DEMOS

Centre for the Governance of AI

UK AI Strategy Beyond 2024

Background reading for the Demos/GovAI UK AI Strategy Roundtable

Jan 10, 2024

Compiled by Elizabeth Seger

Contents

1. Introduction

Context
This briefing was prepared for the January 10, 2024 roundtable held by GovAI and Demos to reflect on the UK's AI strategy to date, and how we should prepare for 2024 and beyond. The roundtable brought together government leaders, AI experts, and stakeholders from industry and civil society. *High-level roundtable reflections are summarised in Section 6.*

Last year the UK made a mark in the global AI governance landscape with the AI Safety Summit. The next step is for the UK to implement its own concrete AI policies to keep pace with ongoing developments in the US and EU. These policies need to be based on clearly articulated priorities to promote UK national and public interests. These priorities define the UK's AI Strategy.

The Government originally published its National AI Strategy in September 2021. However, much has happened in the last couple years, and especially in the last few months with respect to AI development and governance. As we approach the 2024 elections, it is time to think critically about what we have achieved so far and about what goals, values, and priorities are guiding government efforts on AI. It is a good time to take stock.

The following brief provides an overview of recent AI governance activities in the UK to date.

To make the most of this document, keep the following questions in mind as you read:

- What does it mean for the UK to be a global leader in AI?
- What are we doing well? What, if anything, have we overlooked?
- What policy areas should we prioritise going forward?
- How do we translate all of this into regulatory practice?
- *[Participants of the Ditchley Jan 25-27 conference might also ask: How has the UK been preparing to adapt to the AI revolution to preserve democracy and democratic values and to best serve public interest? What can the UK do better in this respect?]*

2. What is AI?

The purpose of the section is to establish a shared level of technical understanding of AI, and the different varieties thereof, to underpin further conversation.

2.1. AI Definition

OECD Definition of AI System (also adopted in the EU AI Act):

> *An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*

The UK's AI White Paper defines AI by reference to 2 characteristics that are posited to generate the need for bespoke regulatory response.

- Adaptivity: The adaptivity of AI can make it difficult to explain the intent or logic of the system's outcomes… AI systems often develop the ability to perform new forms of inference not directly envisioned by their human programmers.
- Autonomy: The autonomy of AI can make it difficult to assign responsibility for outcomes. Some AI systems can make decisions without the express intent or ongoing control of a human.

The combination of adaptivity and autonomy can make it difficult to explain, predict, or control the outputs of an AI system, or the underlying logic by which they are generated. It can also be challenging to allocate responsibility for the system's operation and outputs.

This characteristic-based definition of AI is intended to age well such that it does not become quickly outdated with the rapid evolution of AI.

2.2. Kinds of AI

The above definitions are very broad, encapsulating a wide variety of machine learning enabled-technologies. However, not all AI is created equal.It is worth distinguishing between different kinds or classes of AI that may therefore require different regulatory approaches.

Size & Capability

In general, larger models are more highly capable. Increases in training compute translate to increases in capability.

Due to high costs of compute and employing top AI talent, only most well resourced actors - mainly major frontier AI labs (e.g. OpenAI, Google DeepMind, Anthropic) - currently have the financial resources necessary to push the frontier foundation model capability. For example, the operating costs alone to run a compute cluster to train a model like GPT-4 is about $1 Billion. In 2023 OpenAI received $10 Billion in funding from Microsoft. In comparison OneTrust, the UK's largest AI company measured by funding, raised $200 million in 2023.[1]

Improvements in compute efficiency and algorithmic efficiency may lower costs such that smaller actors can also develop and run smaller yet still highly-capable models.[2] Similar outcomes might also be achieved through advancements in fine-tuning whereby smaller models are trained on high-quality and more task-specific data sets or on the outputs of more highly-capable models.

Generality

Most of the hype in 2023 has been about general purpose foundation models, sometimes referred to as general-purpose AI or GPAI. These are machine learning models, very often large language models like GPT-4, that demonstrate a base of general capabilities that allow them to be adapted to perform a wide range of downstream tasks. These capabilities can include natural language conversation, behaviour prediction, image analysis, and media generation, which can be used to develop or be directly integrated into other AI systems, products, and models.

