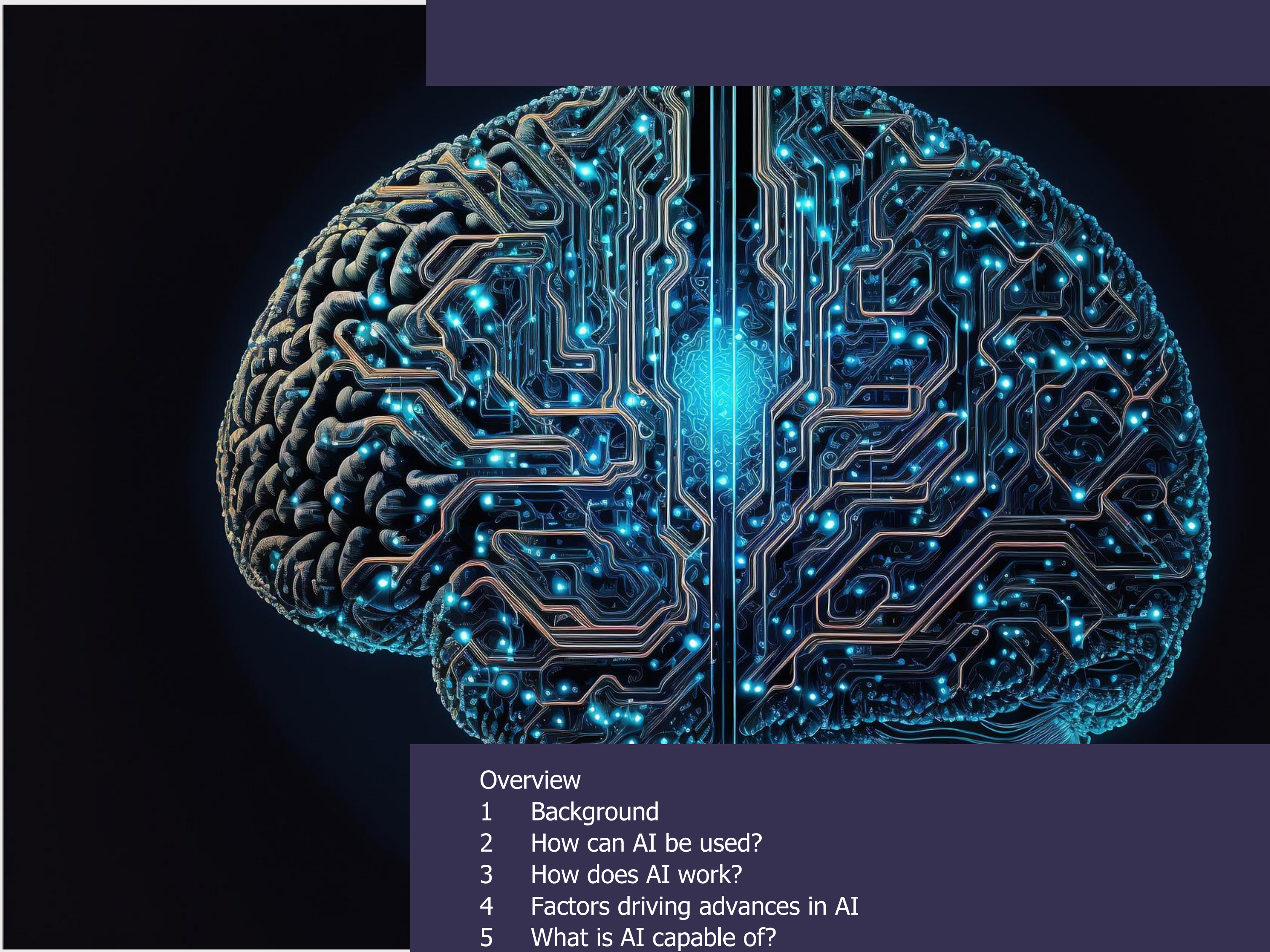
UK Parliament
POST

POSTbrief 57

By Devyani Gajjar

14 December 2023

Artificial intelligence: An explainer



post.parliament.uk

Feedback
If you have any comments on our briefings, please email papers@parliament.uk. Please note that we are not always able to engage in discussions with members of the public who express opinions about the content of our research, although we will carefully consider and correct any factual errors.

If you have general questions about the work of the House of Commons email hcenquiries@parliament.uk or the House of Lords email hlinfo@parliament.uk.

Suggested citation
Parliamentary Office of Science and Technology (POST). 2023. POSTbrief 57: Artificial Intelligence – An explainer. UK Parliament.

DOI: https://doi.org/10.58248/PB57

Image Credits
Image by Kohji Asakawa from Pixabay

Contents

Overview

Artificial Intelligence (AI) technologies are capable of:

- interpreting, processing and generating realistic human-like speech and text
- interpreting, processing and generating images, videos and other visuals
- independently performing tasks in the real world, such as when they are paired with machinery such as robots

AI technologies can be found in a wide range of everyday applications, including virtual assistants, search engines, navigation software, online banking and financial services, and facial recognition systems.

As a result, they can be applied in a wide range of sectors, such as healthcare, finance, education and commerce and can assist in tasks, such as decision making and improving productivity.

A 2023 report by McKinsey estimated that deploying and applying generative AI technologies has the potential to add between $2.6 trillion to $4.4 trillion annually to the global economy (greater than the entire GDP in the UK in 2021 of approximately $3.1 trillion).[1]

Many AI technologies are underpinned by 'machine learning,' which works by finding patterns in existing data (known as 'training data') and using these patterns to inform the processing of new data to make predictions or generate other outputs.

Some AI technologies, known as generative AI, can generate realistic outputs, such as text, audio, code, pictures, videos and music. Many AI technologies are designed to perform a specific task and cannot be adapted to other tasks.

Foundation Models are a type of machine learning model that can increasingly be adapted to a wide range of tasks, including generating realistic outputs.

Large Language Models are Foundation Models that carry out a range of language related tasks, such as processing and generating text.

Recent advances in AI technologies have been driven by: greater availability and volume of training data; computing power; computing investments; and new technology uses.

Concerns about AI technologies include:

- who has access to the biggest Large Language Models, such as a few technology companies

- the source, management and sharing of data that is used to train AI models, and the related privacy, security and discrimination implications
- impacts on the environment of training and running AI models
- challenges around the supply of AI hardware
- the ability of AI models to generate false information, which could lead to disinformation, biased decisions or discriminatory outcomes
- a lack of understanding about how large AI models make recommendations or decisions
- implications of AI for the economy and a lack of specialised AI skills to meet the growing demand in the UK workforce
- employment conditions for outsourced workers involved in developing large AI models

Issues such as bias in AI systems and additional policy and regulatory issues will be covered in the POSTnote on the policy implications of AI, due to be published in 2024.

In the past few years, various research has been conducted by academia, industry, NGOs and the public sector to determine public understanding of AI.

Experts have varying views on if, how and when future forms of AI are achievable and what nature these forms will take.

1 Background

There is no universally agreed definition of artificial intelligence (AI) or AI technologies.[2] Whilst this lack of a precise definition has helped AI to be adapted and advanced in different scenarios,[3] definitions can aid regulators.[4,5]

The UK Government's 2023 policy paper on 'A pro-innovation approach to AI regulation' defined AI, AI systems or AI technologies as 'products and services that are 'adaptable' and 'autonomous.' The 'adaptability' of AI refers to AI systems, after being trained, often developing the ability to perform new ways of finding patterns and connections in data that are not directly envisioned by their human programmers. The 'autonomy' of AI refers to some AI systems that can make decisions without the intent or ongoing control of a human.[*] This POSTbrief uses the same definition.

