UK Government

INTRODUCTION TO THE AI SAFETY SUMMIT

Introduction to the AI Safety Summit

AI Safety Summit – Introduction

We are in the midst of a technological revolution that will fundamentally alter the way we live, work, and relate to one another. Artificial Intelligence (AI) has begun and promises to further transform nearly every aspect of our economy and society, bringing with it huge opportunities but also risks that could threaten global stability and undermine our values.

The opportunities are transformational - advancing drug discovery, making transport safer and cleaner, improving public services, speeding up and improving diagnosis and treatment of diseases like cancer and much more. To seize these opportunities however, we must grip the risks, not just here but as a global endeavour which is why we are organising the first Global Summit on AI Safety.

The AI Safety Summit will focus on how to best manage the risks from the most recent advances in AI ('*Frontier AI' – see below*). These risks necessitate an urgent international conversation given the rapid pace at which the technology is developing. The dynamic nature of AI means it can be hard to predict the risks. Different variables - how the AI is designed, the uses to which it is put, the data on which it is trained - can interact in ways that are near impossible to predict. Governments, academics, companies and civil society groups need to work together to understand these risks, and possible remedies.

In the road to the Summit, there will be several opportunities for people to contribute to an inclusive debate. The Summit itself will bring together a small number of countries, academics, civil society representatives and companies who have already done thinking in this area to begin this critical global conversation. This is the focus of the first AI Safety Summit hosted by the UK at Bletchley Park and a conversation we will be continuing thereafter.

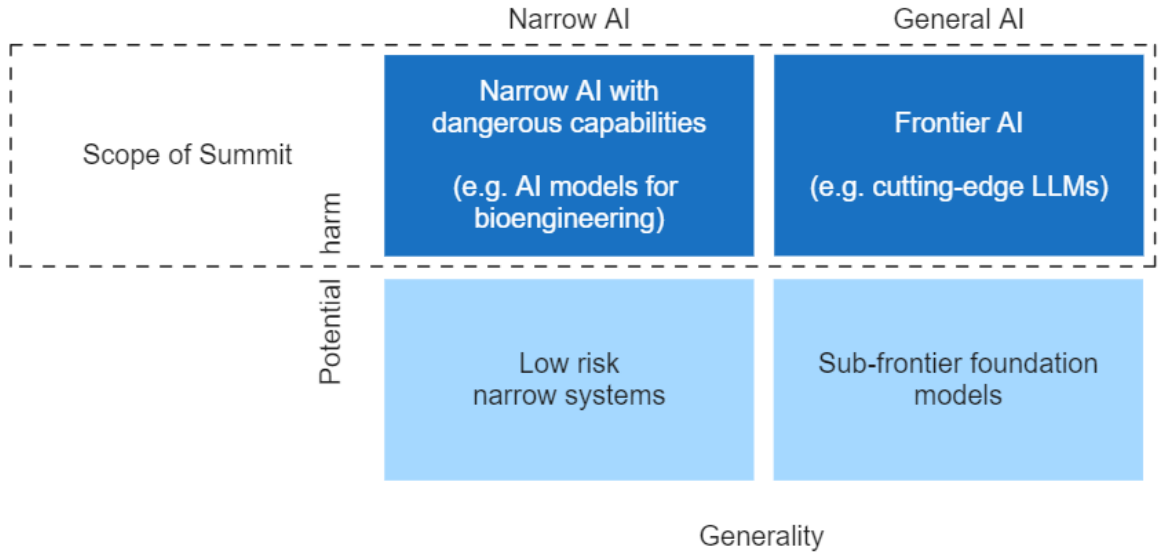
Scope of the AI Safety Summit – what is Frontier AI?

The Summit will focus on certain types of AI systems based on the risks they may pose. As set out below, these risks could stem from the most potentially dangerous capabilities of AI, which we understand to be both at the 'frontier' of general purpose AI, as well as in some cases specific narrow AI which can hold potentially dangerous capabilities.

We understand frontier AI to be highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models. This builds on a number of existing definitions used by various civil society and corporate actors. As well as discussing our understanding of this with Summit attendees, the UK is taking forward work to specifically identify *where* the frontier of AI development currently is and where it may advance to in the future.

AI technologies are evolving with unprecedented speed. Soon, models many times more powerful than what is currently available may be released. The capabilities of these models are very difficult to predict – sometimes even to those building them - and by default they could be made available to a wide range of actors, including those who might wish us harm. This pace of change means that urgent action on AI safety is needed. We are at a crossroads in human history and to turn the other way would be a monumental missed opportunity for mankind.

Whilst the Summit's focus will be on frontier AI, it will also be important to consider these alongside the risks of this technology in certain use cases – werefer to this sometimes as narrow AI systems. This is because we cannot fully predict the development of technology and cannot know for certain how narrow or general the riskiest AI systems of the future will be, or the extent to which risks may be created through the use of narrow AI as tools for general AI.



AI safety does not currently have a universally agreed definition and it is best considered as the prevention and mitigation of harms from AI. These harms could be deliberate or accidental; caused to individuals, groups, organisations, nations or globally; and of many types, including but not limited to physical, psychological, or economic harms.

