on safety should remain with AI vendors.

UK Government Policy on AI

UK AI National Strategy (2021) aims to continue UK's leadership as a science and AI superpower through long-term investment and planning; support transition to an AI-enabled economy, ensuring AI benefits all sectors and regions; and ensure UK gets the national and international governance of AI technologies right to encourage innovation, investment, and protect the public and our fundamental values.

At the Summit, PM Rishi Sunak unveiled government's new AI Safety Institute chaired by Ian Hogarth. The Institute will carefully test new types of frontier AI before and after they are released to address the potentially harmful capabilities of AI models. In undertaking this research, the AI Safety Institute will look to work closely with the Alan Turing Institute, as the national institute for data science and AI.

Labour Party Reaction to Summit

Peter Kyle, Labour's Shadow Secretary of State for Science, Innovation and Technology called for binding regulation on the big-tech funded companies developing the most powerful 'frontier AI'. Kyle warned that the Prime Minister

“must not hesitate to regulate ” frontier models of AI - the most powerful AI models being developed by a handful of companies such as Google DeepMind, OpenAI and Anthropic.

A Labour government would urgently introduce binding regulation of those companies developing the most powerful ‘frontier' AI. This would include requirements to:

- Report before they train models over a certain capability threshold.
- Conduct safety testing and evaluation on these models, with independent oversight.
- Maintain strong information security protections, to limit the unintended spread of dangerous models.

Peter Kyle unveiled plans to speed up regulatory decision-making for innovation through Labour’s new Regulatory Innovation Office, which would hold regulators accountable by setting target timelines for regulatory decisions.

-The End-