UK Government

Introducing the AI Safety Institute

Introducing the AI Safety Institute

Presented to Parliament

by the Secretary of State for Science, Innovation and Technology

by Command of His Majesty

November 2023

CP 960

This publication is available at www.gov.uk/official-documents.

Any enquiries regarding this publication should be sent to us at correspondence@dsit.gov.uk.

ISBN 978-1-5286-4538-6

E03012924 11/23

Printed on paper containing 40% recycled fibre content minimum

Printed in the UK by HH Associates Ltd. on behalf of the Controller of His Majesty's Stationery Office

Ministerial Foreword

The release of ChatGPT was a Sputnik moment for humanity – we were surprised by rapid and unexpected progress in a technology of our own creation. With accelerating investment into and public adoption of advanced AI, these systems are becoming more powerful and consequential to our lives.

These systems could free people everywhere from tedious routine work and amplify our creative abilities. But they could also change our future labour markets and economy more quickly than any other technological advance in history. They could help our scientists unlock bold new discoveries, opening the door to a world without cancer and with access to near-limitless clean energy. But they could also further concentrate unaccountable power into the hands of a few, or be maliciously used to undermine societal trust, erode public safety, or threaten international security.

Some of these risks already manifest as harms to people today and are exacerbated by advances at the frontier of AI development. The existence of other risks is more contentious and polarising. But in the words of mathematician I.J. Good, a codebreaking colleague of Alan Turing at Bletchley Park, "It is sometimes worthwhile to take science fiction seriously."

We must always remember that AI is not a natural phenomenon that is happening to us, but a product of human creation that we have the power to shape and direct. Accordingly, we are not waiting to react to its impacts but are choosing to be proactive in defining the trajectory of its development, to ensure public safety and human flourishing for years to come. This is why the UK is building the AI Safety Institute.

The Institute is the first state-backed organisation focused on advanced AI safety for the public interest. Its mission is to minimise surprise to the UK and humanity from rapid and unexpected advances in AI. It will work towards this by developing the sociotechnical infrastructure needed to understand the risks of advanced AI and enable its governance. Its work will move the discussion forward from the speculative and philosophical, further towards the scientific and empirical.

This is our contribution to addressing a shared challenge posed to all of humanity. In doing so, we can safely capture the existential upsides of AI for future generations to come.

What we are building here could be truly historic – and it's worth reflecting on where we started from. 73 years ago, Alan Turing dared to ask if computers would one day think. From his vantage point at the dawn of the field, he observed that "we can only see a short distance ahead, but we can see plenty there that needs to be done."

We can see further yet today, and with ever more that needs to be done. So let's get to work.

RT HON MICHELLE DONELAN MP

Secretary of State for Science, Innovation and Technology

Department for Science, Innovation and Technology

Introduction

Advances in artificial intelligence (AI) over the last decade have been impactful, rapid, and unpredictable. Today, harnessing AI is an opportunity that could be transformational for the UK and the rest of the world. Advanced AI systems have the potential to drive economic growth and productivity, boost health and wellbeing, improve public services, and increase security.

The UK Government is determined to seize these opportunities. In September, we announced Isambard AI as the UK AI Research Resource, which will be one of Europe's most powerful supercomputers purpose-built for AI. The National Health Service NHS is running trials to help clinicians identify breast cancer sooner by using AI. In the workplace, AI promises to free us from routine tasks, giving teachers more time to teach and police officers more time to tackle crime. There is a world of opportunity for the UK that we will explore.

But advanced AI systems also pose significant risks, as detailed in the Government's paper on Capabilities and Risks from Frontier AI published in October. AI can be misused – this could include using AI to generate disinformation, conduct sophisticated cyberattacks or develop chemical weapons. AI can cause societal harms – there have been examples of AI chatbots encouraging harmful actions, promoting skewed or radical views, and providing biased advice. AI generated content that is highly realistic but false could reduce public trust in information. Some experts are concerned that humanity could lose control of advanced systems, with potentially catastrophic and permanent consequences.