Narrow AI

Other AI systems are narrow "expert systems" trained on narrow data sets to execute specific tasks in specific contexts. For example, DeepMinds AlphaFold is a powerful narrow AI system that can predict protein structures which may significantly accelerate drug discovery. Narrow AI systems tend to be smaller and cheaper to train, though still require access to high-quality training data.

Frontier AI

For the purpose of the UK AI Safety Summit government defines Frontier AI as "highly capable general-purpose AI models, like large language models, that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models"

---

[1] https://www.crunchbase.com/organization/onetrust/company_financials
[2] Pilz et al. (2023). 'Increased Compute Efficiency and the Diffusion of AI Capabilities'

Commercial AI

Though 'commercial AI' is not a widely used term, we think it may provide a helpful distinction from frontier foundation models. Frontier foundation model development is cutting edge and only accessible to the most well resourced actors due to massive and increasing model size and compute costs.

Comparatively, we might think of commercial AI as referring to AI development and application activities taking place a good distance behind the frontier. These models are smaller and therefore more easily accessible to less well resourced actors. They also underpin much of the commercial AI development, application, and integration activities happening today as driven by smaller developers and startups.

While concerns around extreme risks posed by AI often pertain to the emergence of new capabilities at the frontier, commercial AI applications are already presenting real world harms. Instances discriminatory algorithmic decision-making, breaches in privacy, and job displacement are on the rise as existing commercial AI systems are applied as productivity boosters, to decide insurance premiums and parole hearings, to make medical diagnoses, and to survey crowds for criminal behaviour.

3. The UK's AI Sector[3]

- **The UK is home to leading academic AI research institutions** in Oxford, Cambridge, London, Bristol, and Edinburgh that draw top AI talent to the UK.

- **The UK's AI start-up industry** is growing rapidly. There are currently over 2,600 AI startups in the UK with just over 1,300 in London alone.[4] These startups mainly focus on the development and/or application of machine learning systems to address target problems and improve productivity.

  Successful UK AI startups are often acquired by US-founded global digital companies as in the case of Google acquisition of DeepMind, Amazon's acquisition of Evi, Microsoft's acquisition of Swiftkey, and Apple's acquisition of VocallQ. When this happens, the companies and their expertise largely stay in the UK and continue to attract talent and investment to the UK. This is encouraging, but cannot be guaranteed as global companies will move assets to where they will be used most effectively.

- **Frontier AI research and development** primarily take place in the offices of mainly US-founded global AI companies with UK based offices (e.g. Google DeepMind, Amazon, Microsoft).

- **The UK Government continues to grow its in-house AI expertise** by onboarding global talent to the AI Safety Institute (AISI).

- **The UK is a hub for AI governance and policy expertise.** These experts are housed within institutions such as DSIT, The Alan Turing Institute, the Centre for the Governance of AI, the Ada Lovelace Institute, the Leverhulme Centre for the Future of Intelligence, the Oxford Internet Institute, the Oxford Martin AI Governance Initiative, the Minderoo Centre for Technology and Democracy, Demos, Chatham House, and others organisations.

[3] Though it is now a bit dated, the 2017 report 'Growing the artificial intelligence Industry in the UK' provides a good overview of the UK's AI sector. One of the more significant updates needed pertains to the UK's AI startup sector which has grown significantly since publication from about 200 companies to over 2,600.

[4] https://thedatacity.com/blog/the-uks-top-artificial-intelligence-ai-hotspots/

4. UK Government on AI: Highlights

This section chronologically highlights activities having taken place within the UK government toward establishing domestic and international AI governance approaches.