AI incorporates many different aspects of intelligence, such as reasoning, decision-making, learning from mistakes, communicating, problem-solving, and independently performing tasks in the real world.'[6] There are a wide variety of AI technologies, and multiple aspects of intelligence are often combined to deliver a task.

AI technologies can be found in a variety of everyday applications,[6,8] and AI has the potential to bring many social and economic benefits, such as increased labour productivity and improved services across a wide range of sectors (PN633, PN637 PN681, PN692).[1,9] AI could improve public sector services such as health and education, with implications for improved health outcomes, education delivery and cost savings (PN637, upcoming PN on the policy implications of AI, upcoming PN on AI in education delivery).

A 2023 report by McKinsey predicted that 75% of the global value that generative AI use cases could deliver would fall across customer operations, marketing and sales, software engineering and research and development.[1] Examples include supporting interactions with customers, generating creative content for marketing and sales, drafting computer code based on natural-language prompts, amongst others.[1] Other impacts could fall across banking, high tech and the life sciences amongst others.[1]

[*] The Organisation for Economic Co-operation and Development (OECD) also mentioned the adaptability and autonomy of AI in its November 2023 AI definition as '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 Alan Turing Institute[*] defined AI as "the design and study of machines that can perform tasks that would previously have required human (or other biological) brainpower to accomplish".[6] This definition is similar to how some academic institutions, such as Stanford University Institute for Human-Centred Artificial Intelligence,[7] have referred to AI.

However, AI can also create social and individual harms, and ethical challenges and issues (upcoming PN on the policy implications of AI), such as:[10–18]

- discrimination and inequalities from biases in AI systems[10]
- the spread of false information
- potential existential risks*
- data security and privacy challenges
- challenges around liability and transparency in AI systems and how AI should be regulated
- issues around unequal access to AI systems
- increased mistrust in online information
- environmental issues around the resources required to train and run AI systems
- copyright issues from AI outputting copyrighted material

The use of AI technologies could also impact the labour market.

This POSTbrief discusses technical information about AI in general. Some sections focus on Foundation Models and some on generative AI (see Appendix for definitions).

1.1 UK Government public policy developments

There have been significant UK Government public policy developments relating to AI in the last few years.

A National AI Strategy was published in 2021 with a 10-year plan to make the UK a 'global AI superpower' and an AI Action Plan followed in July 2022.[20,21] The Office for AI in the Department of Science, Innovation and Technology (DSIT) is responsible for overseeing the implementation of the National AI Strategy.

In March 2023, DSIT identified AI as one of five critical technologies in its Science and Technology Framework, outlining the Government's approach to making the UK a science and technology superpower by 2030.[22] This was followed by a White Paper on 'A pro-innovation approach to AI regulation.'[23] AI is governed via existing laws in different sectors of society (upcoming PN on the policy implications of AI, House of Commons Research Briefing on AI

* The Cambridge University Centre for the Study of Existential Risk refers to this as "risks that could lead to human extinction or civilisational collapse."[19]

and employment law), and devolution of AI governance operates according to existing sector policies.

The Prime Minister hosted a global summit on AI safety in November 2023. This resulted in 28 countries agreeing to The Bletchley Declaration on AI safety and both the UK and US Governments announcing new national AI Safety Institutes.[24–26] The UK Government has also announced significant investments over the past year, such as £54 million to develop trustworthy AI research.[27]

There has been a range of parliamentary activity around AI. In October 2022 the House of Commons Science, Innovation and Technology committee opened an inquiry into the governance of AI,[28] and in July 2023 the House of Lords Communications and Digital Committee opened an inquiry into Large Language Models[29] (Appendix).

2 How can AI be used?

AI has uses across several sectors, including:

- **Agriculture:** choosing optimum crops for weather conditions; monitoring crops and conditions; improving crop quality and resource efficiency; forecasting prices; and employing automated workers, such as robots that distribute fertiliser.[30,31]
- **Education:** Assisting teachers with lesson planning; scheduling; marking; responding to queries; diagnosing learners' needs; and identifying learners' needs and appropriate learning materials (upcoming PN on the use of AI in education delivery and assessment).[14,32,33]
- **Engineering:** Providing networks; forecasting; routing; employing in maintenance and security; and managing in network quality in energy, water, wastewater, transport and telecommunications infrastructure.[34,35]
- **Finance:** Enhancing data and analytic capabilities; increasing operational efficiency; detecting and flagging fraudulent activities; modelling investments; assessing risks; approving loans; and automating compliance (House of Commons Debate Pack on AI).[8,36]
- **Freight and transport:** Supporting freight management; monitoring goods; transporting and managing last minute deliveries; monitoring traffic flows; providing traffic status; and navigating (PN 692).[8]
- **Healthcare:** Medical imaging; helping clinicians make decisions; monitoring patient health; assisting surgeons in medical procedures; identifying patients at high risk of developing certain conditions; diagnosing diseases; devising personalised treatments; and identifying and developing new drugs (PN 637).[18,37–39]
- **Justice system, policing and security:** Predicting crimes; assisting visa-issuing authorities; and employing facial recognition tools to assess whether separate images are the same person (House of Lords report on AI and the justice system, AI in policing and security).
- **Manufacturing:** Processing data; monitoring, predicting, modelling, optimising and controlling processes; diagnosing faults; and estimating how long tools can be used[40]
- **Marketing and sales:** Gathering and analysing market trends and customer information; drafting personalised marketing and sales communications; assisting with marketing campaigns; and employing virtual sales representatives.[1]
- **National security and military operations:** Gathering intelligence; analysing data; and employing AI in weapons systems (PN 681 House of

Commons Research Briefing on Emerging and disruptive defence technologies).

- **Personal contexts:** Employing in search engines; chatbots; virtual personal assistants; activity trackers; and recommendation systems.[8]
- **Recruitment and management:** Devising job adverts; sourcing candidates; filtering CVs; allocating tasks; managing performance; surveillance; and monitoring of the workforce (House of Commons Research Briefing on AI and employment law).[41]

3 How does AI work?

All AI technologies are underpinned by an algorithm or a set of algorithms (PN 633). An algorithm is a set of instructions used to perform tasks (such as calculations and data analysis) usually using a computer or another smart device (PN 633).[42,43] AI often involves retraining algorithms with new data.

3.1 What is machine learning and how does it work?

Many AI applications, such as chatbots, predictive text and web recommendations, are underpinned by machine learning and its subset, deep learning (Appendix).

Machine learning systems learn by finding patterns in sample data.[6] They then create a model (with algorithms) encompassing their findings. This model is then typically applied to new data to make predictions or provide other useful outputs, such as translating text.[6] Sample data can be labelled (for example, pictures of cats and dogs labelled 'cat' or 'dog' accordingly) or unlabelled.

Training machine learning systems for specific applications can involve different forms of learning, such as supervised, unsupervised, semi-supervised and reinforcement learning (Box 1).

Box 1: Forms of machine learning

Machine learning can have varying degrees of autonomy.