We will only unlock the benefits of AI if we can manage these risks. At present, our ability to develop powerful systems outpaces our ability to make them safe. The first step is to better understand the capabilities and risks of these advanced AI systems. This will then inform our regulatory framework for AI, so we ensure AI is developed and deployed safely and responsibly.

The UK is taking a leading role in driving this conversation forward internationally. We launched the Frontier AI Taskforce – the first state body dedicated to the safety of advanced AI, investing more than any other nation – and hosted the world's first major AI Safety Summit. Responsible government action in an area as new and fast-paced as advanced AI requires governments to develop their own sophisticated technical and sociotechnical expertise.

The Artificial Intelligence Safety Institute (AISI) is the next step in this process. It will advance 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. It will conduct fundamental research on how to keep people safe in the face of fast and unpredictable progress in AI. The Institute will make its work available to the world, enabling an effective global response to the opportunities and risks of advanced AI.

Mission and Scope

The Institute is the first state-backed organisation focused on advanced AI safety for the public interest. Its mission is to minimise surprise to the UK and humanity from rapid and unexpected advances in AI. It will work towards this by developing the sociotechnical infrastructure needed to understand the risks of advanced AI and enable its governance.

This mission stems from our conviction that governments have a key role to play in providing publicly accountable evaluations of AI systems and supporting research. While developers of AI systems may undertake their own safety research, there is no common standard in quality or consistency. Beyond this, only governments can run evaluations on issues related to national security as they require access to very sensitive knowledge. Governments will only be able to develop effective policy and regulatory responses to AI if they understand the technology better than they do today. By building a body of evidence on the risks from advanced AI, the Institute will lay the foundations for technically grounded international governance.

The Institute will focus on the most advanced current AI capabilities and any future developments, aiming to ensure that the UK and the world are not caught off guard by progress at the frontier of AI in a field that is highly uncertain. It will consider open-source systems as well as those deployed with various forms of access controls. Both AI safety and security are in scope.

The research of the AI Safety Institute will inform UK and international policymaking and provide technical tools for governance and regulation. Possible examples of technical tools include secure methods to fine-tune systems with sensitive data, platforms to solicit collective input and participation in model training and risk assessment, or techniques to analyse training data for bias (see Box 2).

The Institute is not a regulator and will not determine government regulation. It will collaborate with existing organisations within government, academia, civil society, and the private sector to avoid duplication, ensuring that activity is both informing and complementing the UK's regulatory approach to AI as set out in the AI Regulation White Paper. It will provide foundational insights to our governance regime and be a leading player in ensuring that the UK takes an evidence-based, proportionate response to regulating the risks of AI.

The Institute will establish the UK as a global hub for safety research, deepening the UK's stake in this strategically important technology. By improving the safety of advanced AI, the Institute will pave the way for increased adoption of advanced AI in this country, so that the UK is well-placed to seize these benefits.

Functions

The Institute will adjust its activities within the scope of its headline mission to ensure maximum impact in a rapidly evolving field. It will initially perform three core functions:

- **Develop and conduct evaluations on advanced AI systems**, aiming to characterise safety-relevant capabilities, understand the safety and security of systems, and assess their societal impacts.
- **Drive foundational AI safety research**, including through launching a range of exploratory research projects and convening external researchers.
- **Facilitate information exchange**, including by establishing – on a voluntary basis and subject to existing privacy and data regulation – clear information-sharing channels between the Institute and other national and international actors, such as policymakers, international partners, private companies, academia, civil society, and the broader public.

Each of these functions is considered in greater detail below.

Develop and Conduct AI System Evaluations

AI system evaluations are thorough assessments of a system's safety-relevant properties. These properties include:

- Capabilities most relevant to AI misuse, such as the ability to meaningfully lower barriers for a human attacker seeking to cause real-world harm.
- Capabilities that might exacerbate existing and future societal harms, such as psychological impacts, manipulation and persuasion, impacts on democracy, biased outputs and reasoning, or systemic discrimination.
- System safety and security, such as understanding the efficacy and limitations of system safeguards and the adequacy of cybersecurity measures.
- Abilities and tendencies that might lead to loss of control, such as deceiving human operators, autonomously replicating, and adapting to human attempts to intervene.