4.1. UK Industrial Strategy (2017)

The UK Industrial strategy identified "putting the UK at the forefront of the artificial intelligence and data revolution" as the first of 4 Grand Challenges in which the UK could lead the world. The industrial strategy on AI included support for:

- A "sandbox" approach to developing regulation that stimulates and facilitates innovation;
- Fostering, attracting, developing and retaining leading AI talent;
- The establishment of an AI Council to champion research and innovation, accelerate AI uptake across the economy, and to increase awareness of the advantages of AI and advanced data analytic technologies;
- Taking international leadership in safe and ethical AI development and regulation.

4.2. National AI Strategy (2021)

The National AI Strategy outlines the UK Government's vision for the UK to lead in AI. The Strategy reads, "The UK is a global superpower in AI and is well placed to lead the world over the next decade as a genuine research and innovation powerhouse, a hive of global talent and a progressive regulatory and business environment."

The UK's National AI Strategy is supported by the pursuit of three aims or "pillars":

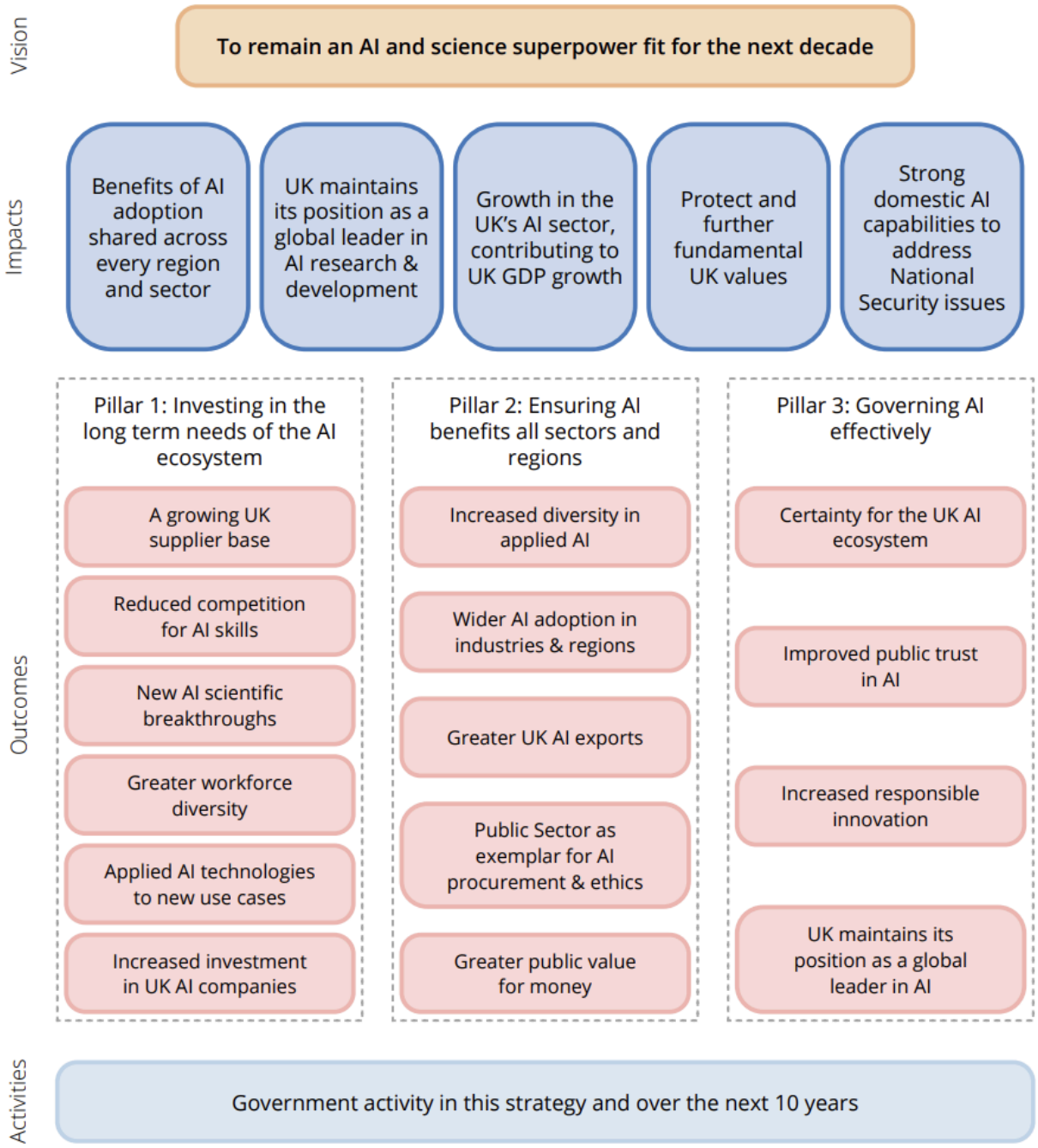
1. Invest and plan for the long-term needs of the AI ecosystem to continue our leadership as a science and AI superpower.

