Department for
Science, Innovation
& Technology

AI SAFETY
SUMMIT
HOSTED BY THE UK
1-2 NOVEMBER 2023

Capabilities and risks from frontier AI

A discussion paper on the need for further research into AI risk

October 2023

Acknowledgements

We would like to thank the expert review panel, Yoshua Bengio, Sara Hooker, Arvind Narayanan, William Isaac, Paul Christiano, Irene Solaiman, Alexander Babuta and John McDermid for their insightful comments and feedback.

This report is a discussion paper to support the AI Safety Summit, and does not represent a policy position of HMG or represent the views of the expert review panel above, who only provided comments for consideration.

Contents

Introduction

We are in the midst of a technological revolution that will fundamentally alter the way we live, work, and relate to one another. Artificial Intelligence (AI) promises to transform nearly every aspect of our economy and society. The opportunities are transformational - advancing drug discovery, making transport safer and cleaner, improving public services, speeding up and improving diagnosis and treatment of diseases like cancer and much more.

Developments in frontier AI are transforming productivity and software services, which will multiply the productivity of many industries and sectors.[1] This progress in frontier AI in recent years has been rapid, and the most advanced systems can write text fluently and at length, write well-functioning code from natural language instructions, make new apps, score highly on school exams, generate convincing news articles, translate between many languages, summarise lengthy documents, amongst other capabilities. The opportunities are vast, and there is great potential for increasing the productivity of workers of all kinds.

However, these huge opportunities come with risks that could threaten global stability and undermine our values. To seize the opportunities, we must understand and address the risks. AI poses risks in ways that do not respect national boundaries. It is important that governments, academia, businesses, and civil society work together to navigate these risks, which are complex and hard to predict, to mitigate the potential dangers and ensure AI benefits society.

The UK Government believes more research into AI risk is needed. This report explains why. It describes the current state and key trends relating to frontier AI capabilities, and then explores how frontier AI capabilities might evolve in the future and reviews some key risks. There is significant uncertainty around both the capabilities and risks from AI, including some experts who believe that some of these risks are overstated. This report focuses on evidence for risks and concludes that doing further research is necessary.

This report covers many risks, but we wish to emphasise that the overarching risk is a loss of trust in and trustworthiness of this technology which would permanently deny us and future generations its transformative positive benefits. In discussing the other risks, we do so in order to galvanize action to mitigate them, such that we can capture the full benefits of frontier AI.

Defining AI is challenging as it remains a quickly evolving technology. For the purposes of the Summit we define "frontier AI" 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* (see Figure 1).[2] Today, this primarily includes large language models (LLMs)[3] such as those underlying ChatGPT,[4] Claude,[5] and Bard.[6] However, it is important to note that, both today and in the future, frontier AI systems may not be underpinned by LLMs, and could be underpinned by another technology.

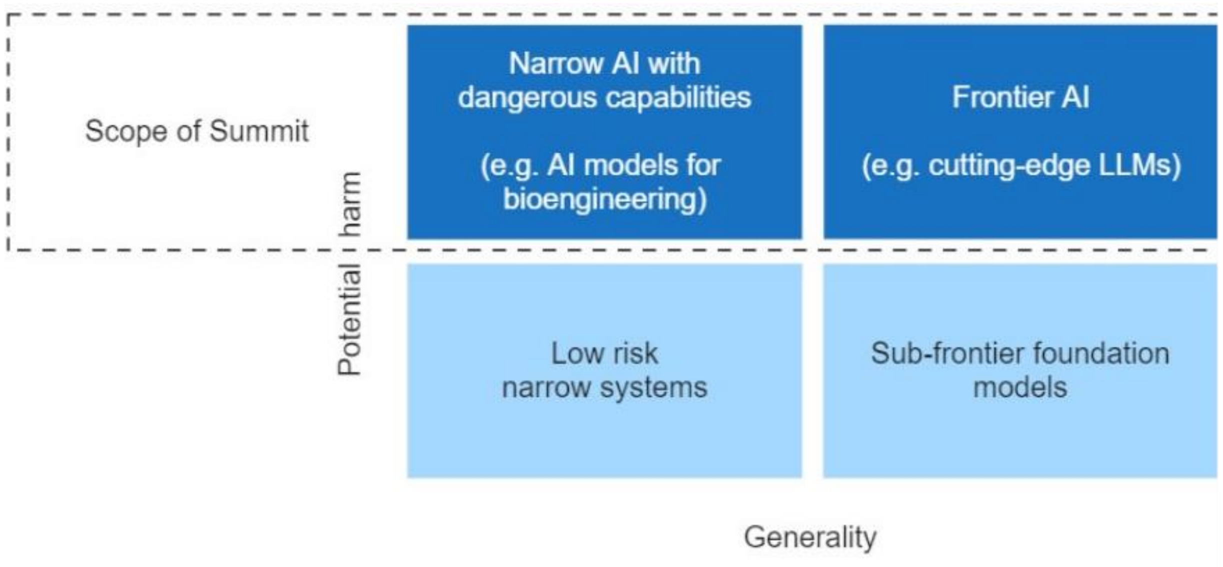

Scope of Summit
Narrow AI with dangerous capabilities
(e.g. AI models for bioengineering)
Frontier AI
(e.g. cutting-edge LLMs)
Low risk narrow systems
Sub-frontier foundation models
Potential harm
Generality

Figure 1: Scope of the AI Safety Summit - 2023

The limited focus of this report means we do not cover powerful narrow AI[7] systems like AlphaGo, AlphaFold or DALL·E 3 which cannot perform as wide a variety of tasks.[8]

There are already a number of existing international efforts and initiatives which touch upon the capabilities and risks of frontier AI. The upcoming AI Safety Summit will provide space for a focused and deep discussion on AI safety at the frontier and what further action needs to be taken, complementing existing initiatives, and this report is intended to be a resource for all.

This report is by no means conclusive; there are many risks we omit and we encourage readers to view it as the start of a conversation.

What is the current state of frontier AI capabilities?

Frontier AI can perform a wide variety of tasks, is being augmented with tools to enhance its capabilities, and is being increasingly integrated into systems that can have a wide impact on the economy and society. Although these models still have major limitations such as their factuality and reliability, their current capabilities are impressive, may be greater than we have been able to assess, and have appeared faster than we expected.

How frontier AI works

Frontier AI companies such as OpenAI, DeepMind and Anthropic develop large language models (LLMs) such as GPT-4 in two phases: pre-training and fine-tuning.

During pre-training, an LLM “reads” millions or billions of text documents.[9] As it reads, word by word,[10] it predicts what word will come next. At the start of pre-training it predicts randomly, but as it sees more data it learns from its mistakes and improves its predictive performance. Once pre-training is over, the model is significantly better than humans at predicting the next word of a randomly chosen text document.[11]

During fine-tuning,[12] the pre-trained AI is further trained on highly curated datasets, which are focused on more specialised tasks, or are structured to direct model behaviour in ways which are in alignment with developer values and user expectations [13]

Increasingly, frontier AI models are multi-modal. In addition to text, they can generate and process other data types such as images, video, and sound.[14]

The key inputs to development are computational resources ("compute"[15]) to train and run the model, data for it to learn from, the algorithms that define this training process, and talent and expertise that enable all of this.[16] The vast majority of compute is spent on pre-training, which is when most core capabilities are learnt by a model.[17]

The total development costs for the most capable frontier AI models today runs into the tens of millions of pounds,[18] with costs expected to soon reach into the hundreds of millions or even billions of pounds.[19] While the best performing models are developed by a small number of well-resourced organisations, a larger number of smaller entities build products on top of these frontier models for specific markets.[20]

The below diagram outlines the inputs to, and stages of, the development and deployment of frontier AI.

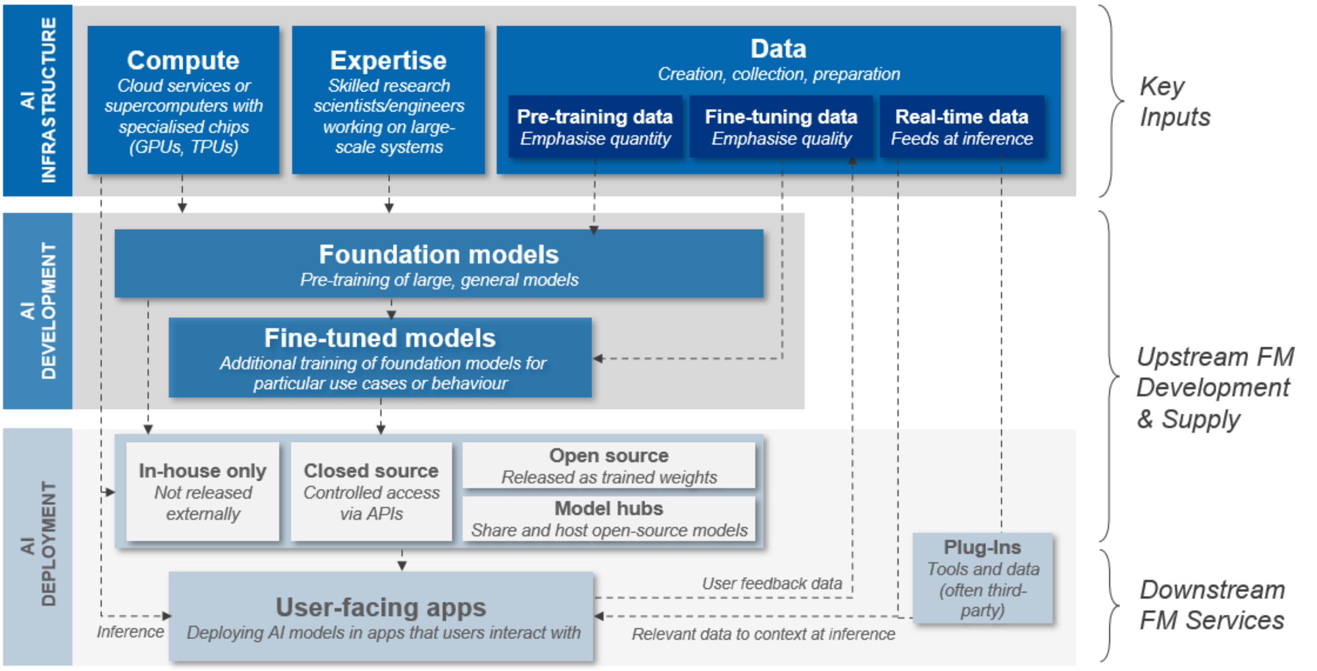

AI INFRASTRUCTURE
Compute
Cloud services or supercomputers with specialised chips (GPUs, TPUs)
Expertise
Skilled research scientists/engineers working on large-scale systems
Data
Creation, collection, preparation
Pre-training data
Emphasise quantity
Fine-tuning data
Emphasise quality
Real-time data
Feeds at inference
Key Inputs
AI DEVELOPMENT
Foundation models
Pre-training of large, general models
Fine-tuned models
Additional training of foundation models for particular use cases or behaviour
Upstream FM Development & Supply
AI DEPLOYMENT
In-house only
Not released externally
Closed source
Controlled access via APIs
Open source
Released as trained weights
Model hubs
Share and host open-source models
Plug-Ins
Tools and data (often third-party)
User-facing apps
Deploying AI models in apps that users interact with
Inference
User feedback data
Relevant data to context at inference
Downstream FM Services

Figure 2. An overview of foundation model development, training and deployment. From AI Foundation Models: initial review, CMA, 2023.

Frontier AI can perform many economically useful tasks

Simply from being trained to predict the next word across diverse datasets, models develop sophisticated capabilities.[21] For example, frontier AI can (with varying degrees of success and reliability):

- Converse fluently and at length, drawing on extensive information contained in training data.
- Write long sequences of well-functioning code from natural language instructions, including making new apps.[22]
- Score highly on high-school and undergraduate examinations in many subjects.[23]
- Generate plausible news articles.[24]
- Creatively combine ideas together from very different domains.[25]
- Explain why novel sophisticated jokes are funny.[26]
- Translate between multiple languages.[27]
- Direct the activities of robots via reasoning, planning and movement control.[28]
- Analyse data by plotting graphs and calculating key quantities.[29]
- Answer questions about images that require common-sense reasoning.[30]
- Solve maths problems from high-school competitions.[31]
- Summarise lengthy documents.[32]

These capabilities show potential to be applied across a wide array of economic use-cases. In addition to some of the applications above, frontier AI has been used to:

- Improve the performance of leading consultants in developing go-to-market plans.[33]
- Automate a wide variety of legal work.[34]
- Support leading wealth managers.[35]
- Increase the productivity of call-centre workers.[36]
- Accelerate academic research, for example in economics.[37]

Annex A provides more detail on AI capabilities in content creation, computer vision, theory of mind, memory, mathematics, physical intuition, and robotics.

Frontier AI models can be augmented with tools to make them more autonomous

Frontier AI models are more useful when augmented with other tools and software.