Further detail can be found in Box 1.

As agreed at the 2023 Global AI Safety Summit, ensuring the safety of advanced AI systems is a shared responsibility across all steps from early AI development to its use, and in particular between the actors developing and deploying them. Developers both have responsibility to devise and conduct safety testing through evaluations, transparency, and other appropriate measures, and the technical means of mitigating risks and addressing vulnerabilities. We see a key role for government in providing external evaluations independent of commercial pressures and supporting greater standardisation and promotion of best practice in evaluation more broadly. This is also reflected in our publication on Emerging Processes for Frontier AI Safety, which details the role independent, external evaluations can play in ensuring safety.

AI safety research, and work related to evaluations, is becoming more prevalent in academia. There are also a range of private sector efforts to build tools to evaluate AI systems, such as those highlighted by the Department for Science, Innovation and Technology's (DSIT) portfolio of AI assurance techniques. However, only a small number of private organisations are currently evaluating the most advanced AI systems. Most of these evaluations are taking place inside the top AI tech companies. Governments and any external parties are unable to verify the results of these evaluations.

The Institute will develop and run system evaluations, independently and in partnership with external organisations, while also seeking to address a range of open research questions connected to evaluations. Evaluations may not be able to fully understand the limits of capabilities or assure that safeguards are effective. The goal of the Institute's evaluations will not be to designate any particular AI system as "safe", and the Institute will not hold responsibility for any release decisions. Nevertheless, we expect progress in system evaluations to enable better informed decision-making by governments and companies and act as an early warning system for some of the most concerning risks. The Institute's evaluation efforts will be supported by active research and clear communication on the limitations of evaluations. The Institute will also convene expert communities to give input and guidance in the development of system evaluations..

Box 1. Evaluation Priorities

Early evaluations by the Institute will likely cover the following four areas of interest. As the Institute grows, these focuses are likely to expand.

1. **Dual-use capabilities:** As AI systems become more capable, there could be an increased risk that malicious actors could use these systems as tools to cause harm. Evaluations will gauge the capabilities most relevant to enabling malicious actors, such as aiding in cyber-criminality, biological or chemical science, human persuasion, large-scale disinformation campaigns, and weapons acquisition. Such evaluations will draw heavily from relevant expertise inside and outside of government.

2. **Societal impacts:** As AI is integrated into society, existing harms caused by current systems will likely increase, requiring both pre and post-deployment evaluations. These evaluations will seek to investigate psychological impacts, privacy harms, manipulation and persuasion, biased outputs and reasoning, impacts on democracy and trust in institutions, and systemic discrimination. Such evaluations may be conducted in part post-deployment, drawing from usage data and incident reporting. Evaluations will build on existing work in the UK ecosystem, such as by the Centre for Data Ethics and Innovation, the Ada Lovelace Institute, the Turing Institute, and the Bridging Responsible AI Divides (BRAID) and Responsible AI UK (RAI UK) programmes.

3. **System safety and security:** Current safeguards are unable to prevent determined actors from misusing today's AI systems, for example by breaking safeguards or taking advantage of insecure model weights. Safety and security evaluations will seek to understand the limitations of current safeguard methodologies and research potential mitigations. These evaluations will range from automated or human-crafted real-world attacks on full AI systems, to more intensive examinations of individual safeguard components. Evaluation protocols will draw from relevant expertise, including from areas like safety-critical infrastructure and best practices in auditing.

4. **Loss of control:** As advanced AI systems become increasingly capable, autonomous, and goal-directed, there may be a risk that human overseers are no longer capable of effectively constraining the system's behaviour. Such capabilities may emerge unexpectedly and pose problems should safeguards fail to constrain system behaviour. Evaluations will seek to avoid such accidents by characterising relevant abilities, such as the ability to deceive human operators, autonomously replicate, or adapt to human attempts to intervene. Evaluations may also aim to track the ability to leverage AI systems to create more powerful systems, which may lead to rapid advancements in a relatively short amount of time.