2. Support the transition to an AI-enabled economy, by capturing the benefits of innovation in the UK, and ensuring AI benefits all sectors and regions.

3. Ensure the UK gets the national and international governance of AI technologies right to encourage innovation, investment, and protect the public and our fundamental values.

The National AI Strategy announces the establishment of the Office for AI to oversee the implementation of the National AI Strategy. The Office for AI is now housed within the Department of Science, Innovation, and Technology (DSIT).

The UK's National AI Strategy



4.3. The AI White Paper: A Pro-innovation Approach to AI Regulation (March 2023)

The AI white paper follows on from the National AI Strategy by outlining a regulatory approach that aims to balance the real risk AI can pose against the opportunities and benefits the technology can generate. It is described as a "pro-innovation framework designed to give consumers the confidence to use AI products and services, and to provide businesses the clarity they need to invest in AI and innovate responsibly".

The framework aims to achieve 3 objectives:

1. Drive growth and prosperity by making responsible innovation easier and reducing regulatory uncertainty… "to achieve this objective we must act quickly to remove existing barriers to innovation and prevent the emergence of new ones".

2. Increase public trust in AI by addressing risks and protecting our fundamental values. Trust is a critical driver for AI adoption. If people do not trust AI, they will be reluctant to use it which would reduce demand for AI produce and hinder innovation.

3. Strengthen the UK's position as a global leader in AI. The UK can play a central role in the global conversation by shaping international governance and regulation to maximise opportunities and build trust in the technology, while mitigating potential cross-border risks and protecting our democratic values. There is also an important leadership role for the UK in the development of the global AI assurance industry, including in auditing and safety.

The framework implements 5 principles to guide the responsible development and use of AI in all sectors. These principles have been issued initially on a non-statutory basis to be implemented by individual regulators in the way best tailored to the specific domain.[5] Following an initial period of implementation, and when parliamentary time allows, the plan is for statutory duties on regulators to be introduced that require regulators to have due regard for the principles.

5 Principles:

1. Safety, security and robustness - AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed.

2. Appropriate transparency and explainability - An appropriate level of transparency and explainability will mean that regulators have sufficient information about AI systems and

[5] The CMA has made a standout effort in this regard. See the CMA AI Foundation MOdels: Initial Report

their associated inputs and outputs to give meaningful effect to the other principles (for example, to identify accountability).

3. Fairness - AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes.
4. Accountability and governance - Governance measures should be in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle.
5. Contestability and redress - Where appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm.

The white paper also mandates the establishment of the Central Risk Function tasked with identifying, assessing, prioritising and monitoring AI risks that may require government intervention drawing on expertise across government, industry, and academia. The central risk function operates within the Department for Science, Innovation, and Technology (DSIT).

4.4. House of Commons Interim Report on AI Governance (Aug 2023)

The House of Commons Interim Report on AI Governance responds to the Government's AI white paper.

It welcomes the AI white paper as an initial effort to engage with the complex task of AI governance but raises the concern that the non-statutory, principle-based framework is already falling behind the pace of AI development and regulation as the EU and US take the lead in setting international standards.

The report posits that unless the UK makes a serious and rapid effort to establish the right governance frameworks and to take a leading role in the international initiatives, the governance frameworks laid down by other jurisdictions will likely become the default even if less effective than what the UK could have offered.