Frontier *AI models*, before they are augmented, respond to a request simply by producing a snippet of text. By contrast, autonomous[38] *AI agents*[39] can take long sequences of actions in pursuit of a goal, without requiring human involvement.

Researchers have built software programs called “scaffolds”[40] that allow frontier AI models to power autonomous AI agents. The scaffold prompts the AI model to create a plan for achieving a high-level goal and to then execute the plan step by step. The scaffold augments the AI model with tools like web browsers, allowing it to execute each step autonomously. The resultant system, built out of the AI model and the scaffold, is an AI agent. AutoGPT is the most well-publicised example of such an AI agent as of late 2023.[41]

Today’s AI agents currently struggle to perform most tasks – they often get stuck in loops and cannot self-correct, or fail at crucial steps. However, they do allow frontier AI to perform some entirely new tasks. Examples of tasks that AI agents can currently do include:

- Find specific information by browsing the internet.[42]
- Organise parties in simulated ‘The Sims’-like environments.[43]
- Solve complex problems in open-world survival games like Minecraft[44] and Crafter[45].
- Support the synthesis of chemicals by searching the web for relevant information and writing code to operate robotic hardware.[46]

Many leading AI researchers and companies explicitly aim to build AI agents whose general capabilities would exceed those of humans.[47]

Frontier AI could be more capable than evaluations indicate

Researchers and users frequently uncover surprising capabilities for frontier AI models which pre-deployment evaluation did not uncover.[48]

The capabilities of frontier AI models are likely to be further enhanced in many ways in the future, such as through:

- **Better prompts.**[49] The way that a question is phrased can significantly affect a frontier AI system’s response. For example, encouraging a model to think through its answer “step by step” significantly improves performance on maths and logic problems.[50]
- **Better tools**. Frontier AI models can be trained to use tools like web browsers, calculators, knowledge databases, or robot actuators, and can competently use entirely new tools when provided text instructions on how to use them[51]. These tools and resources can significantly improve capabilities at relevant tasks or endow them with entirely novel capabilities, such as the ability to directly manipulate physical systems.[52]
- **Better scaffolds.** Scaffolding software programs (“scaffolds”) structure the information flow of an AI model, leaving the model itself unchanged.[53] Better scaffolds could, for example, help an AI agent self-correct when they have made a mistake,[54] or improve their long-term memory.
- **New fine-tuning data.** Fine-tuning on high-quality data can significantly improve AI capabilities in a given domain, at a tiny fraction of the cost of pre-training.

- **Team-work between AI systems.** Multiple different AI systems, including both narrow models and more general models, could collaborate to perform tasks.[55]

Unlike pre-training, these improvements do not require significant computational resources and so **a wide range of actors could cheaply improve frontier AI capabilities**, provided they have easy access to pre-trained models**.**

Limitations of frontier AI

There is ongoing debate about the limitations of frontier AI systems, including whether their performance is driven more by general reasoning or by a combination of memorisation and following basic heuristics[56].

General reasoning abilities are evidenced by frontier AI producing remarkably apt responses to novel questions, For example, PaLM's ability to understand the humour behind jokes which had never before been told.[57]

However, there is also evidence that models rely heavily on memorisation and basic heuristics:

- LLMs perform less well when a question is reworded to make it different from text that is in their training data.[58]
- LLMs often solve complex problems using overly-simple heuristics that would fail to solve other similar problems.[59]
- There are instances where LLMs fail to apply information from their training data in very basic ways.[60]

Beyond an uncertain ability to generalise to new contexts, other key limitations of current frontier AI models include:

- **Hallucinations:** AI systems regularly produce plausible yet incorrect answers and state these answers with high confidence.[61] This might be addressed by systems using knowledge repositories,[62] improved fine-tuning, or new methods for teaching the model what it does and does not know.
- **Coherence over extended durations**: AI models are less reliable on tasks that require long-term planning or taking a large number of sequential steps (e.g. writing a novel).[63] This is partially due to their restricted context length and the scarcity of long-duration task training data.[64] These limitations might be addressed by algorithmic innovations to give AI a source of long-term memory, creating more data on long-horizon tasks, better scaffolds that help AI agents spot and correct their own errors,[65] or improved techniques for breaking long tasks into multiple small steps[66].
- **Lack of detailed context:** Many tasks in the real economy require extensive context about a particular company, project, or code-base. Current frontier systems are generically competent, but lack this specific context and cannot learn it from the available data. This might be addressed by access to additional private data sources, new data generation techniques, more data-efficient fine-tuning techniques, new "model-based" learning methods,[67] or simply by increasing the compute and data used to develop the system.

It remains uncertain how these limitations will evolve. Some argue that these limitations will permanently limit frontier AI development in certain applications. On the other hand, recent progress in AI has greatly surpassed expert predictions in many domains, while underperforming in other areas.[68]

How might frontier AI capabilities improve in the future?

Recent AI progress has been rapid and will likely continue. This is due to predictable improvements in the performance of frontier AI models when developed with more compute, more data and better algorithms. Unexpected new capabilities may also emerge. Advanced general-purpose AI agents could be developed in the not too distant future – although this is a subject of debate, especially regarding the timing.

Recent AI progress has been rapid

The recent pace of AI progress has surprised forecasters and machine learning experts alike.[69] Problems that frustrated the AI community for decades have rapidly fallen to ever-more-capable models.

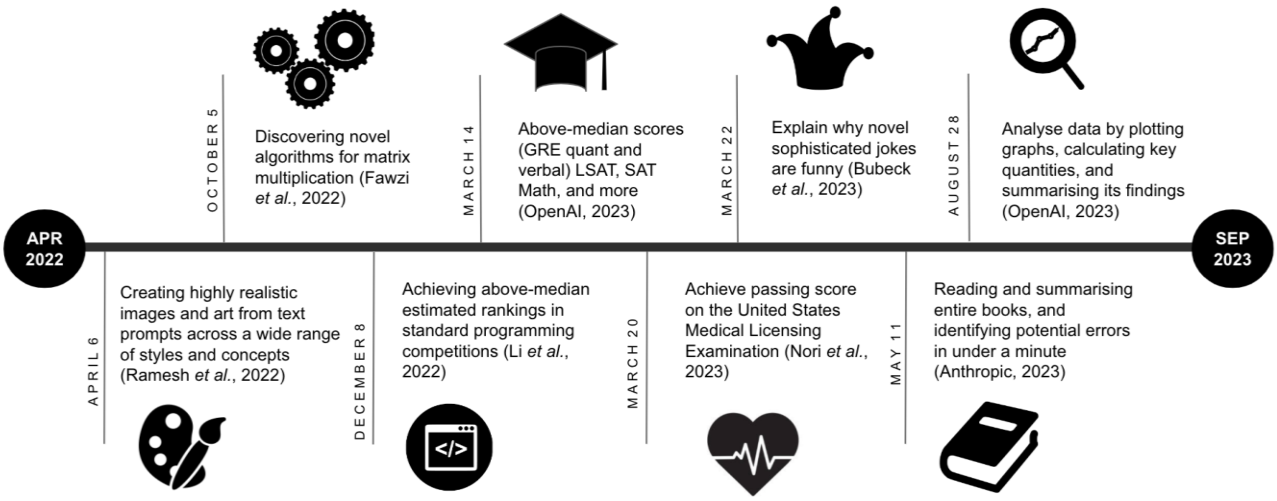

APR 2022
APRIL 6
Creating highly realistic images and art from text prompts across a wide range of styles and concepts (Ramesh et al., 2022)
OCTOBER 5
Discovering novel algorithms for matrix multiplication (Fawzi et al., 2022)
DECEMBER 8
Achieving above-median estimated rankings in standard programming competitions (Li et al., 2022)
MARCH 14
Above-median scores (GRE quant and verbal) LSAT, SAT Math, and more (OpenAI, 2023)
MARCH 20
Achieve passing score on the United States Medical Licensing Examination (Nori et al., 2023)
MARCH 22
Explain why novel sophisticated jokes are funny (Bubeck et al., 2023)
MAY 11
Reading and summarising entire books, and identifying potential errors in under a minute (Anthropic, 2023)
AUGUST 28
Analyse data by plotting graphs, calculating key quantities, and summarising its findings (OpenAI, 2023)
SEP 2023

Figure 3. An overview of notable AI achievements from 2022-2023 across diverse domains, Epoch 2023

Recent advances in frontier AI are the continuation of a longer-running trend: the rapid progress since 2012 in its parent field of deep learning across computer vision, game playing, and language modelling.[70] In 2014, AI could only generate simple, blurry images. However, by 2022, models like DALL-E 2 and Imagen could generate high-quality, creative images from text prompts (see figure 4a). Substantial advances were seen in the shift from GPT-3.5 to GPT-4, released just months apart. For example, on calculus questions GPT-3.5 scored below most humans, but GPT-4 improved significantly and scored around the median human level.

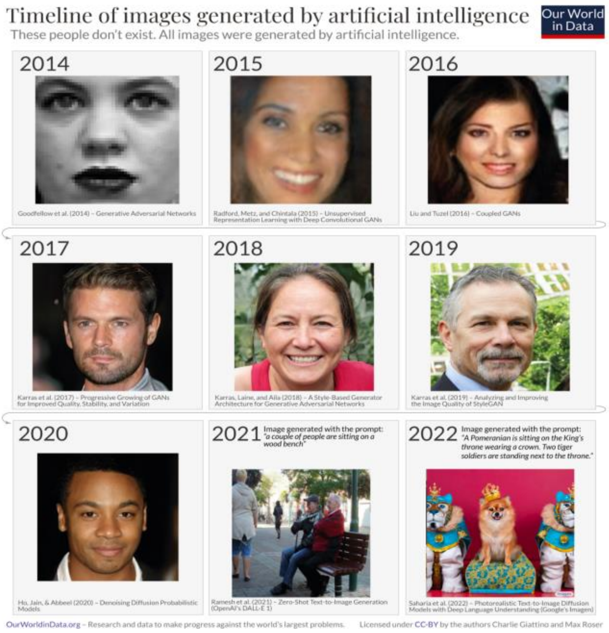
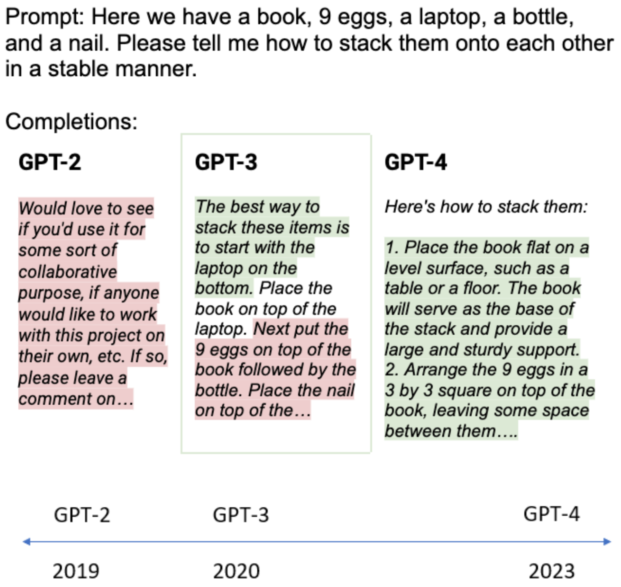

Timeline of images generated by artificial intelligence
These people don't exist. All images were generated by artificial intelligence.
Our World in Data
2014
Goodfellow et al. (2014) – Generative Adversarial Networks
2015
Radford, Metz, and Chintala (2015) – Unsupervised Representation Learning with Deep Convolutional GANs
2016
Liu and Tuzel (2016) – Coupled GANs
2017
Karras et al. (2017) – Progressive Growing of GANs for Improved Quality, Stability, and Variation
2018
Karras, Laine, and Aila (2018) – A Style-Based Generator Architecture for Generative Adversarial Networks
2019
Karras et al. (2019) – Analyzing and Improving the Image Quality of StyleGAN
2020
Ho, Jain, & Abbeel (2020) – Denoising Diffusion Probabilistic Models
2021 Image generated with the prompt: "a couple of people are sitting on a wood bench"
Ramesh et al. (2021) – Zero-Shot Text-to-Image Generation (OpenAI's DALL-E 1)
2022 Image generated with the prompt: "A Pomeranian is sitting on the King's throne wearing a crown. Two tiger soldiers are standing next to the throne."
Saharia et al. (2022) – Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Google's Imagen)
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the authors Charlie Giattino and Max Roser

Figure 4a. Timeline of images generated by image models from Our World in Data



Prompt: Here we have a book, 9 eggs, a laptop, a bottle, and a nail. Please tell me how to stack them onto each other in a stable manner.

Completions:

GPT-2

Would love to see if you'd use it for some sort of collaborative purpose, if anyone would like to work with this project on their own, etc. If so, please leave a comment on…

GPT-3

The best way to stack these items is to start with the laptop on the bottom. Place the book on top of the laptop. Next put the 9 eggs on top of the book followed by the bottle. Place the nail on top of the…

GPT-4

Here's how to stack them:

1. Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
2. Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them….