Driving Foundational AI Safety Research

System evaluations alone are not sufficient to ensure safe and beneficial development and deployment of advanced AI. There may be fundamental limitations in the ability of evaluations to assess risks, and effective governance requires capabilities other than risk assessment.

The Institute will therefore pursue foundational AI safety research to advance global understanding of the risks that advanced AI systems pose and develop the technical tools necessary for effective AI governance. Examples of these research topics can be found in Box 2.

Box 2. AI Safety Institute Research

The Institute's research will support short and long-term AI governance. It will ensure the UK's iterative regulatory framework for AI is informed by the latest expertise and lay the foundation for technically grounded international governance of advanced AI. Projects will range from rapid development of tools to inform governance, to exploratory AI safety research which may be underexplored by industry. Some examples of projects the Institute may pursue include:

1. **Building products for AI governance.** Effective governance of AI systems may require developing new real-world tools. Such tools could include secure methods to prompt or fine-tune systems with sensitive data, techniques to analyse training data for bias or otherwise concerning properties, processes that enable broader input into core development decisions, or assurance methods to verify compliance with the UK's or other countries' regulatory frameworks for AI.

2. **Improving the science of evaluations.** In parallel to efforts to rapidly implement existing AI system evaluations, the Institute will conduct research aimed at developing future evaluations, as well as characterising the claims that can be supported by those evaluations. For example, the Institute may work to develop multidisciplinary sociotechnical evaluations aimed at measuring diffuse and hard-to-measure effects of integrating AI into society; or work to address the evaluation-capability gap, where system capabilities are underestimated by evaluators.

3. **Novel approaches to safer AI systems.** In cases where promising research directions are underexplored by other actors, the Institute will conduct and support fundamental AI safety research. Such efforts may include technical scoping of emergent capabilities, including studying the effects of human curation, synthetic data, and training on data generated by deployed AI systems; new methods for reducing filter bubble effects of personalised assistants; and proposing best practices for safe development and deployment of advanced AI systems, including developing methods to enable responsible open-source innovation.

Research at the Institute will draw upon experience from across the AI ecosystem. The Institute will partner with existing organisations or initiatives – including internationally. It will focus on research that cannot or is not taken forward by other actors in academia or industry. The Institute expects to solicit input from a broad range of partners on its initial research agenda. It will also draw on the international research ecosystem to assess and synthesise existing research and aims to help forge scientific consensus around the state of AI and associated risks.

Facilitating Information Exchange

Due to technical complexity, competitive pressures, legal issues, and safety concerns, there are currently large insight gaps between industry, governments, academia, and the public. The Institute's evaluations and research are the first step in addressing this issue - improving understanding of the capabilities, safeguards, and societal impact of advanced AI systems. To ensure that relevant parties receive the information they need to effectively respond to rapid progress in AI, the Institute will appropriately share its findings with policymakers, regulators, private companies, international partners, and the public. This includes sharing the outcomes of the Institute's evaluations and research with other countries where advanced AI models will be deployed, where sharing can be done safely, securely and appropriately - as agreed at the AI Safety Summit.

The Institute will work with other UK government functions, such as DSIT's recently established Central AI Risk Function, to feed up to date information from the frontier of AI development and AI safety into government. This will ensure the UK's regulatory framework remains fit for purpose as AI technologies develop at pace.

Effective information sharing requires a trusted actor with deep connections across all parts of the AI ecosystem. There is currently a lack of clear channels for developers of advanced AI to share information with government. Competition laws and sensitivities around intellectual property can meanwhile limit information sharing between firms. The Institute could act as a trusted intermediary, enabling responsible dissemination of information as appropriate.