The interim report called for a tightly-focused AI Bill to be introduced in the King's Speech to help position the UK as an AI governance leader. (The Speech made only a minor reference to AI)

The interim report also identified Twelve Challenges of AI Governance that policy makers and the frameworks they design ought to meet:

1. The Bias challenge: AI can introduce or perpetuate biases that society finds unacceptable.
2. The Privacy challenge: AI can allow individuals to be identified and personal information about them to be used in ways beyond what the public wants.
3. The Misrepresentation challenge: AI can allow the generation of material that deliberately misrepresents someone's behaviour, opinions or character.
4. The Access to Data challenge: The most powerful AI needs very large datasets, which are held by few organisations.
5. The Access to Compute challenge: The development of powerful AI requires significant compute power, access to which is limited to a few organisations.
6. The Black Box challenge: Some AI models and tools cannot explain why they produce a particular result, which is a challenge to transparency requirements.
7. The Open-Source challenge: Requiring code to be openly available may promote transparency and innovation; allowing it to be proprietary may concentrate market power but allow more dependable regulation of harms.
8. The Intellectual Property and Copyright Challenge: Some AI models and tools make use of other people's content: policy must establish the rights of the originators of this content, and these rights must be enforced.
9. The Liability challenge: If AI models and tools are used by third parties to do harm, policy must establish whether developers or providers of the technology bear any liability for harms done.
10. The Employment challenge: AI will disrupt the jobs that people do and that are available to be done. Policy makers must anticipate and manage the disruption.
11. The International Coordination challenge: AI is a global technology, and the development of governance frameworks to regulate its uses must be an international undertaking.
12. The Existential challenge: Some people think that AI is a major threat to human life. If that is a possibility, governance needs to provide protections for national security.

See UK Government's response to the Interim Report.

4.5. UK AI Safety Summit at Bletchley Park (Nov 2023)

At the UK AI Safety Summit, the 28 countries[6] in attendance agreed to the Bletchley Declaration on AI Safety. The Bletchley Declaration emphasises an urgent need to understand and collectively manage potential risks through a new joint global effort to ensure AI is developed and deployed safely and responsibley to benefit the global community.

Attending countries and frontier AI labs recognised through a joint statement the need to collaborate on testing next generation AI models against a range of critical national security, safety, and societal risks. Soon after the Summit, the UK's National Cyber Security Centre (NCSC), together with 22 other international agencies, announced new guidelines for secure AI system development.

Countries represented at Bletchley Park also agreed to support the development of a 'State of the Science' report to help build international consensus on the capabilities and risks of frontier AI.

The next virtual summit will be co-hosted by the Republic of Korea 6 months after the UK summit, and France will host the next in person Summit 6 months after that.

4.6. UK Government Bodies Concerned with AI

The AI Council (not a government body, but still relevant) was established in 2019 to provide advice to the government and high-level leadership of the AI ecosystem. In January 2021, the committee of independent experts published its 'AI Roadmap' providing 16 recommendations to the government on the strategic direction for AI. The council also led a programme of engagement with the wider AI community to inform the development of the National AI Strategy. The final meeting on the AI Council was held on 21 June, 2023.

The Office for Artificial Intelligence was established in 2021 to oversee the implementation of the National AI Strategy. It now sits within the Department for Science, Innovation and Technology (DSIT)

Department for Science Innovation and Technology (DSIT). In the February 2023 cabinet reshuffle, the Department for Digital, Culture, Media and sport (DCMS) split into the Department for Culture, Media, and Sport (DCMS) and the Department for Science Innovation

[6] 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

and Technology (DSIT). DSIT combines the old DCMS's tech and digital work with the R&D, science, and innovation policy parts of the also dissolved Department for Business Energy and Industrial Strategy (BEIS).

DSIT is responsible for positioning the UK at the forefront of global scientific and technological advancement and for driving technological innovation that improves lives and sustains economic growth. DSIT houses the Office for AI and the central risk function.