1. **Supervised learning:** In a training phase, an AI system is fed labelled data. The system trains from the input data, and the resulting model is then tested to see if it can correctly apply labels to new unlabelled data (such as if it can correctly label unlabelled pictures of cats and dogs accordingly).[2] This type of learning is useful when it is clear what is being searched for,[2] such as identifying spam mail.[44]

2. **Unsupervised learning:** An AI system is fed large amounts of unlabelled data, in which it starts to recognise patterns of its own accord.[2] This type of learning is useful when it is not clear what patterns are hidden in data,[2] such as in online shopping basket recommendations ("customers who bought this item also bought the following items").[44]

3. **Semi-supervised learning**: An AI system uses a mix of supervised and unsupervised learning and labelled and unlabelled data.[44] This type of learning is useful when it is difficult to extract relevant features from data and when there are high volumes of complex data,[44] such as identifying abnormalities in medical images, like potential tumours or other markers of diseases.[44–47]

4. **Reinforcement learning:** An AI system is trained by being rewarded for following certain 'correct' strategies and punished if it follows the 'wrong' strategies.[2] After completing a task, the AI system receives feedback, which can sometimes be given by humans (known as 'reinforcement learning from human feedback')[48,49]. In the feedback, positive values are assigned to 'correct' strategies to encourage the AI system to use them, and negative values are assigned to 'wrong' strategies to discourage them, with the classification of 'correct' and 'wrong' depending on a pre-established outcome.[50,51] This type of learning is useful for tweaking an AI model to follow certain 'correct' behaviours, such as fine-tuning a chatbot to output a preferred style, tone or format of language.[52]

3.2 What are Foundation Models and how are they developed and deployed?

Foundation Models constitute a shift in AI model development and deployment.[12] They are a type of machine learning model that can be adapted to a range of general tasks such as translating and summarising text, responding to queries and generating new text, images, audio or visual content based on text or voice prompts (generative AI).[6,53] In contrast, standard AI models are typically designed by researchers and companies for a single specific application.[54,55,56]

Foundation Models include Large Language Models, which are trained on vast amounts of text to carry out a range of language related tasks, such as processing and generating text (see section 5.1 and Appendix). Cutting-edge Large Language Models, such as those underlying:

- ChatGPT (OpenAI),
- Claud (Anthropic),
- and Bard (Google),

have been referred to by the UK Government as 'Frontier AI' (Appendix), which was the focus of the November 2023 Global Summit on AI safety.[57]

In April 2023, the Government announced £100 million in funding for a Foundation Model Taskforce to "ensure sovereign capabilities and broad adoption of safe and reliable Foundation Models." The Taskforce would bring together Government and industry experts. The Government specified this funding would be invested in "Foundation Model infrastructure and public service procurement, to create opportunities for domestic innovation".[58]

A few academics postulate that 'sovereign capabilities' could manifest in a range of ways, such as UK companies and researchers building a Foundation Model from scratch, adapting existing software to UK needs, and licensing Foundation Model technology from existing suppliers on 'suitable terms'.[59]

Following the November 2023 Global summit on AI safety, this Taskforce evolved to become the AI Safety Institute, a new UK-based global hub tasked with testing the safety of emerging types of AI.[25]

Training, developing, hosting and fine-tuning Foundation Models

Multiple forms of learning, including from labelled and unlabelled data can be used in the training stage (Box 1). Once trained and developed, the same Foundation Model can be shared and reused across many applications.[12,60]

Companies who develop Foundation Models can make them directly available to consumers and to other developers seeking to adapt the models to a specific application. Some models are private and hosted inside a company.[53] Some models (or parts of them) are made publicly available for anyone to download, modify, and distribute under a licence.[53]

Some models are hosted on cloud computing platforms and made accessible through a user interface.[53] A user interface allows other developers (who did not develop nor own the Foundation Model) and users to access and fine-tune, but not fundamentally modify the underlying Foundation Model.[53]

Fine-tuning a model involves developers training it further on a specific set of data to improve its performance for a specific application. Fine-tuning can often involve supervised or reinforcement learning (Box 1). For example, OpenAI has used reinforcement learning to fine-tune ChatGPT models and reduce inaccurate or harmful (such as violent or biased) outputs they generate.[61]

Debate around open-source AI models and concerns around access to the largest models

There is debate around whether AI models should be open-source. Although definitions vary, open-source often means the underlying code used to run AI models is freely available for testing, scrutiny and improvement.[62]

Open-source models can aid with transparency in how models work and often support scrutiny by a larger developer community who can play an important role in spotting biases, risks or faults.[63] Open-source models can also be tailored for specific user needs.[63]

In 2023, the UK Government released guidance on how public sector organisations could provide transparent information about the algorithmic tools they use, and why they are using them.[64]

However, some experts have said that in the specific case of Frontier AI (Appendix), having them open-source may increase the risk of misuse by malicious actors, such as cyber-attacks on national infrastructure.[62]

Only a few large private sector technology companies have developed Frontier models due to the scale of computing power and data required.[53]

A 2023 report by the Government Office for Science predicted that, in the near future, the development of Frontier AI models is highly likely to be carried out by a select few companies with the required resources, such as funding for computing power, skills and data.[63] These include OpenAI, Google, Anthropic and Meta.[63] The report also predicted that a few other companies with significant research and development budgets could enter the market in the next 18 months, such as Amazon and Apple.[63]

Due to high costs, concerns exist around the inaccessibility of developing Frontier models for small companies, open-source communities and academia, and the concentration of market power by a few private sector organisations.[15,59,65]

4 Factors driving advances in AI

Drivers of recent AI developments include greater availability of training data, increases in computing power, infrastructure investments and new uses of algorithms.[66,67]

4.1 Volume and quality of data

Increased data availability has allowed machine learning systems to be trained on larger and larger datasets.[66] Frontier Large Language Models have been trained on billions or even trillions of bits of data. For example, the large language model underpinning ChatGPT 3.5 (released to the public in November 2022) was trained using 300 billion words obtained from internet text.[68]

Research emphasises the importance of high-quality data for advancing AI capabilities, such as data that is correctly labelled, findable, accessible, reusable, explainable and un-biased.[69–72] See section 5.1 for concerns around using poor quality data to train machine learning systems.

4.2 Increasing computing power

The amount of computing power used to develop and run significant machine learning models has increased exponentially in the past half-decade.[73,74] For example, a report by the Centre for Security and Emerging Technology noted that a Foundation Model released in 2020 used 600,000 times more computing power than a noteworthy model in 2012.[75]

There have been environmental concerns around the increasing computing power needed (see section 5.2).

There have been improvements in the efficiency of computing resources and in machines working in parallel to share the load. These improvements have helped meet computing requirements of training and running large machine learning models, particularly Large Language Models.[73,76,77]

To train these models, large clusters of graphical processing units (similar to central processing units found in a typical home computer) are used with specialised 'accelerator' chips.[*] [80] These chips are capable of processing data across billions of units in parallel, which is particularly useful for Large Language Models, where traditional computer hardware is less capable of handling the vast amounts of data needed.

The large number of clusters of graphical processing units needed are expensive and scarce, making it difficult for businesses to acquire and maintain this hardware.

Therefore, much processing work occurs using cloud computing[80] involving the use of pooled computing resources provided by cloud companies to customers on-demand (see PN 629).

Inadequate computing power to advance AI capabilities further

Some reports suggest that it may not be feasible to increase computing power at its current rate past 2030, due to cost, limited supplies of 'accelerator' chips (see section 6.3), and technical difficulties, such as difficulties in managing large quantities of graphical processing units.[57,75] Computing power may be a barrier to further advancements in AI as a result.[75]

Some researchers are trying to develop AI algorithms that would require less computing power.[81–83] Experts suggest that future developments in more efficient 'accelerator' chips and quantum computing (PN 552) could improve computing power efficiency and the ability and speed of machine learning algorithms to process large amounts of data.[81,84–88]

The UK Government's National Quantum strategy, released in 2023, stated that "high performance computing is required in the UK to accelerate and steer the development of frontier AI" and that "over the next ten years, quantum computing will be an important addition to the UK's high performance computing ecosystem."[89]

[*] Graphical Processing Units (GPUs) have been used since the 1970s in gaming applications and have been designed to accelerate computer graphics and image processing.[78,79] In the past decade, GPUs have been increasingly applied in the training of large machine learning models after they were found to be effective in parallel processing.[78,79] Due to the high technical threshold and significant investment needed, companies such as Nvidia, Intel and Advanced Micro Devices hold the majority of GPU market shares.[79]