Additional approaches to support information exchange could include:

- Supporting the establishment of a clear process for academia and the broader public to report harms and vulnerabilities of deployed AI systems, such that government and other relevant actors are made adequately aware of the impact of AI on society.
- Where not provided by existing regulatory bodies, supporting the establishment of a clear process for AI tech companies to disclose information about their systems to bodies responsible for public safety.
- Supporting the creation of a panel of geographically diverse, multidisciplinary experts to contribute to risk assessment and red teaming.
- Supporting the assessment of societal impacts of AI by collating and sharing data on deployment and usage.
- Supporting information sharing between governments, to enable a global response to AI developments.
- Providing relevant parts of the UK government with the technical support needed to understand and respond to AI systems.

Several of these approaches have parallels to well-established processes in other sectors, such as for cybersecurity, nuclear power and food safety.

Partnerships

International Partners

The risks arising from AI are inherently global in nature and action to address them requires international cooperation. We welcome the international community's cooperation on the responsible development of AI systems. However, there is still a gap when it comes to reaching a consensus on how to develop and direct the field of advanced AI safety. To address this challenge, the 2023 Global AI Safety Summit was convened to establish international collaboration on identifying and mitigating safety risks from advanced AI.

Countries represented at the Summit agreed to the development of a "State of the Science" Report on the capabilities and risks of advanced AI, as part of their continued cooperation as an informal network. As host of the AI Safety Summit, the UK Government has commissioned Yoshua Bengio, a pioneering and Turing Award winning AI academic, to Chair the writing group that will draft the Report. This group will be composed of a diverse group of leading AI academics, supported by an Expert Advisory Panel made up of representatives from countries attending the Summit. The Institute will house the Secretariat for the Chair and we envisage that the Institute's cutting-edge research will also inform the Report.

The "State of the Science" Report will help build international consensus on the risks and capabilities of advanced AI. Rather than producing new material, it will summarise the best of existing research and identify areas of research priority, providing a synthesis of the existing knowledge of risks from advanced AI. It will not make policy or regulatory recommendations but will instead help to inform both international and domestic policy making. In focusing on advanced AI, it is also intended to help inform and complement other international initiatives.

Industry

To guarantee that the Institute is linked to the cutting edge of AI development, the Institute will work with leading AI tech companies. The research and evaluations conducted at the Institute will depend on access to frontier AI systems. Earlier this year, the Prime Minister announced that the leading AI tech companies had pledged to provide the Taskforce with priority access to their systems. We will seek that these companies will also provide access for the Institute so that its research team can undertake unhindered safety evaluations and share the results, as appropriate. In addition, we are developing processes for companies to share their expertise, including through potential secondment arrangements and close engagement to enable the Institute to retain expertise in developments at the frontier; and respond to the outputs of the Institute, taking action where governments identify potential risks.

In addition, the Institute will work with leading private sector organisations that deliver research and evaluations. We are looking forward to supporting and collaborating with the nascent AI assurance ecosystem in the UK and beyond to ensure we incorporate their valuable expertise. The Institute aims to support and complement private sector efforts, rather than competing with existing AI assurance and evaluation companies.

Academia and Civil Society

The Government welcomes the range of research on AI and AI safety taking place across civil society, including in universities and other research organisations. The Institute will build on existing work as far as possible. The Institute will establish partnerships with leading academics and civil society organisations in the UK and beyond.

Development of advanced AI is too often kept out of reach of academia and civil society. The Institute will work to facilitate their involvement which will support the safe and beneficial development of advanced AI. This will allow us to leverage the expertise of the UK's world-leading researchers.

National Security

The Institute will draw on the specialist expertise of the defence and national security community to support its work in assessing potential national security risks associated with advanced AI capabilities.

Establishment

The Institute is an evolution of the UK's Frontier AI Taskforce. The Frontier AI Taskforce was announced by the Prime Minister and Technology Secretary in April 2023. Since then, the Taskforce has assembled a globally recognised research team at the heart of government. The overarching objective of the Taskforce - to enable the safe and reliable development and deployment of advanced AI systems - has only become more pressing. The Taskforce will therefore become a permanent feature of the AI ecosystem. As of today, the Taskforce will become the AI Safety Institute, a new institution established for the long-term.