Figure 4b. Completions from GPT-2 to GPT-4. GPT-4 completion from Bubeck et al., 2023.

Recent progress was driven by systematic trends in compute, data and algorithms

A standard analysis of progress in AI capabilities considers three key factors: computing power, data, and improvements in the underlying algorithms.[71]

Computing power ("compute" for short) refers to the number of operations that are performed, usually in the context of training AI systems. The amount of compute used during training has expanded over the past decade by a factor of 55 million: from systems trained by single researchers at the cost of a few pounds, to systems trained on multiple GPU clusters by companies at the cost of many millions of pounds.[72] This trend is mostly the result of spending more money on compute, as well as the result of significant technological improvements to computing hardware.[73]

Training algorithms have also improved substantially over the past decade, so that today's machine learning models can achieve the same performance with less compute and data than those of the past. Research suggests that better algorithms roughly halved compute requirements each year for vision and language models.[74] Massive amounts of data have also played an important role in recent AI progress. AI developers have tapped into readily available datasets scraped from the internet, with the amount of training data used growing at over 50% per year.[75]

Enhancements applied after initial training have further augmented system capabilities. These post-training enhancements include improved data for fine-tuning,[76] equipping models with tools like calculators,[77] web browsers[78], and better prompts.[79] Post-training enhancements can significantly improve performance in specific domains at a small fraction of the original training cost,[80] and so a wide range of actors can use them to improve frontier AI capabilities.

Scaling laws: performance improves predictably with increased compute and data

The key driver for the increase in compute and data is that frontier AI model performance predictably improves with model scale. Researchers have discovered so-called "scaling laws",[81] which can predict, given a particular amount of compute and data, a frontier AI model's performance at the specific task of predicting the next word (the task used to train these models).

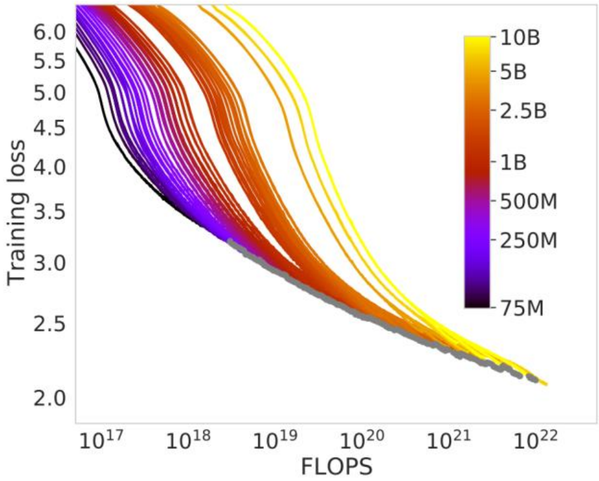


Figure 5a. Training error reduces predictably with compute across a broad range of empirically-studied training runs. Figure from Hoffmann et al, 2022.

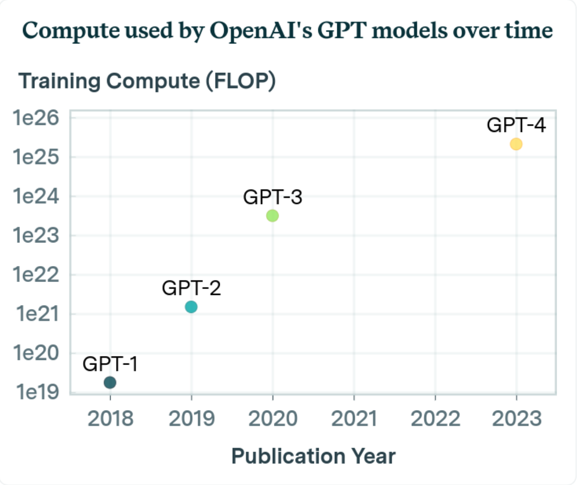
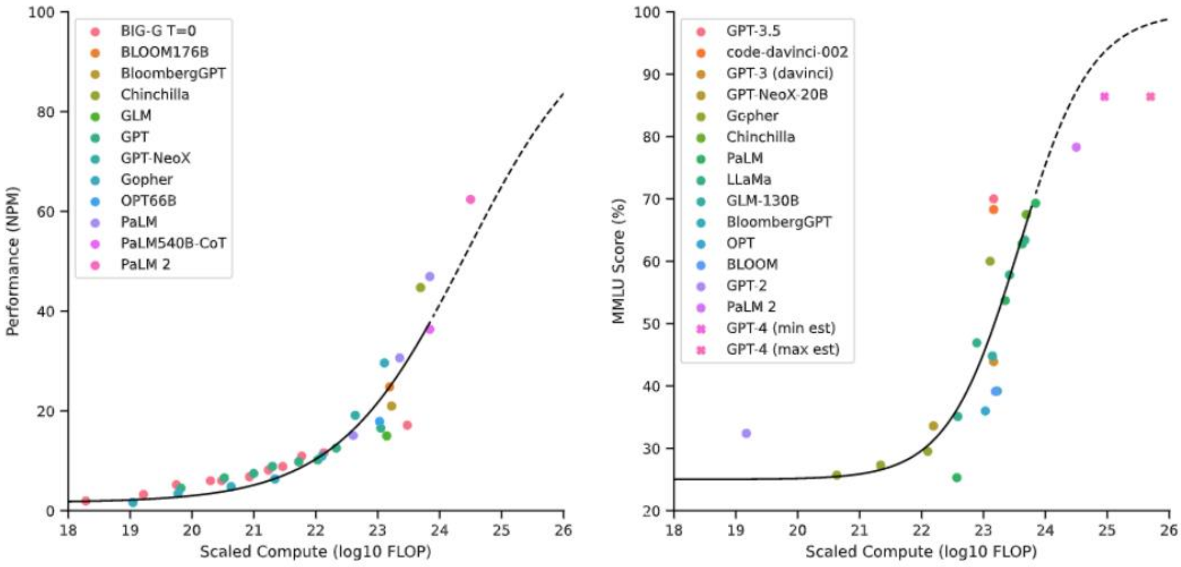


Figure 5b. Exponential increase in training compute for OpenAI's GPT models from 2018 to 2023.[82] *Epoch.*

Next word prediction has continually improved over time as developers have scaled their training compute and data. It is uncertain how long this trend will continue, but it has held over many orders of magnitude of compute and dataset size increases without breaking.

While the next word prediction task is not itself what we care about, it is used as an indicator of model capabilities since it is strongly correlated with performance in many downstream tasks.[83] For example, if a model is extremely good at next word prediction on code and mathematics data, it is more likely to be good at solving programming puzzles and mathematics problems.



(a) BIG-Bench

(b) MMLU

Figure 6. Performance on broad benchmarks such as BIG-Bench and MMLU improves with more training compute. This figure was taken from Owen 2023.

Although *average performance*, aggregated across many downstream tasks, improves fairly predictably with scale, it is much harder to predict performance improvements at *specific real-world problems*. The development of frontier AI systems has involved many examples of surprising capabilities, unanticipated by model developers before training and often only discovered by users after deployment. There are documented examples of unexpected capabilities where models were not showing any signs of improvement before a certain scale and then rapidly improved suddenly[84] – though the interpretation of these examples is contested.[85] In any case, we cannot currently reliably predict ahead of time which specific new capabilities a frontier AI model will gain when it is trained with more compute and data.

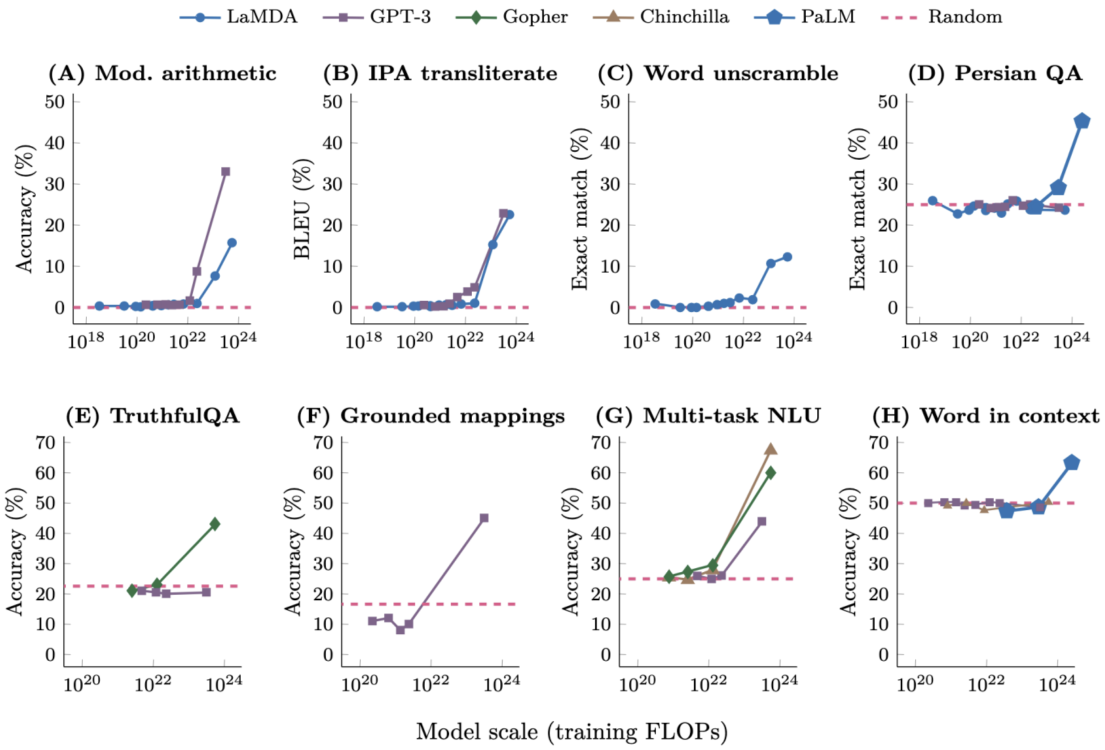


Figure 7. Individual capabilities may appear suddenly or unexpectedly as the compute used to develop AI increases. Figure from Wei et al, 2022.

Rapid AI progress is likely to continue for several years

The recent improvement in AI capabilities is not the result of a single breakthrough but rather a concerted advancement across multiple dimensions, including algorithms, spending on compute, improvements in hardware performance, and post-training enhancements. All of these factors can independently enhance progress, meaning that challenges or limits in any single one of them is unlikely to stop progress in AI as a whole.

Investments in AI will continue to grow rapidly over the next few years.[86] Leading AI developers like Anthropic and OpenAI have garnered significant funding and established cloud partnerships, in large part to support further scaling of compute.[87] Hardware manufacturers like TSMC are reportedly expanding their production of AI chips, again suggesting that more computational resources will be available for training.[88]

However, sustaining the rate of recent rapid scale up of compute and data past 2030 is likely to require new approaches. Developers would have to i) spend hundreds of billions of pounds on compute for a single training run[89] and ii) find ways to generate sufficient high-quality data going beyond what is readily available on the internet.[90] Having said this, improvements in algorithmic efficiency may reduce compute needs, such that compute might not be a binding constraint.

Novel research directions that could further accelerate frontier AI progress include:

- Enriched training data – e.g. expert human feedback, AI generated synthetic feedback, and data pruning – may increase data efficiency, improve capabilities on challenging scientific problems, and reduce costs.[91]
- Multimodal training, which may offer increasing synergies between the different modalities and the potential for frontier AI to process and produce text, images, audio and video.[92]
- Training frontier AI to act as an autonomous agent that navigates the internet as a human and performs long sequences of actions, using the above techniques to generate cheap data for learning these skills.[93]

Importantly, there is also the prospect that AI systems themselves accelerate AI progress. Frontier AI is already helping AI researchers to create synthetic data for training,[94] write new code,[95] and even improve model architectures.[96] While AI research is currently mostly non-automated, increased automation by future frontier AI systems may accelerate the pace of AI progress significantly.[97] This could mean we develop very capable AI systems sooner that we would otherwise expect, and have less time to prepare for the associated risks.

Advanced general-purpose AI agents might be developed in the future

Recent progress in AI has prompted discussion regarding the potential near-term development of advanced general-purpose, highly autonomous AI agents that can perform most economically valuable tasks better than human experts.

Several leading AI companies explicitly aim to build such systems,[98] and believe that they may succeed this decade.[99] Some surveys of published machine learning researchers have found the median respondent predicts a greater than 10% chance of human-level machine intelligence by 2035, though these surveys have been critiqued.[100] Attempts at forecasting the development of human-level machine intelligence based on historic trends in computing costs and growth in AI research inputs sometimes conclude that there is a greater than 10% probability by 2035.[101]

However, there is a large amount of uncertainty about the timeline to these capabilities. Many, if not most, other researchers do not expect AI systems that generally match human performance within twenty years and do not agree that it is a concern.[102] Historically, and frequently, there have been predictions of imminent AI breakthroughs that did not come to pass.[103]

What risks do frontier AI present?