4.3 Computing infrastructure investments

In November 2022, the Independent Review of the Future of Compute commissioned by the UK Government said the UK ranked 10 internationally in terms of compute* capabilities.[90]

In March 2023 the UK Government announced a £900 million investment towards a new AI research resource and a new supercomputer, dubbed Isambard-AI, hosted at the University of Bristol. This will be several times more powerful than current UK computers and be one of Europe's most powerful supercomputers.[91,92] This supercomputer will be made up of thousands of "state-of-the-art" graphical processing units and "will be able to train the Large Language Models that are at the forefront of AI research and development today."[92] The funding will also go towards a new exascale computer at the University of Edinburgh.[93]

In November 2023, the UK Government announced it was increasing investment towards the AI research resource from £100 million to £300 million to boost British supercomputing 30-fold.[94]

In 2023, UK Research and Innovation (UKRI) also announced investments for advanced computing facilities. These include a £10 million UKRI award to a group of universities and industries, including the University of Bristol and Hewlett Packard Enterprise,[95] and a £30 million UKRI award to the Science and Technology Facilities Council's Daresbury Laboratory in Cheshire.[96]

4.4 New uses of algorithms

Two artificial neural network architectures called Transformers and generative adversarial networks (Appendix), developed in the past few years, have greatly improved generative AI.[97]

* Compute is defined by the Independent review as 'the systems assembled at scale to tackle computational tasks beyond the capabilities of everyday computers. This includes both physical supercomputers and the use of cloud provision to tackle high computational loads.'[90]

5 What is AI capable of?

5.1 Interpreting, processing and generating language

Whilst not all language related tasks require Large Language Models, many recent language processing capabilities (Appendix) are due to Large Language Model developments.

A single Large Language Model can be adapted to achieve many linguistic tasks, such as speech-to-text converters, online tools that summarise text, chatbots, speech recognition and translations.[12] As a result, language generated by Large Language Models is becoming more difficult to distinguish from language generated by humans.[12,98]

Limitations to using Large Language Models include their restricted ability to process linguistic differences[12,99], their lack of consistency in constructing accurate phrases,[12] their limited ability to understand contexts,[99] and the large resources required for training.[12,100]

5.2 Computer vision

Advances in AI have improved computer vision tasks (Appendix), such as object recognition, medical imaging analysis and navigation.[12,101]

These advances also have the potential to reduce the cost of training by making use of large quantities of data to understand the visual world.[12,101]

The training of computer vision capabilities can be very labour-intensive, time consuming and computationally complex as it has traditionally required expensive and carefully labelled data and supervised learning.[12,102]

5.3 AI use in robotics

AI is increasingly being integrated into robotics (Appendix), and are improving robots' abilities to learn, adapt, improve their performance over time, interact with their environments and perform complex tasks.[103,104] For example, robots are assisting surgeons during complex medical procedures with greater precision and accuracy.[105]

A longstanding challenge is giving robots the ability to handle the numerous conditions they will encounter in real-world settings, such as unexpected obstacles that can appear when driving.[12] Challenges to achieving this ability include collecting large quantities of data in the physical world that covers diverse environments and tasks, and ensuring the safety and robustness of such systems.[12]

6 Concerns

6.1 Data sources and management

AI systems require large volumes of data in order to be trained. The use of such large amounts of data has raised several concerns, including:[12,69]

- how this data is sourced, managed and shared
- licensing
- data quality
- biases in datasets and potential discrimination issues
- potential security and privacy issues
- and potential copyright issues (PN 633, upcoming PN on the Policy implications of AI). For example, generative AI could output copyrighted material present in its training data, leading to intellectual property rights issues (upcoming PN on the Policy implications of AI).

Data can be personal and non-personal. In the UK, The Data Protection Act 2018 and the UK GDPR regulate the collection and use of personal data (upcoming PN on the Policy Implications of AI).

Details about the training data for Large Language Models are subject to companies disclosing them.[106] Some articles suggest much of the data has come from of publicly available information on the web, such as Wikipedia or the discussion website Reddit.[68,106]

A few studies have found that future AI model capabilities could erode if they are trained on large proportions of AI-generated data, as AI generated data will become progressively less precise and diverse.[107,108] This may happen in the future as more internet content becomes AI generated.[107,109]

6.2 Impact on the environment

Research suggests that AI can both positively and negatively impact the environment.[74] For example, machine learning models can be used to optimise energy usage and improve the efficiency of logistics.[74] A 2019 report by PwC anticipated that applications of AI in energy, water, transport and agriculture could lead to a 4% reduction in greenhouse gas emissions by 2030.[110]

However, concerns exist around the environmental costs of training and running large AI models.[63,74,111,112] The amount of energy used in training large machine learning models depends on multiple factors such as the type of model, geographic location, the way data is processed, cloud computing, and the artificial neural network algorithms used.[106]

One academic study published in 2021 estimated that training ChatGPT-3 led to 1,287 MWh of energy consumption, which is equivalent to the annual energy consumption of around 477 average UK households[113,106] Researchers have found that more energy and carbon intensive tasks include generating new content compared to classifying tasks, and tasks involving images compared to those involving text alone.[114,115]

Increased model sizes have led to calls for developers to document and reduce their energy use (PN 677).[116,117] In 2023 some companies, such as Meta, released reports estimating the carbon footprint of their models to improve transparency.[118]

6.3 Supply of computing hardware

Some computing hardware, such as 'accelerator chips' (section 4.2), required to train and use AI is dependent on supply chains that are highly concentrated and at risk of disruption.[63,75] Cost changes or disruption to hardware or cloud computing could impact the training, use and deployment of AI models.[63]

6.4 Inaccurate results (hallucinations)

Large Language Models, such as ChatGPT, generate text by predicting the most likely words and phrases that go together based on patterns they have seen in training data. [119,120]

However, they are unable to identify if the phrases they generate make sense or are accurate.[121] This can sometimes lead to inaccurate results, also known as 'hallucination' effects, where Large Language Models generate plausible sounding but inaccurate text.[121,122] Hallucinations can also result from biases in training data or the model's lack of access to up-to-date information.[121]

Hallucinations can cause problems where the results of an AI are used to take decisions without proper consideration of the risk that the results are inaccurate.

This can be particularly problematic if individuals rely too heavily on the results. There is evidence to suggest that humans tend to favour automated decisions or advice,[123] which can lead to discriminatory outcomes or disinformation (upcoming PN on the policy implications of AI).

A 2023 academic review of literature suggested that hallucinations could be addressed by combining human judgement with AI evaluation systems, fine-tuning models, and improving the ability of AI to check data for biases and inaccuracies.[121]

6.5 Deepfakes, misinformation, disinformation and AI watermarks

AI systems can generate realistic text, images and videos. This can enable the creation of 'deepfakes': pictures and video that are deliberately altered to generate misinformation and disinformation.[103,124,125]

The UK Government defines disinformation as the "deliberate creation and spreading of false and/or manipulated information that is intended to deceive and mislead people, either for the purposes of causing harm, or for political, personal or financial gain". It defines misinformation as "the inadvertent spread of false information".[126]

These advances have copyright, disinformation, privacy, trust, crime and societal implications. There is concern that malicious actors will find it easier to produce disinformation at scale, with implications for public trust in online content, including election information (upcoming PN on the policy implications of AI).