AI Safety Institute (AISI) (formerly the Frontier AI Taskforce, formerly the Foundation Model Taskforce) is a state-backed organization focused on advanced AI safety for the public interest. It is not a regulator. Its mission is to "minimise surprise to the UK and humanity from rapid and unexpected advances in AI". It works towards these goals by:

- developing the sociotechnical infrastructure needed to understand the risks of advanced AI and enable its governance.
- advancing the world's knowledge of AI safety by carefully examining, evaluating, and testing new types of AI, so that we understand what each new model is capable of.
- conducting fundamental research on how to keep people safe in the face of fast and unpredictable progress in AI.

AISI defines AI safety as:

> *The understanding, prevention, and mitigation of harms from AI. These harms could be deliberate or accidental; caused to individuals, groups, organizations, nations or globally; and of many types, including but not limited to physical, psychological, social, or economic harms.*

5. International Happenings

5.1. United States

On October 30, 2023 the Biden Administration issued an Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence with the aim of protecting citizens from potential risks from AI.

The EO includes stipulations for:

- Notifying the government of training runs and sharing safety test results with the government
- Developing safety evaluation and risks assessment standards, tools, and tests.
- Preventing the use of AI to engineer dangerous biological materials
- Studying the risks and benefits of openly sharing model weights
- Evaluating and improving privacy preserving techniques
- Developing best practices for use of AI throughout the criminal justice system to prevent discrimination and preserve human rights.
- Developing best practices to mitigate harms and maximize the benefits of AI for workers and to address job displacement.
- Promoting a fair, open, and competitive AI ecosystem
- Accelerating the rapid hiring of AI professional into government departments

5.2. European Union

On December 9, 2023 the European Parliament and Council reached an agreement on the EU AI Act: "This regulation aims to ensure that fundamental rights, democracy, the rule of law and environmental sustainability are protected from high risk AI, while boosting innovation… The rules establish obligations for AI based on its potential risks and level of impact."

The Act prohibits certain AI applications such as emotional recognition in the workplace and education institutions, social scoring, AI systems that manipulate human behaviour to circumvent free will, and biometric categorization that uses sensitive characteristics (e.g. race, sexual orientation, political, religious beliefs).

The EU AI Act also outlines transparency requirements for general purpose AI systems as well as model evaluation and risks assessment obligations for high-risk and high-impact systems.

6. Demos/GovAI Roundtable Reflections

On the what it means for the UK to be a global AI leader:

The UK National AI Strategy (2021) stated an intention for the UK to remain an "AI Superpower fit for the next decade". This commitment to AI leadership was upheld in the AI White Paper (2023) which outlined a plan for a pro-innovation approach to AI regulation.

The roundtable began with a discussion of how we should think about the UK leading in AI. Is being an AI superpower a realistic goal for the UK? How else might we think about AI leadership?

One participant raised the concern that being a leader in science and innovation may be at odds with being a leader in responsible practice. The former requires moving fast and cutting red tape, the latter requires the opposite approach. Roundtable discussion on this point noted that in the shadow of the Postmaster scandal - the effort to rapidly introduce new automated accounting software which resulted in false accusations of fraud, loss of livelihood, and loss of life - the UK public mentality has veered from a "move fast and break things" perspective will likely be more supportive of efforts to lead in safe and responsible AI development and deployment.

Furthermore the UK may not be well placed to lead in AI development given the extremely high costs of frontier model development and the significant head start help by industry mainly in the US in China. Discussion turned instead to how the UK can play to its strengths to demonstrate global leadership, if not in frontier AI development, in the larger AI landscape.

In particular, the UK has a historical strength in convening global conversations (a tradition continued with the 2023 UK AI Safety Summit) as well as an extremely high concentration of AI Governance expertise within organisations such as The Alan Turing Institute, DSIT, The Centre for the Governance of AI, the Ada Lovelace Institute, the Leverhulme Centre for the Future of Intelligence, and the Oxford Internet Institute. This makes the UK well-placed to facilitate international governance conversations, especially between the US, EU, and China; participants observed that had the US hosted the first AI Safety Summit, it is highly unlikely that China would have been brought to the table.