We must understand the risks associated with frontier AI to safely access and seize the opportunities and benefits the technology brings.

In this section, we first review several cross-cutting *risk factors* – technical and societal conditions that could aggravate a number of particular risks. We then discuss individual risks under three headings: societal harms, misuse and loss of control.

We do not comprehensively cover all important AI risks and only highlight some salient examples.

Cross cutting risk factors

There are many long-standing technical challenges[104] to building safe AI systems, evaluating whether they are safe, and understanding how they make decisions. They exhibit unexpected failures and there are barriers to monitoring their use.

Adequate safety standards have not yet been established for AI development, there may be insufficient economic incentives for AI developers to invest in safety measures, and significant market concentration might exacerbate various risks.

It is difficult to design safe frontier models in open-ended domains

Frontier AI systems operate in open-ended domains, such as free-form dialogue or code generation. The complexity of open-ended domains makes it difficult to design safe systems or exhaustively evaluate all downstream use cases. While we can restrict the behavioural repertoire of an AI (for instance to text outputs from a limited vocabulary), this limits performance so may be uncompetitive and AI systems often use their behavioural repertoire in unanticipated ways, realising unexpected -- and potentially dangerous -- outcomes.[105]

In general, frontier AI systems are not robust, i.e. they frequently fail in situations sufficiently unlike their training data.[106] In particular, safeguards to prevent frontier AI models from complying with harmful requests (such as designing cyberattacks)[107] are not robust, and “adversarial” users who aim to bypass these safeguards have succeeded. Simple “jailbreaking” approaches, such as prompting the model to respond affirmatively to a request, are often sufficient, although more unusual prompts can be more effective.[108] AI that processes visual inputs may be especially vulnerable,[109] and methods for automatically generating adversarial prompts may worsen the situation.[110] Though AI robustness is a well-developed research field with thousands of published papers, in practice, lack of robustness is still an unsolved problem that affects all kinds of machine learning models, including language models,[111] image models,[112] and other AI agents.[113]

Preventing AI systems from pursuing unintended goals is an unsolved research problem, known as the “specification problem”. It is generally not possible to completely express complex behaviours, concepts, or goals directly in code, and so teaching AI which behaviours are desirable or undesirable must be done indirectly and can only be learned approximately; giving rise to potential specification and assurance gaps.[114] Current approaches to solving the specification problem involve training AI to behave in ways that score highly according to some metrics derived from data about human preferences.[115] Existing methods suffer from known limitations, and may not scale to highly advanced AI systems.[116] In addition, even if a solution to the technical specification problem was found, there are further social and technical challenges given the wide variation in people’s values.[117]

Evaluating the safety of frontier AI systems is an open challenge

Safety testing and evaluation of frontier AI is ad-hoc, with no established standards, scientific grounding or engineering best practices. Broadly, we can use techniques like interpretability to try to inspect a model’s inner functioning to understand whether it will behave as intended; or, we can try to evaluate a model’s behaviour by running experiments to see what outputs it gives in response to certain inputs. Both of these approaches have several limitations.

When building software, developers can precisely describe instructions for specific behaviours. This enables them to predict the system’s behaviour and understand its limitations. By contrast, frontier AI developers merely specify a learning process. The system produced by that process is not interpretable even to the system’s developers: hundreds of billions of parameters (numbers), which do not map cleanly to human-interpretable concepts.[118] For this reason, frontier AI systems are “black boxes” to their developers, who can observe their behaviour but have little understanding of the internal mechanisms that produce them. This lack of mechanistic understanding makes it challenging to know how to change, much less how to predict, the behaviour of an AI system.[119]

The nascent field of mechanistic interpretability aims to understand a model’s inner functioning, not just the behaviour in response to individual inputs. However, the field has so far only managed to explain a small fraction of behaviours in toy models much smaller and less capable than those used in practice.[120] Alternative techniques like saliency maps, which aim to identify which parts of input are salient to an AI model, have been shown to be unreliable or misleading.[121] Other approaches provide developers with an incomplete understanding of the model.[122]

Because of the above interpretability issues, many have turned to behavioural evaluations which simply involve observing the model’s response to certain inputs. However, such behavioural evaluations cannot exhaustively explore all possible vulnerabilities, and reliably extrapolating from those that have been explored is an open problem.[123] Due to their lack of robustness, frontier AI systems are likely to exhibit novel failure modes during deployment,[124] as they encounter novel situations not covered by previous evaluations.[125]

Formal verification techniques can prove the correctness of software (subject to assumptions). Some degree of robustness to small modifications to the input can be proven for AI systems using such techniques.[126] But in general, there may be many differences in an input that humans consider unimportant but that have major effects on an AI systems’ behaviour, and vice versa.[127] So using formal verification to ensure expected behaviour would require better methods of specifying what aspects of an input humans consider behaviourally (ir)relevant.

It may be difficult to track how frontier AI systems are deployed or used

Tracking the use of frontier AI models is important for monitoring misuse, noticing malfunctions or establishing liability for harms caused in part by frontier AI models.

Two common forms of deployment are: (i) an “open release” of the entire model, (ii) releasing a limited Application Programming Interface (API)[128] by which users can interact with the model in a particular way, e.g. receiving responses to user inputs.[129],[130] Either mechanism opens up the ability for third parties to develop applications using the frontier models, potentially enhancing model capabilities.

Open release (often referred to as open source) makes a model permanently available for other actors to copy, fine-tune, and use as they see fit. It is currently relatively cheap to fine-tune a model to enhance its capabilities or remove any safety features that have been put in place.[131] This could be misused,[132] but is also essential for innovation and enabling broader research into both AI safety and AI for good.

API access is reversible and allows a deployer to maintain control over a model and monitor its use. However, some of the model's capabilities might still be extracted.[133] As an example, responses from GPT-3.5 were used to train the open release Alpaca model, which does not have the same safety features as GPT-3.5.[134] Although open release generally has fewer reliable guardrails in place, APIs may also require less skill and resources to make use of, potentially lowering the bar for misuse.

Frontier AI systems, especially open release models, could also be used privately, and such use would likely remain undetected.

Frontier AI models embody extremely valuable intellectual property. Even if frontier developers intend to limit deployment, the information security practices of frontier developers will influence the likelihood that the full model is exfiltrated by employees or external actors. Much more investment in security would be needed for frontier AI developers to defend against attacks from the most well-resourced actors.[135] After exfiltration, attackers might then be able to use and modify the model without detection.

Unintended behaviours or dangerous capabilities might also be introduced in models via supply chain vulnerabilities such as training data poisoning or vulnerabilities in the hardware or software used to train or operate frontier models. For example, recent research has demonstrated that it is possible to automatically construct adversarial attacks on LLMs, that cause some systems to ignore their safeguards and obey user commands even if doing so produces harmful content.[136] At all levels of the supply chain, from hardware, to data ingestion, training, deployment and monitoring, vulnerabilities exist that could be deliberately exploited, or accidentally neglected.

AI safety standards have not yet been established

Researchers have argued that the breadth of potential use-cases for foundation models makes them a general-purpose technology, similar to electricity.[137] These industries can create systemic risks and are sometimes subject to dedicated regulators and have extensive standards, codes of practice and certification regimes.[138] Some researchers have argued that the AI industry should draw on practices observed in highly safety focussed industries such as healthcare, aviation, and nuclear engineering.[139]

But AI safety standards are still at an early stage. Work by standard development organisations such as IEEE, ISO/IEC and CEN/ CENELEC is still ongoing in many areas.[140] Similarly, while external assurance of models prior to and after deployment has been identified as an important mechanism for managing AI risks.[141] There is currently little government capacity for this and more work is required to build a mature ecosystem.[142] One challenge is that systems are often developed in one country and then deployed in another, enhancing the need for global coordination.[143]

Insufficient incentives for AI developers to invest into risk mitigation measures

Market failures are observed in many global challenges, such as climate change. When a company produces carbon emissions, the harms are not only incurred by them, but by the world. They do not incur the full cost, so there is an externality.[144] As a consequence, the company lacks sufficient incentive to reduce the harm.

Similarly, safe AI development may be hindered by market failure among AI developers and collective action problems among countries because many of the harms are incurred by society as a whole.[145] Individual companies may not be sufficiently incentivised to address all the potential harms of their systems. In recent years there has been an intense competition between AI developers to build products quickly.[146] Competition on AI has raised concern about potential "race to the bottom" scenarios, where actors compete to rapidly develop AI systems and under-invest in safety measures.[147] In such scenarios, it could be challenging even for AI developers to commit unilaterally to stringent safety standards, lest their commitments put them at a competitive disadvantage.[148] The risks from this "race" dynamic will be exacerbated if it is technologically feasible to maintain or even accelerate the recent rapid pace of AI progress.

There may be significant concentration of market power in AI

Researchers and regulators have begun to explore the likelihood of high concentration of market power among frontier AI developers.[149] The high upfront costs associated with training frontier AI models appear to create economies of scale and significant barriers to entry for smaller players. Established leaders benefit from better access to the cutting-edge computing resources and specialised talent required to develop frontier AI models. In addition, an early lead might grow over time, e.g. because the leader gathers data from their users they can use in training or because the leader uses their AI systems to accelerate their own progress.[150]

A considerable concentration of market power could weaken competition, reducing innovation and consumer choice. A loss of consumer choice also means users have less say in the use of their personal data, potential behavioural manipulation, surveillance, and an erosion of democratic norms.[151]

Societal harms

There is a wide range of potential societal harms arising from the use of AI.[152] This has sparked a debate around the ethics of AI, with a wide proliferation of ethical frameworks and principles.[153] We focus here on only a few societal harms, but this is not to downplay the importance of others.

Degradation of the information environment

Frontier AI can cheaply generate realistic content which can falsely portray people and events. There is potential risk of compromised decision-making by individuals and institutions who rely on inaccurate or misleading publicly available information, as well as lower overall trust in true information.

Information abundance leads to information saturation – people turn off and ignore information, including whether it is verified or not. A study by Ofcom reveals that 30% of UK adults who go online are unsure about, or do not even consider, the truthfulness of information.[154] The attention economy means on the supply side, trade-offs are made between the truth orientation of information and attention-grabbing strategies.[155] Additionally, frontier AI can be known for its tendency to generate false information, sometimes called 'hallucinations', without users being aware; meaning they could spread it unintentionally.[156] Meanwhile, adults and children overestimate their ability to spot misinformation.[157] Against this backdrop, the risk of frontier AI degrading the information environment is significant.

Impacts will be felt first where the truth is critical, news reporting, legal processes, and public safety.[158] There are examples already of outlets concerned that real images and videos are increasingly likely to be regarded as unreliable given they may have been AI generated.[159] There have been examples of AI hallucinating dangerous information, inadvertently radicalising individuals, and nudging users towards harmful actions as an unintended consequence of model design.[160] Long-term consequences, particularly as frontier AI becomes more embedded in mainstream applications and more accessible to children and vulnerable people, are highly uncertain.

Frontier AI may also result in indirect consequences that further degrade the information environment. For example, AI-generated functionalities and content are increasingly being integrated into search engines, which may lower traffic to news articles, harming the business models of news organisations that play an important role in debunking misinformation.[161]

Many of the harms arising from this degradation of the information environment would not be novel to AI, but the use of frontier AI may accelerate existing trends. Some examples of potential harm caused by frontier AI degrading the information environment include:

- Encouraging individuals to make dangerous decisions, for example through suggesting toxic substances as medicine.
- Exposing young or vulnerable people to high-risk information and age-restricted content, or significantly shaping their information diet.
- Promoting skewed or radical views as a result of model features — i.e. sycophancy[162] — that could lead to criminal or other harmful behaviours.
- Reducing public trust in true information, institutions, and civic processes such as elections.
- Contributing to systemic biases in online media as a result of bias in AI-generated content.[163]
- Inciting violence.[164]
- Exacerbating public health crises.[165]
- Increase political divisiveness, through malicious and non-malicious mechanisms.[166]

On the other hand, some have been using frontier AI to try to *improve* the information environment. For example, frontier AI chat assistants have been used to improve conversations about divisive topics, including political divisiveness.[167]

Authentication solutions (e.g. ‘watermarking’) are under development,[168] but should not be considered fully reliable yet as there are techniques that may allow users to escape detection.[169] Watermarking, like other solutions, may introduce new risks, that must be weighed up in balance with those which they mitigate. For example, watermarking may require new verification institutions or standards bodies with the major players involved, which could lead to a reinforcement or further concentration of power.

Labour market disruption

Economists view disruption and displacement in labour markets as one of the risks through which rapid advances in AI may affect citizens and reduce social welfare.[170] Technological change can also bring about improvements to working conditions, historically reducing the demand for human labour in more dangerous occupations.[171] While the impacts on labour markets remain uncertain and shapeable,[172] economists have identified potential risks and opportunities from AI to labour markets. AI has already begun to reduce the administrative burden of some roles and has the potential to accelerate this considerably including in areas such as teaching and medicine.