AI watermarks can be used to embed a recognisable unique signal into AI generated content to identify it as such.[127] Such watermarks can protect against the spread of deepfakes and mis- and disinformation, and can also indicate authorship and establish authenticity of content.[127,128]

However, concerns include fake watermarks being added to content and that AI watermarks could be used to track an individual's use of generative AI, potentially compromising privacy.[127,129]

There are technical challenges to the introduction of AI watermarks, including:

- robustness (text watermarks can be easy to remove)
- watermarks working only on certain datasets (and therefore being limited to fine-tuned models)
- watermarks degrading the accuracy of AI generated outputs (for example, AI generated text emphasising certain words due to watermarks and therefore sounding un-natural)[127,129,130]

Some AI companies are researching how to produce robust AI watermarks.[130]

6.6 A lack of transparency

Some machine learning models, particularly those trained with deep learning, are so complex that it may be difficult or impossible to know how the model produced the output (PN 633).

The complexity arises because the model's decision is calculated through a path across billions of 'neurons' and interconnected layers in its network. This path is determined by how the model was trained. These models are commonly referred to as 'black box' machine learning models.[131]

This lack of transparency raises several concerns about the fairness, safety, reliability, liability and potential existential risks when using AI (upcoming PN on policy implications of AI),[131,132] particularly in high-risk scenarios such as healthcare (PN 637). For example, an individual adversely affected by an AI system may not know how it works, what went wrong, who is liable, and how to exercise any rights they may have (upcoming PN on policy implications of AI).

Approaches to improving how machine learning systems can be explained include designing systems using simpler methods and using tools to gain an insight into how complex systems function (PN 633).[133–135]

6.7 Implications for the economy

A range of bodies have published reports estimating the future impact of AI on the economy and on jobs (House of Commons Library briefing on the potential impact of AI on the labour market, upcoming PN on policy implications of AI).[1,9,136,137] In 2021, PwC published a report commissioned by the then Department for Business, Energy and Industrial Strategy that estimated that 7% of existing UK jobs could face a high probability of automation in the next 5 years, 18% in 10 years, and just under 30% after 20 years.[9,136] The research also reported that many jobs would be created through AI-related productivity and economic growth, such as in health and personal care.[9,136]

6.8 Lack of skills in the UK

Across the UK workforce, there is a growing demand for specialised skills in AI, machine learning, and data science (involving data collection, processing, storage, analysis and modelling) (PN 697). This demand could affect companies' capacity to use and apply AI in the future. The 2021 National AI Strategy recognised 'skills and talent' as core to UK sectors being able to apply AI.[21] The UK Government has launched several initiatives to develop specialised data skills, such as £117 million to train PhD students in AI at UK-based research organisations from 2024/25 (PN 697), and publishing guidance in November 2023 to support businesses upskilling employees so they can use AI to carry out tasks in the workplace.[138] A 2022 inquiry by the

House of Lords Science and Technology Committee raised concerns that there is a mismatch between the scale of the UK's Science Technology Engineering and Mathematics (STEM) skills gap and the solutions posed by the Government (PN 697).

6.9 Concerns around employment conditions for outsourced workers

Data used in training large machine learning models can be unlabelled or labelled. Labelling can be done automatically in some cases, or manually, either by developers and companies themselves, or through outsourcing.[53]

Concerns have been raised about the employment conditions of some of the outsourced workers involved in data labelling. For example, over the past few years, some AI companies have outsourced data labelling to workers in Kenya, who contributed to filtering toxic content for ChatGPT. This work led to widespread criticism about their working conditions, pay and the negative impact of the work on their mental health (upcoming PN on the Policy Implications of AI).[14,139–141]

7 Perceptions of AI

7.1 Public perceptions of AI applications

Over the past few years, a variety of research has been conducted by academia, industry, NGOs and the public sector to determine public understanding of, and opinions about, AI globally, in the UK, and in specific contexts (such as healthcare).[142–152] The findings can depend on study contexts and terminology used.

Key messages that have emerged from: in-depth interactions between members of the public, specialists, and policy makers (public dialogues);[150] a November 2022 survey by the Ada Lovelace Institute and the Alan Turing Institute of 4000 nationally representative adults in Britain;[145] and December 2023 survey results from the Centre for Data Ethics and Innovation and DSIT,[146] include:

- people often have questions about why the application was developed, whether it is necessary, and who benefits[150]
- high levels of awareness for visible AI applications such as facial recognition[145]
- self-reported awareness of AI increased significantly following the emergence of Large Language Models into public view in late 2022[146]
- perceptions of risks and benefits vary according to the way in which AI can be used in different applications[150] (section 3)
- AI applications in detecting cancer risks and other healthcare applications were seen as beneficial by most people[145]
- Key areas of concern:
  - displacement of jobs (particularly amongst non-graduates)[146]
  - impact on human creativity and problem-solving skills[146]
  - a potential negative impact on fairness of society[146]
  - the use of AI applications such as driverless cars and autonomous weapons[145].
- Members of the public with low digital familiarity felt less in control of their data relative to the overall UK population, however they were increasingly recognising the benefits of data use in society and trust accountability mechanisms for misuse[146]

7.2 Perceptions on future forms of AI

Experts have varying opinions on if, how and when Artificial General Intelligence and Superintelligence (Appendix) are achievable.

In 2023 Microsoft released a paper claiming some new Large Language Models have "more general intelligence than previous models."[153] Others dispute this claim.[154] Some experts say human intelligence and current AI are fundamentally different, such as AI being unable to learn abstract concepts, and cannot be compared.[98,154,155]

Risks and opportunities posed by future forms of AI depends on their capabilities, how they are used, geopolitics, access, ownership, safety measures and public attitudes.[63]

Appendix: Definitions of common terms

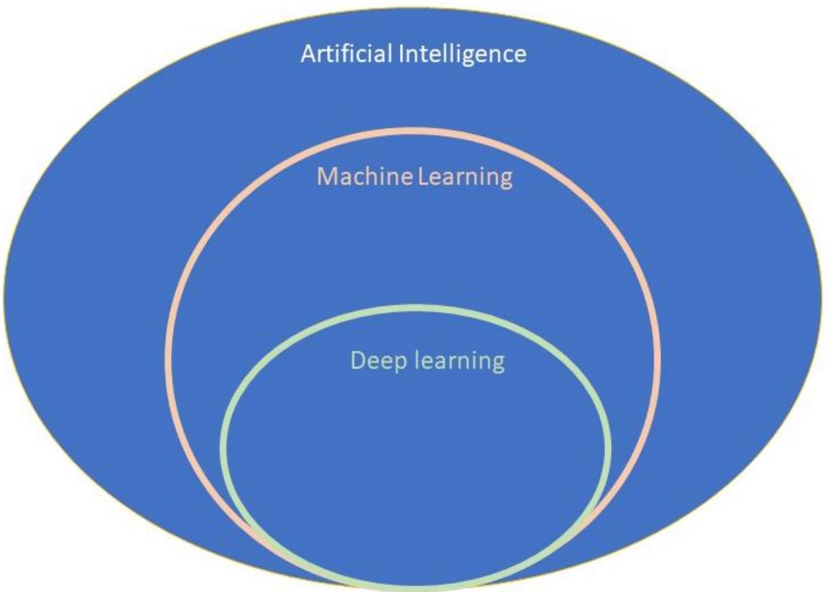
Definitions are not universally agreed, move at a fast pace, and are interlinked. In 2022, the International Organisation for Standardisation and International Electrotechnical Committee published a set of cross-sectorial definitions related to AI.[156]

AI types

- **Machine learning:** A type of AI that allows a system to learn and improve from examples without all its instructions being explicitly programmed (PN 633)

- **Deep learning:** A type of machine learning that uses artificial neural networks (see algorithms below) to recognise patterns in data and provide a suitable output, for example, a prediction.[6] Deep learning is suitable for complex learning tasks, and has improved AI capabilities in tasks such as voice and image recognition, object detection and autonomous driving (PN 633)[6]

Figure 1: AI types



Algorithms

- **Algorithm:** A set of instructions used to perform tasks (such as calculations and data analysis) usually using a computer or another smart device (PN 633)[42,43]