The Institute will continue the Taskforce's safety research and evaluations. The other core parts of the Taskforce's mission will remain in DSIT as policy functions: identifying new uses for AI in the public sector; and strengthening the UK's capabilities in AI.

Running evaluations and advancing safety research will also depend on access to compute. The Institute will receive priority access to state-of-the-art compute provided by the AI Research Resource (AIRR), which will deliver specialised compute capacity for use by the AI research community. The AIRR will integrate the recently announced Isambard-AI compute cluster at Bristol University, which will be one of the most powerful AI supercomputers in Europe. The Government is committed to supporting a thriving compute environment that maintains the UK's position as a leader across science, innovation and technology.

Talent is a key input for the Institute. The Taskforce research team will become the initial core of the Institute. The Institute will seek to attract new members to build an interdisciplinary team including, but not limited to, additional technical experts. We are grateful to the companies and civil society organisations that have already expressed an interest in seconding people to the Institute.

Ian Hogarth will continue as Chair of the AI Safety Institute and the External Advisory Board for the Taskforce will now advise the AI Safety Institute. A process for appointing the Chief Executive of the Institute will launch shortly.

Ensuring the development of advanced AI is safe is essential for harnessing the extraordinary opportunities of AI. The UK Government is therefore prepared to put significant investment behind the AI Safety Institute over the coming decade. With the initial £100m investment in the Frontier AI Taskforce, the UK is providing more funding for AI safety than any other country in the world. The Institute will be backed with a continuation of the Taskforce's 2024/25 funding as an annual amount for the rest of this decade, subject to it demonstrating the continued requirement for that level of public funds. This will be funded as part of the Government's record investment into R&D, which next year will have increased to £20bn.

Definitions

There is debate and disagreement relating to several key terms used in this document. For the purpose of this document, we are using the following working definitions:

- **Artificial Intelligence:** The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Modern AI is usually built using machine learning algorithms. The algorithms find complex patterns in data which can be used to form rules.
- **Machine Learning:** Algorithms that allow computers to recognise patterns in data, understanding the relationship between the data they are given and the problem the algorithm designer is trying to solve, without the rules having to be explicitly programmed by a human. Machine Learning is a sub-field of AI.
- **AI system:** The complete hardware and software setup through which one or more machine learning models is developed, deployed and/or made available to downstream users.
- **Advanced/frontier AI:** The terms 'advanced AI' and 'frontier AI' are contested. The intention of this document, where both terms are used, is to capture the cutting edge of technological advancement in AI – therefore offering the most opportunities but also presenting new risks. The scope of the AI Safety Institute includes both highly capable general-purpose AI models and narrow AI that is designed to perform a specific task, if the narrow system has high potential for harm. This matches the scope of the 2023 Global AI Safety Summit. Ahead of the Government's response to the AI Regulation White Paper, we intend to work to define terms more clearly in the context of fast-paced research developments.
- **AI safety:** The understanding, 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, social, or economic harms.
- **AI security:** Protecting AI models and systems containing AI components from attacks by malicious actors that may result in the disruption of, damage to, theft of, or unauthorised leaking of information about those systems and/or their related assets. This encompasses protecting AI systems from standard cybersecurity threats as well as those arising from novel vulnerabilities associated with AI workflows and supply chains (known as adversarial machine learning).

- **Sociotechnical:** Considering both technical and social aspects of an issue, and their interactions. For example, advanced AI systems can contain and magnify biases ingrained in the data they are trained on, or cheaply generate realistic content which can falsely portray people and events, with a risk of lowering societal trust in true information. Likewise, measures to improve safety, such as evaluating bias in AI systems or establishing a red teaming network, require multidisciplinary expertise beyond the technical.
- **Evaluations:** Systematic assessments of an AI system's safety-relevant properties. This does not constitute a pass/fail test or mandates conditions for deployment, but aims to improve understanding of the system's capabilities, behaviours, and safeguards.

E03012924

978-1-5286-4538-6