Throughout history, technological progress has always resulted in some level of change within the labour market. Introducing new technologies often causes temporary disruption as some workers transition within or between jobs.[173] For example, in 1940, 60% of job categories today didn't exist.[174] Studies suggest that the sectors with greatest exposure to labour market disruption from current AI capabilities are IT, financial, legal while education, manufacturing, agriculture and mining are least exposed.[175] On the other hand, we may return to pre-1980 trends in which worker displacement from automation was roughly offset by creation of new roles.[176]

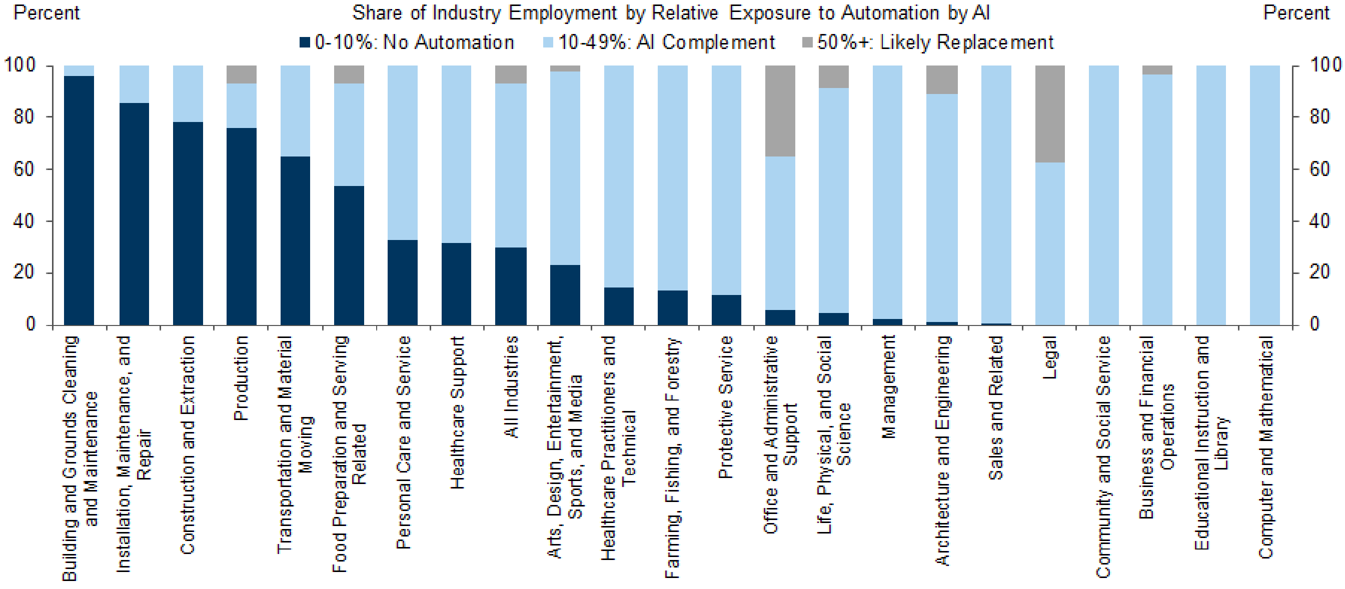


Figure 8. Share of industry employment by relative exposure to automation by AI. Taken from The Potentially Large Effects of Artificial Intelligence on Economic Growth, Goldman Sachs, 2023.

Bias, Fairness and Representational Harms

Frontier AI models can contain and magnify biases ingrained in the data they are trained on, reflecting societal and historical inequalities and stereotypes.[177] These biases, often subtle and deeply embedded, compromise the equitable and ethical use of AI systems, making it difficult for AI to improve fairness in decisions.[178] Removing attributes like race and gender from training data has generally proven ineffective as a remedy for algorithmic bias, as models can infer these attributes from other information such as names, locations, and other seemingly unrelated factors.[179]

Frontier AI models are primarily trained on textual sources, including digitised books and online text. Consequently, they are exposed to derogatory language and stereotypes that target marginalised groups. The training data often mirrors historical patterns of systemic injustice, inequalities in the contexts from which the data is sourced,[180] or it reflects dominant cultures (consider high internet-access regions) and lack data on certain worldviews, cultures and languages.[181] Frontier AI systems have been found to not only replicate but also to perpetuate the biases ingrained in their training data.[182]

When bias manifests in AI outputs, it can do so in subtle and complex ways.[183] Because frontier models lack transparency, it becomes a formidable task to pinpoint the exact mechanisms through which bias has been introduced into their decisions.[184] The complex nature of bias makes it challenging to identify and rectify instances of unfairness.[185] Individuals may therefore question whether their treatment by an AI system was influenced by their gender, race, or other personal characteristics – without insight into the model's inner workings, it is difficult to find answers.

As frontier model capabilities develop, and AI-generated content could come to represent a greater proportion of content available online,[186] there is potential for a reinforcing loop whereby future AI systems are trained on increasingly biased AI-generated content.

AI technologies are increasingly integrated into systems responsible for consequential decision-making, including in sectors where fairness is paramount.[187] Frontier AI technologies have predictable risks when deployed in these settings.[188] Bias in AI systems is particularly concerning in high-stakes real-world domains like job recruitment, financial lending, and healthcare, where biased decisions can have profound consequences.[189] However, there may be cases where taking such factors into account is legitimate, e.g., in some healthcare settings where doses of medication may vary with age; this makes identification of harmful biases even more difficult.[190] Nonetheless, it is possible to mitigate bias, , both at the point of curating the training data and during or after training, when evaluating to what extent outputs are biased.[191]

It is worth noting that discrimination due to model bias can be seen as a kind of alignment problem: AI systems are behaving in ways that its developers did not intend. This highlights the importance of investing in AI alignment and AI ethics research.

Misuse risks

Frontier AI may help bad actors to perform cyberattacks, run disinformation campaigns and design biological or chemical weapons. Frontier AI will almost certainly continue to lower the barriers to entry for less sophisticated threat actors.[192] We focus here on only a few important misuse risks, but this is not to downplay the importance of others.

Dual Use Science risks

Frontier AI systems have the potential to accelerate advances in the life sciences, from training new scientists to enabling faster scientific workflows. While these capabilities will have tremendous beneficial applications, there is a risk that they can be used for malicious purposes, such as for the development of biological or chemical weapons. Experts are in disagreement about the magnitude of risk that AI advances will pose for biosecurity.[193]

Current capabilities

Frontier AI models can provide user-tailored scientific knowledge and instructions for laboratory work which can potentially be exploited for malicious purposes.[194] Studies have shown that systems may provide instruction on how to acquire biological and chemical materials.[195]

Existing frontier AI models have some ability to design and troubleshoot laboratory experiments.[196] These capabilities can be enhanced when equipping LLMs with the ability to access tools such as web search and specialised computational tools.[197] When connected to laboratory robots or cloud labs,[198] LLMs can directly instruct these platforms to carry out experiments.[199]

While our focus is on frontier AI, it is important to note that frontier capabilities can be used in conjunction with the capabilities of narrower biological AI design tools,[200] such as AlphaFold2[201] and RFDiffusion.[202] Narrower AI tools can already generate novel proteins with single simple functions and support the engineering of biological agents with combinations of desired properties.[203] Biological design tools are often open sourced which makes implementing safeguards challenging.[204] Frontier AI can instruct specialised AI systems[205] and make them more accessible, and in the future may itself feature similar abilities.[206]

Projected capabilities

Future frontier AI will likely feature even greater content-level knowledge, reasoning abilities, and capacity to formulate complex plans. Additionally, some expect that future capabilities will make experimental instructions more accessible, including through the ability to generate images and video, and may make science systems more automated.[207] However, it remains unclear whether frontier AI systems add additional capability over just using existing tools such as web search, as studies do not yet control for this.

Potential Risks and Impacts

While the impact of current systems on biological and chemical security risks is still limited, anticipated near-future capabilities have the potential to increase dual-use science capabilities. Current AI systems in particular pose risks where current biological and chemical supply chains already feature vulnerabilities. Significant barriers remain for novel laboratory work.[208] Some of these barriers could be reduced by near-future advances in frontier AI and associated advances in laboratory automation.

Cyber

As the programming abilities of AI systems continue to expand, frontier AI is likely to significantly exacerbate existing cyber risks. Most notably, AI systems can be used by potentially anyone to create faster paced, more effective and larger scale cyber intrusion via tailored phishing methods or replicating malware. Frontier AI's effect on the overall balance between cyber offence and defence is uncertain, as these tools also have many applications in improving the cybersecurity of systems and defenders are mobilising significant resources to utilise frontier AI for defensive purposes.[209] In the future, we may see AI systems both conducting and defending against cyberattacks with reduced human oversight at each step.

Current Cyber Capabilities of Frontier AI

Frontier AI can upskill threat actors by advising on attack techniques, critiquing cyberattack plans, or finding relevant information about a target.[210] Some models have measures to avoid supporting cyber criminals, but these are frequently circumvented through 'jailbreaks'.[211] The uplift provided by current models is limited: they often hallucinate or otherwise give unhelpful answers. As the models improve, this uplift is expected to increase.

Frontier AI systems are saving skilled threat actors time. For example, AI systems have helped create computer viruses that change over time to avoid detection, which previously would have required significant time from experts.[212] Users on underground hacking forums have claimed to be using tools like ChatGPT to help them recreate malware quickly in many different programming languages.[213]

AI improves the effectiveness of existing techniques. AI-enhanced social engineering is already being used by cybercriminals to conduct scams and steal login credentials, with systems that can gather intelligence on targets,[214] impersonate voices of trusted contacts,[215] and generate persuasive spear phishing messages.[216] The risk is significant given most cyber attackers use social engineering to gain access to the victim organisation's networks.[217]

AI systems create new attack techniques. For example, they can power highly targeted malware to act benignly until it detects the intended victim and only then act maliciously.[218] This makes it harder for antivirus software to detect the attack.

Projected Capabilities

Future frontier AI developments will increase the scale and speed of attacks. Current tactics often require human effort which could be replaced by more advanced AI systems, leading to greater scalability of potent cyberattacks. Additionally, AI systems will be able to perform actions more quickly than humans, making human-based defence less effective.

Frontier AI developments will continue to enhance existing attack techniques. For instance, information gathering and targeting is highly likely to become more effective as AI systems are able to be able to process more information simultaneously,[219] and more accurate as models' reasoning capabilities improve,[220] or they are augmented with other tools.[221]

Frontier AI developments may result in systems that can act on the internet to perform their own cyberattacks autonomously.[222] Behaviours such as autonomous replication and self-improving exploit generation are of particular concern, and some work has started to look at how good today's models are at these behaviours.[223]

Cyber defence using frontier AI will likely mitigate some of this risk. In particular, frontier AI is highly likely to help with anomaly detection, security scanning, and mitigating insider threats.[224] However, this defence capability may lag behind offence in the short term, since AI-assisted vulnerability repair and detection may rely on significantly more capable frontier AI systems than currently exist to be effective.[225]

Frontier AI also introduces security vulnerabilities when it is integrated into broader systems. These new digital vulnerabilities – for example corrupting training data ('data poisoning'), hijacking model output ('prompt injection'), and extracting sensitive training data ('model inversion') – will require new and bespoke cybersecurity responses.[226]

Potential Risks and Impact

Critical infrastructure like energy,[227] transportation,[228] healthcare,[229] and finance,[230] are already frequently targeted by cyberattacks today.[231] This can result in the theft of intellectual property, direct theft of funds, data destruction or ransom, privacy breaches, and disruption to operations across the private, public and third sectors.

Cyberattacks also often cause significant harm to the public, including physical,[232] monetary, mental and emotional harms, abuse,[233] discrimination, denial of access to key services and loss of control over personal data.[234] In addition to direct and visible impacts, these harms can lead to erosion of trust in digital systems, limiting people's access to services, preventing responsible innovation, and reducing democratic engagement.

Frontier AI might increase the harms in the above categories and may also create novel harms, such as emotional distress caused by fake kidnapping or sextortion scams.[235] As frontier AI continues to be deployed and used in cybersecurity,[236] it is uncertain where the balance between offence and defence capabilities will end up as frontier AI development continues.

Disinformation and Influence Operations

In addition to unintentional degradation of the information environment (discussed in the section on Societal Harms above), frontier AI can be misused to deliberately spread false information to create disruption, persuade people on political issues, or cause other forms of harm or damage. Although current financial costs for human-generated disinformation remain low,[237] there is already some evidence that cheap, realistic content generated by frontier AI systems is already aiding disinformation campaigns.[238]

Although many emphasise the forms this could take are unpredictable,[239] improved capabilities could stem from several factors. First, the accessibility of cheap, high quality content will lower the price and barrier to entry to creating a disinformation campaign.[240] More actors producing more disinformation could increase the likelihood of high-impact events.