- **Artificial neural networks:** A computer structure inspired by the biological brain, consisting of a large set of interconnected computational units ('neurons') that are connected in layers. Data passes between these units as between neurons in a brain.[6] Outputs of a previous layer are used as inputs for the next (PN 633), and there can be hundreds of layers of units.[6] An artificial neural network with more than 3 layers is considered a deep learning algorithm.[157] Examples of artificial neural networks include Transformers or generative adversarial networks

- **Transformers:** Transformers have greatly improved natural language processing, computer vision and robotic capabilities and the ability of AI models to generate text.[97,158,159] A transformer can read vast amounts of text, spot patterns in how words and phrases relate to each other, and then make predictions about what word should come next. This ability to spot patterns in how words and phrases relate to each other is a key innovation, which has allowed AI models using transformer architectures to achieve a greater level of comprehension than previously possible[106]

- **Generative adversarial networks:** These are made up of two sub artificial neural networks: a generator network and a discriminator network. The generator network is fed sample data and generates artificial data based on patterns in sample data.[6] The discriminator network compares the artificially generated data with the 'real' sample data and feeds back to the generator network where it has detected differences.[6] The generator then alters its parameters. Over time the generator network learns to generate more realistic data, until the discriminator network cannot tell what is artificial and what is 'real' training data and the AI model generates the desired outcomes[6]

AI capabilities

- **Natural language processing:** This focuses on programming computer systems to understand and generate human speech and text.[99] Algorithms look for linguistic patterns in how sentences and paragraphs are constructed and how words, context and structure work together to create meaning.[6] Applications include speech-to-text converters, online tools that summarise text, chatbots, speech recognition and translations[6]

- **Computer vision:** This focuses on programming computer systems to interpret and understand images, videos and other visual inputs and take actions or make recommendations based on that information.[160] Applications include object recognition, facial recognition, medical imaging analysis, navigation and video surveillance[6]

- **Robotics:** Machines that are capable of automatically carrying out a series of actions and moving in the physical world.[6] Modern robots contain algorithms that typically, but do not always, have some form of artificial intelligence.[6] Applications include industrial robots used in manufacturing, medical robots for performing surgery, and self-navigating drones[6]

Machine learning models

- **Foundation Models:** A machine learning model trained on a vast amount of data so that it can easily be adapted for a wide range of general tasks, including being able to generate outputs (generative AI)[6]

- **Large Language Models:** A type of Foundation Model that is trained on vast amounts of text to carry out natural language processing tasks.[6] During training phases, Large Language Models learn parameters from factors such as the model size and training data. Parameters are then used by Large Language Models to infer new content.[161] Whilst there is no universally agreed figure for how large training datasets need to be, Large Language Models often have at least one billion or more parameters[161]

Types of AI outputs

- **Generative AI:** An AI model that generates text, images, audio, video or other media in response to user prompts. It uses machine learning techniques to create new data that has similar characteristics to the data it was trained on.[6] Generative AI applications include chatbots, photo and video filters, and virtual assistants

AI compared with human intelligence and values

- **Narrow AI:** Sometimes known as weak AI, these AI models are designed to perform a specific task (such as speech recognition) and cannot be adapted to other tasks[7,16,162]

- **General-purpose AI:** Often refers to AI models that can be adapted to a wide range of applications (such as Foundation Models)[163,164]

- **Frontier AI:** Defined by the UK Government as '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'.[57,63] Currently, this primarily encompasses a few Large Language Models (see section 2.2)

- **Artificial General Intelligence:** Sometimes known as General AI, Strong AI or Broad AI**,** this often refers to a theoretical form of AI that can achieve human-level or higher performance across most cognitive tasks[63,165]

- **Superintelligence:** A theoretical form of AI that has intelligence greater than humans and exceeds their cognitive performance in most domains[166]

- **Responsible AI:** Often refers to the practice of designing, developing, and deploying AI with certain values, such as being trustworthy, ethical, transparent, explainable, fair, robust and upholding privacy rights[13,167–169]

References and Contributors

References

1. Chui, M. *et al.* (2023). Economic potential of generative AI. McKinsey.
2. Sheikh, H. *et al.* (2023). Artificial Intelligence: Definition and Background. in *Mission AI: The New System Technology*. (eds. Sheikh, H. et al.) 15–41. Springer International Publishing.
3. Stone, P. *et al.* (2016). 'Artificial Intelligence and Life in 2030.' One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford University.
4. Monett, D. *et al.* (2018). Getting Clarity by Defining Artificial Intelligence—A Survey. in *Philosophy and Theory of Artificial Intelligence 2017*. (ed. Müller, V. C.) 212–214. Springer International Publishing.
5. O'Shaughnessy, M. (2022). One of the Biggest Problems in Regulating AI Is Agreeing on a Definition. Carnegie Endowment for International Peace.
6. The Alan Turing Institute (online). Data science and AI glossary.
7. Manning, C. (2022). Artificial Intelligence Definitions. Stanford University Human-Centred Artificial Intelligence.
8. Defence Science and Technology Laboratory (2020). The DSIT Biscuit Book: Artificial Intellience, Data Science and (mostly) Machine Learning.
9. Brione, P. *et al.* (2023). Potential impact of artificial intelligence on the labour market. House of Commons Library.
10. Lee *et al.* (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institute.
11. Kelan, E. K. (2023). Automation Anxiety and Augmentation Aspiration: Subtexts of the Future of Work. *Br. J. Manag.*, Vol 34, 2057–2074.
12. Bommasani, R. *et al.* (2022). On the Opportunities and Risks of Foundation Models. Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI).
13. Singh, S. (2023). What is responsible artificial intelligence and why do we need it? *Warwick Business School*.
14. JISC (2023). AI in tertiary education: A summary of the current state of play.
15. House of Commons Science, Innovation and Technology Committee (2023). The governance of artificial intelligence: interim report.
16. Rough, E. *et al.* (2023). Debate on Artificial Intelligence. House of Commons Library.

17. Tobin, J. (2023). Artificial intelligence: Development, risks and regulation. House of Lords Library.
18. Lacerda, L. (2022). AI, Surgery and Inequalities. Wellcome EPSRC Centre for Interventional Surgical Sciences.
19. University of Cambridge (online). Centre for the Study of Existential Risk.
20. Department for Science, Innovation and Technology *et al.* (2022). National AI Strategy - HTML version.
21. Department for Science, Innovation and Technology *et al.* (2022). National AI Strategy - AI Action Plan.
22. Department for Science, Innovation and Technology *et al.* (2023). UK Science and Technology Framework.
23. Department for Science, Innovation and Technology *et al.* (2023). A pro-innovation approach to AI regulation.
24. Department for Science, Innovation and Technology *et al.* (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.
25. Prime Minister's Office, 10 Downing Street (2023). Prime Minister launches new AI Safety Institute.
26. The White House (2023). FACT SHEET: Vice President Harris Announces New U.S. Initiatives to Advance the Safe and Responsible Use of Artificial Intelligence.
27. Department for Science, Innovation and Technology (2023). £54 million boost to develop secure and trustworthy AI research.
28. House of Commons Science, Innovation and Technology Committee (online). Governance of artificial intelligence (AI) inquiry.
29. House of Lords Science and Technology Committee (online). Large language models inquiry.
30. Javaid, M. *et al.* (2023). Understanding the potential applications of Artificial Intelligence in Agriculture Sector. *Adv. Agrochem*, Vol 2, 15–30.
31. Columbus, L. (2021). 10 Ways AI Has The Potential To Improve Agriculture In 2021. *Forbes*.
32. Jones, S. P. (2023). AI in education.
33. JISC (2023). Generative AI - a primer.
34. McMillan, L. *et al.* (2022). A review of the use of artificial intelligence methods in infrastructure systems. *Eng. Appl. Artif. Intell.*, Vol 116, 105472.
35. McMillan, L. *et al.* (2023). Flow Forecasting for Leakage Burst Prediction in Water Distribution Systems using Long Short-Term Memory Neural Networks and Kalman Filtering. *Sustain. Cities Soc.*, Elsevier BV.
36. Blake, K. *et al.* (2022). Machine learning in UK financial services. Bank of England.
37. Kwint (2023). Artificial intelligence: 10 promising interventions for healthcare. NIHR Evidence.
38. NICE (2023). Artificial intelligence technologies to speed up contouring in radiotherapy treatment planning | News | News.
39. Mock, M. *et al.* (2023). AI can help to speed up drug discovery — but only if we give it the right data. *Nature*, Vol 621, 467–470.
40. Chryssolouris, G. *et al.* (2023). Artificial Intelligence in Manufacturing Processes. in *A Perspective on Artificial Intelligence in Manufacturing*. (eds. Chryssolouris, G. et al.) 15–39. Springer International Publishing.