Focus and objectives

As illustrated in the diagram above, there are a small set of potential harms which are most likely to be realised at the frontier, or in a small number of cases through very specific narrow systems. There are two particular categories of risk that we are focusing on:

1. Misuse risks, for example where a bad actor is aided by new AI capabilities in biological or cyber-attacks, development of dangerous technologies, or critical system interference. Unchecked, this could create significant harm, including the loss of life.

2. Loss of control risks that could emerge from advanced systems that we would seek to be aligned with our values and intentions.

The Summit, in looking specifically at frontier AI safety, will particularly focus on these two areas. This recognises the urgent need for an international conversation on how we can work together to meet the novel challenges these risks pose, combat misuse of models by non-state actors, and promote best practice. This builds on work addressing other risks and harms from AI, including at the OECD, Global Partnership on AI and Council of Europe, and the Hiroshima AI Process.

This is not to minimise the wider societal risks that such AI – both at the frontier and not - can have, including misinformation, bias and discrimination and the potential for mass automation. The UK considers that these are best addressed by the existing international processes that are underway, as well as nations' respective domestic processes. For example, in the UK these risks are being considered through the work announced in the white paper on AI regulation and through wider work across government. It is therefore important that the focus of the Summit is complementary and not duplicative of these existing efforts, and that the UK and other nations continue to work at pace across all these forums to address the full range of risks. We therefore welcome the work taken by our international partners to identify appropriate voluntary measures which may be implemented at pace and on an interim basis, whilst our understanding of the technology continues to evolve.

The first AI Safety Summit has five objectives:

- a shared understanding of the risks posed by frontier AI and the need for action
- a forward process for international collaboration on frontier AI safety, including how best to support national and international frameworks
- appropriate measures which individual organisations should take to increase frontier AI safety
- areas for potential collaboration on AI safety research, including evaluating model capabilities and the development of new standards to support governance
- showcase how ensuring the safe development of AI will enable AI to be used for good globally

Road to the Summit – Engagement

The Summit is only the beginning of a global process to identify, evaluate and ultimately mitigate risks from frontier AI in order for us all to enjoy the public benefits that AI can bring.

Michelle Donelan, Secretary of State for Science, Innovation and Technology, launched the start of formal engagement prior to the Summit in early September, with Jonathan Black and Matt Clifford, the Prime Minister's Representatives for the AI Safety Summit, beginning discussions with countries and a number of leading frontier AI organisations. The Secretary of State also hosted a roundtable with a

cross-section of civil society groups and continues to have bilateral engagements to inform the Summit engagement and programme.

In the coming weeks, the Government will continue to engage with the international and domestic scientific community, civil society and businesses to hear from them on the content of the Summit, and on wider issues relating to AI.

Extending the conversation

To ensure the Summit can achieve the objectives set out, it is necessary to ensure a small and focused discussion at the two-day event itself, which will be limited to around 100 participants. However, we welcome insight from the many more individuals and organisations who have the expertise and desire to contribute to this critical topic. The scope of the Summit is focused on the safety risks of frontier AI, but our wider engagement will ensure that other important issues are also discussed.

To allow a wider range of voices to be heard, we are partnering on four official pre-Summit events with the Royal Society, the British Academy, techUK and The Alan Turing Institute. The outcomes of these workshops will feed directly into the Summit planning, and we will publish an external summary of the engagement.

11 October	**The Alan Turing Institute** Exploring existing UK strengths on AI safety and opportunities for international collaboration.
12 October	**British Academy** Possibilities of AI for the public good: the Summit and beyond.
17 October	**techUK** Opportunities from AI; Potential risks from AI; Solutions that exist in the tech sector.
25 October	**Royal Society** Horizon scanning AI safety risks across scientific disciplines.

Public Engagement

Frontier AI safety is an issue that impacts us all and is not merely a question for academics and technical experts. In the month leading up to the Summit, members of the public from anywhere in the world will be able to ask questions and share their views directly with government.

Event details will be shared on the Department for Science, Innovation and Technology (DSIT)'s social media channels ahead of time with the facility to watch back and post questions and comments on social media channels at other times.

2 October	X Q&A with Matt Clifford, the Prime Minister's Representative for the AI Safety Summit
16 October	LinkedIn Q&A with Secretary of State Michelle Donelan
1 November	Watch keynote speeches on the Summit livestream

On @scitechgovuk on X, LinkedIn, Facebook, Instagram and Threads there will be further regular Summit updates and opportunities to engage in the run up to and throughout the Summit.

Glossary of terms

AI or AI system or AI technologies	Products and services that are 'adaptable' and 'autonomous' in the sense outlined in our definition in section 3.2.1 of the AI White Paper.
Foundation Model	Foundation models are models trained on broad data at scale such that they can be adapted to a wide range of downstream tasks. - *Stanford Centre for Research on Foundation Models*
Frontier AI	Frontier AI is at the cutting edge of technological advancement – therefore offering the most opportunities but also presenting new risks. It refers to highly capable general-purpose AI models, most often foundation models, that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models. It can then enable narrow use cases.
Narrow AI	Narrow artificial intelligence (narrow AI) is AI that is designed to perform a specific task. It is a specific type of artificial intelligence in which a learning algorithm is designed to perform a single task or narrow set of tasks, and any knowledge gained from performing the task will not automatically be applicable or transferable to a wide variety of tasks.

Mass automation	Mass automation is the process which leads to software completing tasks in a faster, more efficient, and cheaper fashion than can be done by humans.