Additionally, AI-generated deepfakes are becoming extremely realistic, meaning they often cannot be identified by individuals or even institutions with advanced detection technologies.[241] Even where AI-generated content is not universally believed, its strategic deployment may cause disruption, confusion, and loss of trust.

Frontier AI can generate hyper-targeted content with unprecedented scale and sophistication.[242] This could lead to "personalised" disinformation, where bespoke messages are targeted at individuals rather than larger groups and are therefore more persuasive.[243] Furthermore, one should expect that as AI-driven personalised disinformation campaigns unfold, these AIs will be able to learn from millions of interactions and become better at influencing and manipulating humans, possibly even becoming better than humans at this.[244] In doing so, they may utilise new manipulation tactics against which we are not prepared because defences have been developed through the influencing attempts of other humans.[245]

Disinformation detection approaches, such as watermarking, discussed above, have been proposed and trialed but still face challenges in effectively detecting false content.[246] Whilst improving media literacy is crucial, it is hard given that the quality of outputs from frontier AI is in many cases indistinguishable even to experts. This is a trend expected to increase with model size – in the GPT-3 paper, authors experiments found humans were better at distinguishing AI generated text for smaller models, but for larger models they could only tell the difference about 52% of the time, barely above random chance.[247]

Loss of control

Humans may increasingly hand over control of important decisions to AI systems, due to economic and geopolitical incentives. Some experts are concerned that future advanced AI systems will seek to increase their own influence and reduce human control, with potentially catastrophic consequences - although this is contested.

There are broadly two factors that could contribute to loss of control:

- Humans increasingly hand over control of important decisions to AIs. It becomes increasingly difficult for humans to take back control.
- AI systems actively seek to increase their own influence and reduce human control.

These are not mutually exclusive – if humans have already handed over significant control to AI systems, it will likely be easier for them to actively gain more influence.

The likelihood of these risks remains controversial, with many experts thinking the likelihood is very low and some arguing a focus on risk distracts from present harms.[248] However, many experts are concerned that losing control of advanced general-purpose AI systems is a real possibility and that loss of control could be permanent and catastrophic.[249]

Humans might increasingly hand over control to misaligned AI systems

Organisations around the world are already deploying misaligned AI systems that are causing harm in unexpected ways.[250] Recommendation algorithms increase the consumption of extremist content.[251] Medical algorithms have been known to misdiagnose US patients,[252] and recommend incorrect prescriptions.[253] Still, we hand over more control to them, often because they are still as - or more - effective than human decision making, or because they are cheaper.

As AI systems become increasingly capable and autonomous, the economic and competitive incentives to deploy them will grow accordingly.[254] Surveys have found that a large number of users overestimate the reliability of generative AI systems.[255] Known biases can lead consumers to over rely on AI applications, including automation bias,[256] confirmation bias,[257] and anthropomorphism.[258] These factors could lead to overreliance on autonomous AI systems that perform increasingly wide-ranging and critical tasks,[259] even if there is a risk the systems are misaligned.[260]

As economic production becomes increasingly dependent on AI systems, the cost of maintaining or reintroducing human control will increase. Advanced AI systems may alter complex systems in ways that are hard to understand,[261] making it hard or risky to extract them. As a result, AI systems may increasingly steer society in a direction that is at odds with its long-term interests, even without any intention by any AI developer for this to happen.[262] Even if many people recognize it happening, it may be difficult to stop (again, the analogy with climate change is illustrative).

Future AI systems might actively reduce human control

Loss of control could be accelerated if AI systems take actions to increase their own influence and reduce human control. This threat model is controversial - experts in AI significantly disagree on how likely it is and those who deem it is likely disagree on the timeframe.

There are two requirements for an AI system to actively reduce human control. First, it must *have the disposition* to take actions that would reduce human control. Second, it must have the *capabilities* to succeed in the face of countermeasures.

Future AI systems may have the *disposition* to reduce human control

AI systems might be disposed to take actions that increase their own influence and reduce human control either because a bad actor instructs them to do so, or because they have unintended goals.

A bad actor could give an AI system an objective that causes it to reduce human control, for example a self-preservation objective.[263] Some groups may simply want to inflict harm on broader society or raise their profile (terrorism).[264] There are people who believe, for a variety of reasons, that the highly advanced AI systems of the future are natural successors to humanity.[265] If there are safeguards in place, bad actors might dismantle them.[266]

Future advanced AI systems with unintended goals may have the disposition to reduce human control. Ensuring that AI systems do not pursue unintended goals, i.e., are not misaligned, is an unsolved technical research problem and one that is particularly challenging for highly advanced AI systems.[267] Many examples of unintended goal-directed behaviour have been observed in the lab.[268] Many possible unintended goals would be advanced by reducing human control.[269] Future AI systems may consistently take actions that advance their goals and so such a system might, without human instruction, be disposed to take actions that reduce human control.[270]

Some researchers are sceptical of our ability to assess the plausibility of hypothetical future scenarios like this,[271] while others believe that this scenario is the default consequence of the current trajectory of AI development.[272]

If future AI systems were *disposed* to take actions that reduce human control – either from human instruction or from unintended goals – this would only pose a risk if they had *capabilities* that could meaningfully reduce human control.

Frontier AI shows early signs of capabilities that could be used to reduce human control

Today's systems have some basic capabilities that could, if rapid AI progress continues, be used to increase their own influence and reduce human control. Currently, these capabilities are not sufficient to pose significant risks and some argue that we are unlikely to ever see the future development of such capabilities.

At present, frontier AI is confined almost exclusively to the digital realm, thus the most immediate risks are likely to arise via manipulating humans or exploiting software vulnerabilities.

Current frontier AI systems have shown early signs of capabilities that could enable:

- **Manipulation**,[273] for example:
  - One social companion chatbot, based on GPT-3, quickly built trust and intimacy with users.[274] A user of the chatbot described themselves as "happily retired from human relationships."[275] Such intimacy could potentially be used to manipulate users.
  - There is evidence that language models tend to respond as though they share the user's stated views, and larger models do this more than smaller ones.[276] The ability to predict people's views and generate text that they will endorse could be useful for manipulation.
  - Frontier AI models can maintain coherent lies in simple deception games, and larger models are more persuasive liars.[277]

    - In an online study, 1500 participants used an opinionated LLM to help them write about a topic. They reported agreeing with the LLM's opinion on the topic considerably more often in a subsequent survey, having changed their opinion to align with it.[278]
- **Cyber offence.** Instead of - or in addition to - manipulating humans, AI systems could acquire influence by exploiting vulnerabilities in computer systems. Offensive cyber capabilities could allow AI systems to gain access to money, computing resources, and critical infrastructure. As discussed earlier in this report, frontier AI is already lowering the barrier for threat actors and future AI agents may be able to execute cyber attacks autonomously.
- **Autonomous replication and adaptation.** Controlling AI systems could become much harder if they could autonomously persist, replicate, and adapt in cyberspace. No current AI systems have this capability, but recent research found that frontier AI agents can perform some relevant tasks.[279]

As discussed earlier in the report, while some experts believe that highly capable general-purpose AI agents might be developed soon, others are sceptical it will ever be possible. If this does materialise such agents might exceed the capabilities of human experts in domains relevant to loss of control, for example political strategy, weapons design, or self-improvement. For loss of control to be a catastrophic risk, AI systems would need to be given or gain some control over systems with significant impacts, such as military or financial systems. This remains a hypothetical and hotly disputed risk.

Conclusion

Understanding the capabilities and risks of frontier AI is critical to unlocking its benefits. That is why we are being proactive in grappling with the risks, rather than waiting for them to transpire.

We have seen that recent progress in frontier AI has been fast and impressive. Frontier AI can perform a wide variety of tasks, and is being augmented with tools to enhance its capabilities. Systematic trends driving recent growth will continue for the next several years. This is due to a correlation between more compute, more data and better algorithms, with the performance of frontier AI. Progress over the next few years could be fast and surprising in certain ways. We cannot predict which specific capabilities will emerge as AI improves. It is possible that advanced general-purpose AI agents could be developed in the not too distant future. On the other hand, some argue there is a lack of evidence that this will happen anytime soon, or that the capabilities we are observing do not trend towards fully general AI.

There are many opportunities from these developments, and these can only be realised if the risks are mitigated. There are several deep, unsolved cross-cutting technical and social risk factors that exacerbate the risks. We outlined examples of societal harms, risks of misuse from bad actors, and even the possibility of losing control of the technology itself if it becomes advanced enough. Some think this is very unlikely, or that if general AI agents did exist they would be easy to control. Regardless of likelihood, these risks require further research – they and can interact with and amplify each other, and could cause significant harm if not addressed. Addressing them, however, will allow us to seize the opportunity, and realize their transformative benefits.

There may not be sufficient economic incentives to develop advanced AI with sufficient guardrails in place, and adequate safety standards have not yet been established for these potential future risks. Therefore it is important that we build a shared understanding of the risks, so that we are equipped to coordinate effectively on preventing and mitigating them as best as possible, and that we continue to work together internationally on frontier AI safety.

Glossary

- **Autonomous:** Capable of operating, taking actions, or making decisions without human oversight.
- **AI agents:** Autonomous AI systems that perform multiple sequential steps – sometimes including actions like browsing the internet, sending emails, or sending instructions to physical equipment – to try and complete a high-level task or goal.
- **AI developers:** Organisations in which scientists, engineers, and researchers work on developing AI models and applications.
- **AI risks:** The potential negative or harmful outcomes arising from the development or deployment of AI systems.
- **Alignment:** the process of ensuring an AI system's goals and behaviours are in line with human values and intentions.
- **Application Programming Interface (API):** a set of rules and protocols that enables integration and communication between AI systems and other software applications.
- **Biological design tools:** AI systems trained on biological data that can help design new proteins or other biological agents.
- **Capabilities:** The range of tasks or functions that an AI system can perform and the proficiency with which it can perform them.
- **Cloud labs:** remotely controlled automatised biochemical laboratories.
- **Cognitive tasks:** Tasks involving a combination of information processing, memory, information recall, planning, reasoning, organisation, problem solving, learning, and goal-oriented decision-making.
- **Compute:** Computational processing power, including CPUs, GPUs, and other hardware, used to run AI models and algorithms.
- **Computer worm**: A type of malicious software that self-replicates and spreads autonomously across computer networks, exploiting vulnerabilities to infect systems and potentially causing damage or disruption.
- **Disinformation:** Deliberately false information spread with the intent to deceive or mislead.
- **Evaluations:** systematic assessments of an AI system's performance, capabilities, or safety features. These could include benchmarking tests, adversarial testing, or user feedback amongst other methods.

- **FLOPS**: are ‘floating point operations per second’ and measure the computing power of a computer
- **Foundation models:** Machine learning models trained on very large amounts of data that can be adapted to a wide range of tasks.
- **Frontier AI:** AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today’s most advanced models.
- **Guardrails:** pre-defined safety constraints or boundaries set up in an attempt to ensure an AI system operates within desired parameters and avoids unintended or harmful outcomes.
- **Heuristic:** a rule-of-thumb, strategy, or a simplified principle that has been developed to solve problems more efficiently when classic methods are too slow or fail to find an exact solution.
- **Input [to an AI system]:** The data or prompt fed into an AI system, often some text or an image, which the AI system processes before producing an output.
- **Large Language Models (LLMs)**: Machine learning models trained on large datasets that can recognise, understand, and generate text and other content.
- **Misinformation**: Incorrect or misleading information spread without harmful intent.
- **Misgeneralisation**: When an AI system trained to perform well in one context fails to perform well in a new context. For instance, if an AI trained mostly on pictures of white cats, labels a black cat as a "dog," it is misgeneralising from its training data.
- **Narrow AI**: an AI system that performs well on a single task or narrow set of tasks, like sentiment analysis or playing Chess.
- **Open ended domains:** Scenarios or environments that have a very large set of possible states and inputs to an AI system – so large that developers cannot test the AI’s behaviour in all possible situations.
- **Prompt:** an input to an AI system, often a text-based question or query, that the system processes before it produces a response.
- **Scaffold:** Software program that structures the information flow, leaving the model itself unchanged.
    - For example, a scaffold allows GPT-4 to power the autonomous AI agent AutoGPT. The scaffold prompts GPT-4 to: break down a high-level task into sub-tasks, assign sub-tasks to other copies of itself, save important information to memory, and browse the internet.
- **Risk factors:** Elements or conditions that can increase downstream risks. For example, weak guardrails (risk factor) could enable an actor to misuse an AI system to perform a cyber attack (downstream risk).
- **Weights:** parameters in a model are akin to adjustable dials in the algorithm, tweaked during training to help the model make accurate predictions or decisions based on input data, ensuring it learns from patterns and information it has seen before.