41. Kelan, E. K. (2023). Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. *Hum. Resour. Manag. J.*, Vol n/a,
42. Dasgupta, S. *et al.* (2006). Algorithms. McGraw-Hill Education.
43. Bruderer, H. (2018). Algorithms Have Been Around for 4,000 Years.
44. Delua, J. (2021). Supervised vs. Unsupervised Learning: What's the Difference? *IBM*.
45. Solatidehkordi, Z. *et al.* (2022). Survey on Recent Trends in Medical Image Classification Using Semi-Supervised Learning. *Appl. Sci.*, Vol 12, 12094. Multidisciplinary Digital Publishing Institute.
46. Huynh, T. *et al.* (2022). Semi-supervised learning for medical image classification using imbalanced training data. *Comput. Methods Programs Biomed.*, Vol 216, 106628.
47. Jiao, R. *et al.* (2022). Learning with Limited Annotations: A Survey on Deep Semi-Supervised Learning for Medical Image Segmentation. arXiv.
48. Bai, Y. *et al.* (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv.
49. Ouyang, L. *et al.* (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.*, Vol 35, 27730–27744.
50. Brooks, R. (2021). What is reinforcement learning? *University of York*.
51. TechTarget (2023). What is Reinforcement Learning?
52. OpenAI (online). Fine-tuning.
53. Jones, E. (2023). Explainer: What is a foundation model? Ada Lovelace Institute.
54. Bommasani, R. *et al.* (2023). AI Accountability Policy Request for Comment. Stanford University Human-Centred Artificial Intelligence.
55. Imran, M. (2020). 21 Best R Machine Learning Packages in 2021 - The Ultimate Guide.
56. McFarland, A. (2022). 10 Best Python Libraries for Machine Learning & AI.
57. Department for Science, Innovation and Technology (2023). Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk.
58. Department for Science, Innovation and Technology *et al.* (2023). Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI.
59. Girolami, M. *et al.* (2023). A sovereign AI capability for the UK. *Found. Sci. Technol. J.*, Vol 23,
60. Schneider, J. (2022). Foundation models in brief: A historical, socio-technical focus. arXiv.
61. OpenAI (2022). Introducing ChatGPT.
62. Seger, E. *et al.* (2023). Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. Centre for the Governance of AI.
63. Government Office for Science (2023). Future Risks of Frontier AI.
64. Central Digital and Data Office *et al.* (2023). Algorithmic Transparency Recording Standard Hub.
65. Government Office for Science (2023). Rapid Technology Assessment: Artificial Intelligence.
66. The Royal Society (2017). Machine learning: the power and promise of computers that learn by example.
67. Amodei, D. *et al.* (2018). AI and compute. *OpenAI*.

68. Hughes, A. (2023). ChatGPT: Everything you need to know about OpenAI's GPT-4 tool. *BBC Science Focus Magazine*.
69. Mazumder, M. *et al.* (2023). DataPerf: Benchmarks for Data-Centric AI Development. arXiv.
70. Yoon, J. *et al.* (2020). Estimating the Impact of Training Data with Reinforcement Learning. *Google*.
71. Toneva, M. *et al.* (2019). An Empirical Study of Example Forgetting during Deep Neural Network Learning. arXiv.
72. Ravi, N. *et al.* (2022). FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. *Sci. Data*, Vol 9, 657.
73. Zhu, S. *et al.* (2023). Intelligent Computing: The Latest Advances, Challenges, and Future. *Intell. Comput.*, Vol 2, 0006. American Association for the Advancement of Science.
74. Maslej, N. *et al.* (2023). AI Index Report 2023. AI Index Steering Committee, Institute for Human-Centered AI.
75. Lohn, A. J. *et al.* (2022). AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress? Centre for Science and Emerging Technology.
76. The Alan Turing Institute (2019). Co-designing computing.
77. Khan, S. M. *et al.* (2020). AI Chips: What They Are and Why They Matter. Centre for Science and Emerging Technology.
78. Pandey, M. *et al.* (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.*, Vol 4, 211–221. Nature Publishing Group.
79. Peddie, J. (2023). Introduction. in *The History of the GPU - Steps to Invention*. (ed. Peddie, J.) 1–30. Springer International Publishing.
80. Härlin, T. *et al.* Exploring opportunities in the gen AI value chain. McKinsey.
81. Morrison, R. (2022). Compute power is becoming a bottleneck for developing AI. Here's how you clear it. *Tech Monitor*.
82. Kim, S. *et al.* (2023). SqueezeLLM: Dense-and-Sparse Quantization. arXiv.
83. Human Brain Project (2023). Learning from the brain to make AI more energy-efficient.
84. Chauhan, V. *et al.* (2022). Quantum Computers: A Review on How Quantum Computing Can Boom AI. in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. 559–563.
85. Bova, F. *et al.* (2021). Quantum Computing Is Coming. What Can It Do? *Harvard Business Review*.
86. Biamonte, J. *et al.* (2017). Quantum machine learning. *Nature*, Vol 549, 195–202. Nature Publishing Group.
87. Abdelgaber, N. *et al.* (2020). Overview on Quantum Computing and its Applications in Artificial Intelligence. in *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 198–199.
88. Gill, S. S. *et al.* (2022). AI for next generation computing: Emerging trends and future directions. *Internet Things*, Vol 19, 100514.
89. Department for Science, Innovation and Technology (2023). National quantum strategy.
90. Department for Science, Innovation and Technology (2022). Independent Review of the Future of Compute.