Furthermore, the UK has a particularly strong AI start-up scene. While leading in cutting edge frontier AI model development felt to many at the table like an unrealistic goal, there was broad agreement the UK is well placed to unlock value from AI by developing AI applications that capitalise on growing AI capabilities to serve a wide variety of consumer needs and to more effectively deliver public services.

On UK Government Activities so far:

While there was concern about the UK aiming to be a global AI superpower, Government activities have reflected a perhaps more grounded interpretation of what it means for the UK to lead in AI. For example, the UK has really leaned into its convening and international coordination role with the UK AI Safety Summit at Bletchley park. The UK AI Safety Institute (AISI) taps into the UK's impressive in-house AI research and governance expertise and has attracted international talent as well to research and inform international AI safety standards.

AISI is not a regulator, and therefore can not enforce AI safety or model release standards. Some participants wondered if by not conferring regulatory power to AISI that the UK government runs the risk of AISI becoming an AI policy research group too far removed from policy impact. Is there a missed opportunity here? However others pointed out that only as a non-regulatory body can AISI play a role coordinating with other international AI safety institutes and efforts.

Some participants also raised concern that the non-statutory, principle-based approach to regulation outlined in the White Paper is too non-specific. Concrete requirements are needed to offer clear corporate guidance and ensure AI safety. Others responded that we should not read too far into the"non-statutory" part of the non-statutory expectations of regulators. We need to build policy at speed to keep up with the pace of AI innovation. This requires a flexible approach to innovating policy that can be solidified as we move forward. This first step is to enable and tool up individual regulators to research and devise plans for fit for purpose regulation in their respective sectors. The intention remains to build towards clear statutory requirements.

It was generally agreed that so far DSIT has done an incredible job moving on issues of AI safety and international AI governance coordination in a very short period of time. There was a worry expressed, however, that what we are seeing is the execution of a DSIT AI strategy as opposed to a wider UK Government AI strategy. If it were a UK Government AI strategy then we would be seeing more regulatory and information gathering activities like those being undertaken by DSIT, spread across numerous government departments. In this respect the UK should perhaps think to emulate more the approach demonstrated by the US Executive Order on AI whereby various government departments have been charged with investigating specific policy challenges and, in some cases, implementing and enforcing solutions.

One key request was that we take a step back to think critically about the overarching purpose of the Government's AI strategy. What is an AI strategy for? Is it meant to be a strategy for regulation, innovation, foreign policy, jobs, community empowerment, or economic development? So far our discussion has represented the upstream regulation of AI research and development with less attention to the downstream impacts of AI deployment and adoption on people. What would a people centred AI strategy look like?

On the UK's AI policy priorities going forward

DSIT is faced with a near impossible task. AI seems to touch everything, and so prioritisation is key. While roundtable participants discussed numerous areas in need of more attention (indeed, as one participant put it, "the most difficult thing is deciding what not to do") focus on AI adoption stood out as both urgent and relatively overlooked.

The UK has traditionally been great with science and innovation, but has then struggled to translate innovation to value for citizens, especially value that is widely distributed. Successful AI adoption across private industry and in the public sector requires not only that a useful AI-enabled application exists, but that users know how to effectively use the tool; they must be provided with adequate training and skills development opportunities to capture the potential value. It will also help to have a clear model for generating public benefit from AI, for example, by boosting productivity, providing people better access to public services, and distributing value produced by AI.

Furthermore, users need to trust the technology; they need reason to believe that the technology is reliable - that it will behave as expected - and that they are not opening themselves to harm by using the technology, for example, by putting one's self at risk of job displacement or of being held responsible harms caused by the technology that they had no way to foresee or prevent. Especially where AI is deployed in public services, many people will not have a choice but to interact with AI. Fit regulation is needed to protect user and consumer well being. Developing clear IP and liability regulation for businesses will be key, in addition to model validation and assurance standards. Civil society and a broad range of stakeholders should be consulted in the development of these standards.