Annex A – Future Risks of Frontier AI (attached)

Annex B - Safety and Security risks from Generative AI (attached)

References

[1] Towards Expert-Level Medical Question Answering with Large Language Models. Singhal et al, 2023.

[2] Introduction to the AI safety Summit, Department for Science, Innovation, and Technology, 2023.

[3] Large Language Models (LLMs): AI models, primarily based on deep learning architectures like transformers, designed to understand, generate, and manipulate human language.

[4] Introducing ChatGPT, OpenAI, 2022.

[5] Introducing Claude, Anthropic, 2023.

[6] An important next step on our AI journey, Pichai, 2023.

[7] Narrow AI: an AI system that performs well on a single task or narrow set of tasks, like sentiment analysis or playing Chess.

[8] Explainer: What is a foundation model?, Ada Lovelace Institute, 2023

[9] Datapoints used to train notable artificial intelligence systems, Our World in Data, 2023.

[10] Technically it is not words, but rather "tokens" (particular sequences of characters) for technical reasons – but this distinction is not important for the purposes of the report.

[11] Language models are better than humans at next-token prediction, Shlegeris et. al, 2022.

[12] Fine-tuning is an optional additional training process that can be applied to pre-trained models to add specific capabilities or improvements by leveraging particular datasets.

[13] *Introduction* in GPT-4 System Card, OpenAI, 2023.

[14] For example, see GPT-4 (vision) system card, OpenAI, 2023.

[15] Compute: Computational processing power, including CPUs, GPUs, and other hardware, used to run AI models and algorithms.

[16] The AI Triad and What It Means for National Security Strategy, Buchanan, 2020.

[17] Section 3.35 AI foundation models: Full report, Competition and Markets Authority, 2023.

[18] The CEO of OpenAI claimed that the cost of GPT-4 exceeded $100 million at an event held at MIT Wired, 2023.

[19] Trends in the dollar training cost of machine learning systems, Cottier, 2023;

Dario Amodei, CEO and Co-Founder of Anthropic, predicts AI models could cost billions of dollars to run in this interview with Logan Bartlett.

[20] Section 4.5 AI foundation models: Full report, Competition and Markets Authority, 2023.

[21] AI capabilities: The range of tasks or functions that an AI system can perform and the proficiency with which it can perform them. These capabilities can span from summarisation to complex problem solving, and evolve over time with advancements.
[22] Evaluation of OpenAI Codex for HPC Parallel Programming Models Kernel Generation, Godoy et al., 2023; OpenAI Codex, OpenAI, 2021; Use Copilot to build and edit apps in Power Apps Studio, Microsoft, 2023.

[23] GPT-4 Technical Report OpenAI, 2023.

[24] Language Models are Few-Shot Learners, Tom B, Brown et al., 2020.

[25] Sparks of Artificial General Intelligence: Early experiments with GPT-4, Bubeck et al., 2023; Pathways Language Model (PaLM), Google, 2022.

[26] Pathways Language Model (PaLM), Google, 2022.

[27] See Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis
[28] Towards Helpful Robots: Grounding Language in Robotic Affordances, Google, 2022;

TidyBot: Personalized Robot Assistance with Large Language Models, Wu et al., 2023.

[29] Language Models and Cognitive Automation for Economic Research, Korinek, 2023.

[30] GPT-4, OpenAI, 2023.
[31] Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification, Zhou et al., 2023;

Solving Quantitative Reasoning Problems with Language Models, Gur-Ari et al., 2022.

[32] Introducing 100K Context Windows, Anthropic, May 2023;

Tools such as ChatGPT threaten transparent science; here are our ground rules for their use, Nature, 2023.

[33] How People Can Create - And Destroy - Value with Generative AI, BCG, 2023.
Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality, Dell'Acqua et al., 2023.

[34] Schwarcz and Choi, 2023;A&O announce exclusive launch partnership with Harvey, Allen & Overy, February 2023.

[35] See Morgan Stanley Wealth Management Announces Key Milestone in Innovation Journey with OpenAI, Morgan Stanley, March 2023.
[36] Generative AI at Work (Working Paper), Brynjolfsson, Li, Raymond, 2023.

[37] Generative AI for Economic Research: Use Cases and Implications for Economists (Forthcoming), Korinek, 2023.

[38] Capable of operating, carrying out sequences of actions, or making decisions without human intervention.
[39] AI agents: AI systems that autonomously perform multiple sequential steps – sometimes including actions like browsing the internet, sending emails, or sending instructions to physical equipment – to try and complete a high-level task or goal.
[40] A scaffold is a software program that structures the information flow between multiple copies of an AI model, leaving the model itself unchanged. For example, a scaffold allows GPT-4 to power the autonomous AI agent

AutoGPT. The scaffold prompts GPT-4 to: break down a high-level task into sub-tasks, assign sub-tasks to other copies of itself, save important information to memory, and browse the internet.

[41] AutoGPT

[42] Evaluating Language-Model Agents on Realistic Autonomous Tasks, ARC Evals, 2023.

[43] Generative Agents: Interactive Simulacra of Human Behaviour, Joon Sung Park et al, August 2023.

[44] Voyager: An Open-Ended Embodied Agent with Large Language Models, Guanzhi Wang et al, May 2023.

[45] SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning, Yue Wu et al., May 2023.

[46] Emergent autonomous scientific research capabilities of large language models, Gomes et al., 2023.

[47] OpenAI Charter, OpenAI, 2018; About, Google DeepMind.

[48] For example, see The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", Evans et al., 2023.

[49] A prompt is an input to an AI system, often a text-based question or query, that the system processes before it produces a response.

[50] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al., 2023.

[51] Chat Plugins, OpenAI, 2023

[52] Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023; Emergent autonomous scientific research capabilities of large language models, Gomes et al., 2023.

[53] For example, it is a scaffold that allows GPT-4 to power AutoGPT. Scaffolds might prompt a frontier model to: break down high-level task into sub-tasks, assign sub-tasks to other copies of itself, save insights to a memory bank, and browse the internet.

[54] Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al., 2023.
Self-critiquing models for assisting human evaluators, Saunders et al., 2022

[55] HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, Shen et al., 2023;

Improving Factuality and Reasoning in Language Models through Multiagent Debate, Du et al., 2023;

ChatGPT can now hear and speak, OpenAI, 2023.

[56] Heuristic: a rule-of-thumb, strategy, or a simplified principle that has been developed to solve problems more efficiently when classic methods are too slow or fail to find an exact solution.

[57] Pathways Language Model (PaLM), Google, 2022. Also see many putative examples in Sparks of Artificial General Intelligence: Early experiments with GPT-4, Bubeck et al., 2023.

[58] Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks, Wu et al., 2023.

[59] Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting; Faith and Fate: Limits of Transformers on Compositionality, Dziri et al., 2023;

Are Emergent Abilities in Large Language Models just In-Context Learning?, Madabushi et al., 2023.

[60] The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", Evans et al., 2023.

[61]For example Meta's Galactica model was withdrawn due to a tendency to generate inaccurate information when assisting scientists: Why Meta's latest large language model survived only three days online, MIT Technology Review, 2022.

See also:
Section 5.9 AI foundation models: Full report, Competition and Markets Authority, 2023;
Fabrication and errors in the bibliographic citations generated by ChatGPT, Walters, Wilder, 2023.

[62] Lamda: Language models for dialog applications, Thoppilan et al., 2022;
WebGPT: Browser-assisted question-answering with human feedback., Nakano et al., 2021;
Retrieval augmentation reduces hallucination in conversation, Shuster et al., 2021.

[63] For example, ARC Evals undertook an open-ended task with four LLM agents based on OpenAI's GPT-4 and Anthropic's Claude. See Evaluating Language-Model Agents on Realistic Autonomous Tasks, Kinniment et al., 2023. These AI systems were inconsistent in scraping and listing the top ten BBC news articles and unsuccessful in identifying recent employee additions to a company when asked to scrape the data. While attempting the tasks, the systems performed poorly at accessing a vast array of websites and software applications.
See also:
- Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change), Valmeekam et al., 2022
- Tree of Thoughts: Deliberate Problem Solving with Large Language Model, Yao et al., 2023

[64] Challenges and applications of large language models, Kaddour et al., 2023.

[65] Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al., 2023.
Self-critiquing models for assisting human evaluators, Saunders et al., 2022

[66] Challenges and applications of large language models, Kaddour et al., 2023.

[67] Scaling in the service of reasoning and model-based machine learning, Bengio, 2023;
GFlowNet-EM for learning compositional latent variable models, Hu et al, 2023.

[68] Scoring forecasts from the 2016 "Expert Survey on Progress in AI", Levermore, 2023; Language models surprised us, Cotra, 2023 and references therein.

[69] See Language models surprised us, Cotra, 2023 and references therein.

[70] See, for example, Visualizing the deep learning revolution, Ngo, 2023

[71] AI and Compute, Amodei and Hernandez, 2018; The AI Triad and What It Means for National Security Strategy, Buchanan, 2020; ML trends, Epoch, 2023.

[72] To sustain the growth in compute used on training runs, AI developers must now procure ever-larger clusters of AI-specialised chips. State of the art clusters being used to train today's largest models require investment in the hundreds of millions or billions of pounds, with refresh cycles every 2-3 years needed to stay at the cutting edge.

Compute Trends Across Three Eras of Machine Learning, Sevilla et al, 2022;

Trends in the dollar training cost of machine learning systems, Cottier, 2023.

[73] The spending on compute used to develop frontier AI models has grown at roughly 200% per year.
The cost of AI-relevant compute is falling at about 30% per year, halving every 2 to 3 years.
Improvements in AI algorithms have roughly halved the training compute needed to achieve key results each year.
As evidenced in Compute Trends Across Three Eras of Machine Learning, Sevilla et al., 2022.

[74] Measuring the Algorithmic Efficiency of Neural Networks Hernandez, Brown, 2020;
Algorithmic progress in computer vision, Besiroglu, Erdil, 2022;
A forthcoming analysis by Epoch indicates similar results for language models as in computer vision.

[75] Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning, Sevilla et al., 2022.

[76] Solving Quantitative Reasoning Problems with Language Models, Gur-Ari et al., 2022.

[77] Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2022.

[78] WebGPT: Improving the factual accuracy of language models through web browsing, OpenAI, 2023.

[79] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al., 2022.

[80] AI capabilities can be significantly improved without expensive retraining, Tom Davidson, Jean-Stanislas Denain and Pablo Villalobos (forthcoming).

[81] Scaling Laws for Neural Language Models, Kaplan et al., 2020;

Training Compute-Optimal Large Language Models, Hoffman et al., 2022.

[82] FLOP/S are 'floating point operations per second' and measure the computing performance of a computer.

[83] Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Srivastava et al., 2022;
Extrapolating performance in language modeling benchmarks, Owen, 2023.

[84] Emergent Abilities of Large Language Models, Wei et al., 2022.

[85] Are Emergent Abilities of Large Language Models a Mirage?, Schaeffer et al., 2023.

[86] AI investment forecast to approach $200 billion globally by 2025, Goldman Sachs, 2023.
Projecting compute trends in Machine Learning, Besiroglu et al., 2022.

[87] Expanding access to safer AI with Amazon, Anthropic, 2023;

Microsoft and OpenAI extend partnership, Microsoft, 2023.

[88] In the near term, supply bottlenecks may delay the expansion of AI hardware. See Supply chain shortages delay tech sector's AI bonanza, Financial Times, 2023

[89] Increased spending has been the most significant recent driver of recent compute scaling. The compute to train GPT4 reportedly cost $50m. Recent spending on compute has grown by an estimated 3X/year. If that trend continues for another 7 years, the compute for a training run would cost $150b.

[90] Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning, Villalobos, 2022.

[91] When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale, Marion et al. 2023;

Let's Verify Step by Step, Lightman et al., 2023;

Solving Quantitative Reasoning Problems with Language Models, Lewkowycz et al., 2022;

LIMA: Less Is More for Alignment, Zhou et al., 2023;

Constitutional AI: Harmlessness from AI Feedback, Bai et al., 2022.

[92] A Generalist Agent, Reed et al., 2022.

[93] ACT-1: Transformer for Actions, Adept AI Blog, September 2022;

WebGPT: Browser-assisted question-answering with human feedback, Hilton et al., 2022.

[94] Constitutional AI: Harmlessness from AI Feedback, Anthropic, December 2022.

[95] Evaluation of OpenAI Codex for HPC Parallel Programming Models Kernel Generation, Godoy et al., 2023;

OpenAI Codex, OpenAI, 2021.

[96] EvoPrompting: Language Models for Code-Level Neural Architecture Search, Chen et al., 2023.

[97] Continuous doesn't mean slow, Davidson, 2023; Reframing Superintelligence, section 1.3, Drexler, 2019.

[98] OpenAI's charter says it intends to build "highly autonomous systems that outperform humans at most economically valuable work." OpenAI Charter (2018). DeepMind describes their mission more succinctly as "solving intelligence."About DeepMind, accessed 27 September 2023. Multiple AI developers are reportedly trying to build autonomous AI agents.