91. Department for Science, Innovation and Technology *et al.* (2023). Government commits up to £3.5 billion to future of tech and science.
92. Department for Science, Innovation and Technology *et al.* (2023). Bristol set to host UK's most powerful supercomputer to turbocharge AI innovation. *GOV.UK*.
93. The University of Edinburgh (2023). Edinburgh to lead new era of UK supercomputing.
94. Department for Science, Innovation and Technology (2023). Technology Secretary announces investment boost making British AI supercomputing 30 times more powerful.
95. University of Bristol (2023). Isambard 3 receives £10 million investment creating one of world's TOP500 supercomputers. University of Bristol.
96. UKRI (2023). New £30 million supercomputer centre at Daresbury Laboratory.
97. Fui-Hoon Nah, F. *et al.* (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *J. Inf. Technol. Case Appl. Res.*, Vol 25, 277–304. Routledge.
98. Nathan, A. *et al.* (2023). Generative AI: Hype or Truly Transformative? *Goldman Sachs*.
99. Khurana, D. *et al.* (2023). Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.*, Vol 82, 3713–3744.
100. Baktash, J. A. *et al.* (2023). Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing. arXiv.
101. Gonog, L. *et al.* (2019). A Review: Generative Adversarial Networks. in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. 505–510.
102. Elasri, M. *et al.* (2022). Image Generation: A Review. *Neural Process. Lett.*, Vol 54, 4609–4646.
103. Littman, M. L. *et al.* (2021). Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. Stanford University.
104. Soori, M. *et al.* (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cogn. Robot.*, Vol 3, 54–70.
105. Panesar, S. *et al.* (2019). Artificial Intelligence and the Future of Surgical Robotics. *Ann. Surg.*, Vol 270, 223.
106. Nield, D. How ChatGPT and Other LLMs Work—and Where They Could Go Next. *Wired UK*.
107. Alemohammad, S. *et al.* (2023). Self-Consuming Generative Models Go MAD. arXiv.
108. Doshi, A. R. *et al.* (2023). Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content.
109. Shumailov, I. *et al.* (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv.
110. Herweijer, C. *et al.* (2019). How AI can enable a sustainable future. PwC.
111. Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, Vol 2, 423–425. Nature Publishing Group.
112. OECD (2022). Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. Vol 341,
113. Ofgem (2023). Average gas and electricity usage.

114. Luccioni, A. S. *et al.* (2023). Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv.
115. Heikkilä, M. (2023). Making an image with generative AI uses as much energy as charging your phone. *MIT Technology Review*.
116. Strubell, E. *et al.* (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv.
117. Kaack, L. H. *et al.* (2022). Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Change*, Vol 12, 518–527. Nature Publishing Group.
118. Touvron, H. *et al.* (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv.
119. Lee, T. B. *et al.* (2023). Large language models, explained with a minimum of math and jargon.
120. Wolfram, S. (2023). What Is ChatGPT Doing … and Why Does It Work?
121. Rawte, V. *et al.* (2023). A Survey of Hallucination in Large Foundation Models. arXiv.
122. McGowan, A. *et al.* (2023). ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res.*, Vol 326, 115334.
123. Alon-Barkat, S. *et al.* (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *J. Public Adm. Res. Theory*, Vol 33, 153–169.
124. van der Sloot, B. *et al.* (2022). Deepfakes: regulatory challenges for the synthetic society. *Comput. Law Secur. Rev.*, Vol 46, 105716.
125. Masood, M. *et al.* (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.*, Vol 53, 3974–4026.
126. Cabinet Office (2023). Fact Sheet on the CDU and RRU.
127. Craig, L. (2023). AI watermarking. *TechTarget*.
128. Grinbaum, A. *et al.* (2022). The Ethical Need for Watermarks in Machine-Generated Language. arXiv.
129. Barrett, C. *et al.* (2023). Identifying and Mitigating the Security Risks of Generative AI. arXiv.
130. Heikkilä, M. (2023). Google DeepMind has launched a watermarking tool for AI-generated images. *MIT Technology Review*.
131. The Royal Society (2019). Explainable AI: the basics.
132. Cassauwers, T. (2020). Opening the 'black box' of artificial intelligence. *Horizon - The EU Research & Innovation Magazine, European Commission*.
133. Anthropic (2023). Decomposing Language Models Into Understandable Components.
134. Saranya, A. *et al.* (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.*, Vol 7, 100230.
135. Hassija, V. *et al.* (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.*,
136. PwC (2021). The Potential Impact of Artificial Intelligence on UK Employment and the Demand for Skills.
137. International Trade Administration (2023). United Kingdom Artificial Intelligence Market 2023.
138. Department for Science, Innovation and Technology *et al.* (2023). New business guidance to boost skills and unlock benefits of AI.
139. BBC News (2023). Kenyan AI worker traumatised from data labelling.

140. Rowe, N. (2023). ‘It’s destroyed me completely’: Kenyan moderators decry toll of training of AI models. *The Guardian*.
141. Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philos. Technol.*, Vol 36, 15.
142. Bristows (2018). Artificial Intelligence: Public Perception, Attitude and Trust.
143. Kelley, P. G. *et al.* (2021). Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 627–637.
144. Neudert, L.-M. *et al.* (2020). Global Attitudes Towards AI, Machine Learning & Automated Decision Making. Oxford Internet Institute, University of Oxford.
145. The Ada Lovelace Institute (2023). How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain.
146. The Centre for Data Ethics and Innovation (2022). Public attitudes to data and AI: Tracker survey.
147. Wu, C. *et al.* (2023). Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. *BMJ Open*, Vol 13, e066322. British Medical Journal Publishing Group.
148. NHS England (online). Surveying public perceptions of AI. *NHS Transformation Directorate*.
149. Ipsos *et al.* (2022). NHS AI Lab Public Dialogue on Data Stewardship.
150. The Royal Society (2018). Portrayals and perceptions of AI and why they matter.
151. The Royal Society (2017). Machine learning: the power and promise of computers that learn by example.
152. Office for National Statistics (2023). Public awareness, opinions and expectations about artificial intelligence: July to October 2023.
153. Bubeck, S. *et al.* (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv.
154. Metz, C. (2023). Microsoft Says New A.I. Shows Signs of Human Reasoning. *The New York Times*.
155. Hasan, M. *et al.* (2023). Relationship between Artificial Intelligence and Human Intelligence.
156. International Organization for Standardization *et al.* (2022). Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.
157. IBM (2023). AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the difference?
158. Vanhoucke, V. (2023). Speaking robot: Our new AI model translates vision and language into robotic actions. *Google*.
159. ADEPT (2022). ACT-1: Transformer for Actions.
160. IBM (online). What is Computer Vision?
161. Kerner, S. M. (2023). What are Large Language Models? *TechTarget*.
162. IBM (online). What is Strong AI?
163. Küspert, S. *et al.* The value chain of general-purpose AI. Ada Lovelace Institute.
164. Future of life Institute (2022). General Purpose AI and the AI Act.
165. Morris, M. R. *et al.* (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv.

166. Müller, V. C. *et al.* (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. in *Fundamental Issues of Artificial Intelligence*. (ed. Müller, V. C.) Vol 376, 555–572. Springer International Publishing.
167. Herrmann, H. (2023). What's next for responsible artificial intelligence: a way forward through responsible innovation. *Heliyon*, Vol 9, e14379.
168. Schindler *et al.* (2023). 10 things governments should know about responsible AI. IBM.
169. Barredo Arrieta, A. *et al.* (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, Vol 58, 82–115.

Contributors

For further information on this subject, please contact Devyani Gajjar. POST would like to thank interviewees and peer reviewers for kindly giving up their time during the preparation of this briefing, including:

Members of the POST Board*

Dr David Busse, Government Office for Science*

Matt Davies, Ada Lovelace Institute*

Dr Yali Du, Kings College London*

Dr Gordon Fletcher, University of Salford

Dr Matthew Forshaw, Newcastle University and The Alan Turing Institute*

Professor Oliver Hauser, University of Exeter*

Elliot Jones, Ada Lovelace Institute*

Dr Clara Martins-Pereira, Durham University*

Dr Shweta Singh, University of Warwick and The Alan Turing Institute*

Adam Leon Smith, British Computing Society, Chair of the Fellows Technical Advisory Group*

Professor Michael Wooldridge, University of Oxford and The Alan Turing Institute

*Denotes people and organisations who acted as external reviewers of the briefing.

The Parliamentary Office of Science and Technology (POST) is an office of both Houses of Parliament. It produces impartial briefings designed to make scientific information and research accessible to the UK Parliament. Stakeholders sometimes contribute to and review POSTbriefs. POST is grateful to these contributors.

POST's published material is available to everyone at post.parliament.uk. Get our latest research delivered straight to your inbox. Subscribe at post.parliament.uk/subscribe.

post.parliament.uk

@POST_UK