[99] Introducing Superalignment, OpenAI, 2023;

Google DeepMind CEO Demis Hassabis Says Some Form of AGI Possible in a Few Years, WSJ, 2023

[100] We are referring to three surveys:
1. Grace et al, 2018, When will AI exceed human performance? Evidence from AI experts, Journal of Artificial Intelligence Research, 62, 729-754.
2. Zhang et al., 2022 Forecasting AI progress: Evidence from a survey of machine learning researchers
3. 2022 Expert Survey on Progress in AI, AI Impacts.
The number of respondents for the three surveys, respectively, were 406, 296 and 734. The third of these surveys has recently been critiqued, in part for having a low response rate of 17%. The first two surveys had response rates of 20%; both searched for evidence of response bias and did not find evidence of significant bias, but there may be a bias from unmeasured variables.

[101] AI Timelines: Where the Arguments, and the "Experts," Stand, Karnofsky, 2021.

[102] What Do NLP Researchers Believe? Results of the NLP Community Metasurvey, Michael et al., 2022; Artificial General Intelligence Is Not as Imminent as You Might Think, Marcus, 2022.

[103] A brief history of AI: how to prevent another winter, Bottino et al., 2021

[104] See, for example, Unsolved Problems in ML Safety, Hendrycks et al., 2021.

[105] In Chu et al. 2017 a model unexpectedly learns to "hide" information about a source image in a generated image in a nearly imperceptible high-frequency signal. In Amodei et al. (2017) an AI trained from human feedback to grasp a ball instead learnt to trick the evaluators into thinking it had grasped the ball, by placing its claw between the ball and the camera. In Bird et. al. (2002), an evolutionary algorithm designed to produce an oscillator instead produced an "antenna" that picked up on signals from the surrounding environment.

[106] Towards Out-Of-Distribution Generalization: A Survey, Liu et al., 2021.

[107] From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads, Charan et al., 2023;

Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, Hazell, 2023.

[108] Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study, Liu et al., 2023.

Jailbroken: How Does LLM Safety Training Fail?, Wei et al., 2023.

[109] Are aligned neural networks adversarially aligned?, Carlini et al., 2023.

Image Hijacks: Adversarial Images can Control Generative Models at Runtime, Emmons et al., 2023.

[110] Universal and Transferable Adversarial Attacks on Aligned Language Models, Zou et al., 2023.

[111] Universal and Transferable Adversarial Attacks on Aligned Language Models, Zou et al., 2023.

[112] Is Robustness the Cost of Accuracy? -- A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, Su et al., 2018
https://openaccess.thecvf.com/content_ECCV_2018/html/Dong_Su_Is_Robustness_the_ECCV_2018_paper.html

[113] Adversarial Policies Beat Superhuman Go AIs, Wang et al., 2022.

Adversarial Policies: Attacking Deep Reinforcement Learning, Gleave et al., 2019.

[114] Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective, Burton et al., 2020.

[115] Scalable agent alignment via reward modeling: a research direction, Leike et al., 2018;

Building safe artificial intelligence: specification, robustness, and assurance, DeepMind Safety Research, 2018;

Constitutional AI: Harmlessness from AI Feedback, Kaplan et al., 2022.

[116] Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, Casper et al., 2023.

[117] Aligned with whom? Direct and social goals for AI systems, Korinek et al., 2022;
Open Problems in Cooperative AI, Dafoe et al. 2020;
Artificial Intelligence, Values, and Alignment, Gabriel, 2020;
A proposal for importing society's values, Leike, 2023.

[118] Zoom In: An Introduction to Circuits, Olah et al., 2020;

Language models can explain neurons in language models, OpenAI, 2023.

[119] Open problems and fundamental limitations of reinforcement learning from human feedback, Casper et al., 2023;

Emergent Abilities of Large Language Models, Wei et al., 2022;

How is ChatGPT's behavior changing over time?, Chen et al., 2023.

Lima: Less is more for alignment, Zhou et al., 2023.

[120] Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, Variengien et al., 2022;

Towards Automated Circuit Discovery for Mechanistic Interpretability, Conmy et al., 2023;

Progress measures for grokking via mechanistic interpretability, Chan et al., 2023;

A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations, Chughtai et al., 2023;

Decomposing Language Models Into Understandable Components, Anthropic, 2023.

[121] Sanity Checks for Saliency Maps, Adebayo et al., 2018;

The (Un)reliability of saliency methods, Kindermans et al., 2017;
A Benchmark for Interpretability Methods in Deep Neural Networks, Hooker et al., 2018.

[122]For example, see Probing Classifiers: Promises, Shortcomings, and Advances | Computational Linguistics, MIT Press.

[123] A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks

[124] Recognition in Terra Incognita

[125] *Limitations and Hazards*, Model evaluation for extreme risks, Shevlane et al., 2023.

[126] Certified Defenses against Adversarial Examples

Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks

Provable defenses against adversarial examples via the convex outer adversarial polytope

Differentiable Abstract Interpretation for Provably Robust Neural Networks

On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models

Certified Adversarial Robustness via Randomized Smoothing

[127] Adversarial Examples Are Not Bugs, They Are Features

[128] An API is a set of rules and protocols that enables integration and communication between AI systems and other software applications.

[129] Structured access: an emerging paradigm for safe AI deployment

[130] See The Gradient of Generative AI Release: Methods and Considerations for more discussion of alternatives and considerations.

[131] Tech leaders including Musk, Zuckerberg call for government action on AI, Washington Post, 2023;

A recent paper proposed a method to increase the cost of fine-tuning for particular capabilities, but it is too early to confidently assess the promise of such approaches. Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models, Henderson et al., 2022

[132] Licences may restrict legitimate use, but are often disregarded cannot alone enforce legitimate use

[133] For example, by training another (e.g. smaller) “student” model to copy the outputs received through the API.Extracting Training Data from Large Language Models, Carlini et al., 2021;

Stealing Machine Learning Models via Prediction APIs, Tramèr et al., 2016.

[134] Alpaca: A Strong, Replicable Instruction-Following Model, Taori et al., 2023.

[135] Nevo & Lahav, forthcoming

[136] Universal and transferable adversarial attacks on aligned large language models. Zou, 2023.

[137] Engines of power: Electricity, AI, and general-purpose, military transformations, Ding & Dafoe, 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models, Eloundou et al., 2023.

Market concentration implications of foundation models, Korinek & Vipra, 2023.

[138] For example, in the UK context, the Telecommunications Security Code of Practice and the Grid Code.

[139] Robust artificial intelligence and robust human organizations, Dietterich, 2019;
Regulating for ‘Normal AI Accidents’, Maas, 2018.

[140] Discerning signal from noise, Schwarz Reisman Institute, 2023;

What will the role of standards be in AI governance?, Ada Lovelace Institute, 2023.

[141] AI risks: The potential negative or harmful outcomes arising from the development or deployment of AI systems.

[142] The roadmap to an effective AI assurance ecosystem, Center for Data Ethics and Innovation, 2021.

[143] Although, many international forums are starting initiatives in this area, such as the Council of Europe, Organisation for Economic Cooperation and Development (OECD), Group of Seven (G7), Global Partnership on AI (GPAI), United Nations (UN), Group of Twenty (G20), and Standards Development Organisations (SDOs), among others. Constructive multilateral and multi-stakeholder engagement across borders will need to continue to effectively address frontier AI risks, as well as capitalise on the opportunities.

[144] Externalities: Prices Do Not Capture All Costs, IMF, 2023.

[145] AI Exemplifies The Free Rider Problem,The Conversation, 2023;

Collective Action on Artificial Intelligence: A Primer and Review, Neufville, 2021;

The role of cooperation in responsible AI development, Askell et al., 2019.

[146] Microsoft CEO Satya Nadella says he hopes Google is ready to compete when it comes to A.I. search: 'I want people to know that we made them dance', Fortune, 2023;

In AI Race, Microsoft and Google Choose Speed Over Caution, New York Times, 2023;

OpenAI reportedly warned Microsoft about Bing's bizarre AI responses, The Verge, 2023.

[147] Modelling and Influencing the AI Bidding War: A Research Agenda, Han et al., 2019;

Artificial intelligence development races in heterogeneous setting, Cimpeanu et al., 2022.

[148] Dafoe, Allan. 'AI Governance: Overview and Theoretical Lenses'. In The Oxford Handbook of AI Governance, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang, 0. Oxford University Press, 2022.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the precipice: a model of artificial intelligence development." AI & Society 31, no. 2 (2016): 201-206., OpenAI Charter (2018): "We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be "a better-than-even chance of success in the next two years."
Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development.

[149] AI foundation models Initial Report, UK Competition and Markets Authority, 2023; there are also concerns about lack of competition on safety, and about regulatory capture.

Market concentration implications of foundation models, Korinek & Vipra, 2023.

[150]- "We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles.": Anthropic's $5B, 4-year plan to take on OpenAI, TechCrunch, 2023.
- Continuous doesn't mean slow, Davidson, 2023

[151] Harms of AI, Acemoglu, 2021.
The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, Zuboff, 2019.

[152] Evaluating the Social Impact of Generative AI Systems in Systems and Society, Solaiman et al., 2019.

[153] An analysis of around 80 ethical frameworks is provided in Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI, Fjeld et al., 2020.

[154] One in three internet users fail to question misinformation. Ofcom, 2022.

[155] Tackling threats to informed decision-making in democratic societies. Seger et al, 2020.

[156] AI foundation models Initial Report, UK Competition and Markets Authority, 2023.

[157] Misinformation: a qualitative exploration. Ofcom, 2021.

[158] People are trying to claim real videos are deepfakes. The courts are not amused, Bond, 2023.
Elon Musk's statements could be 'deepfakes', Tesla defence lawyers tell court, The Guardian, 2023.

[159] For one example of this concern becoming more widespread, see this Guardian article questioning whether you can believe what you see.

[160] Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change, Atillah, 2023.

[161] Google's new A.I. search could hurt traffic to websites, publishers worry, Leswing, 2023.

[162] Discovering Language Model Behaviors with Model-Written Evaluations, Schiefer et al., 2022.

[163] Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence, Birhane, 2022.

[164] Facebook Hosted Surge of Misinformation and Insurrection Threats in Months Leading Up to Jan. 6 Attack, Records Show, ProPublica, 2023.

[165] The Danger of Misinformation in the COVID-19 Crisis, Med, 2020.

[166] How to prepare for the deluge of generative AI on social media. Kapoor and Narayanan, 2023.

[167] AI Chat Assistants can Improve Conversations about Divisive Topics. Argyle et al. 2023.

[168] On the Reliability of Watermarks for Large Language Models, Geiping eal., 2023.

[169] A Systematic Review on Model Watermarking for Neural Networks, Boenisch, 2021;
Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks, Saberi et al., 2023.

[170] Korinek, Anton and Joseph Stiglitz (2019), Artificial Intelligence and Its Implications for Income Distribution and Unemployment. In Ajay Agrawal, Joshua Gans and Avi Goldfarb (eds.), The Economics of Artificial Intelligence, pp. 349-390, NBER and University of Chicago Press, May 2019.

[171] Viewpoint: The future of work in agri-food, Christiaensen et al., 2019.

[172] The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty, Autor, 2022.

[173] Why Are There Still So Many Jobs? The History and Future of Workplace Automation, Autor, 2015.

[174] The labour market impacts of technological change: from unbridled enthusiasm to qualified optimism to vast uncertainty. David Autor, 2022

[175] The Potentially Large Effects of Artificial Intelligence on Economic Growth, Goldman Sachs, 2023, Exhibit 5.

[176] Automation and New Tasks: How Technology Displaces and Reinstates Labor, Acemoglu & Restrepo, 2019.

[177] Ethical and social risks of harm from Language Models, DeepMind, 2021;
Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST, 2022.

[178] Let's talk about biases in machine learning!, Jernite, 2022.

[179] Implications of predicting race variables from medical images, Zou et al., 2023;

What about fairness, bias and discrimination?, ICO.

[180] Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence, Birhane, 2022;

Moving beyond "algorithmic bias is a data problem", Hooker, 2021.

[181] Power to the People? Opportunities and Challenges for Participatory AI, Birhane et al., 2022.

[182] Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models, Rauh et al., 2022;

Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?, The BMJ, 2021.

[183] 1.3 Longer-term deployment and diffusion, Strengthening Resilience to AI Risk, Janjeva et al., 2023.
[184] Towards Measuring the Representation of Subjective Global Opinions in Language Models, Esin Durmus et al., 2023 .
[185] The Right to Explanation, Vredenburgh, 2022.
[186] The Coming Infocalypse: What You Urgently Need To Know, Schick, 2020.

[187] When is automated decision making legitimate?, Barocas, Hardt, Narayanan, 2022;

1.2 Deployment and usage, Strengthening Resilience to AI Risk, Janjeva et al., 2023.

[188] Taxonomy of Risks posed by Language Model, Weidinger et al., 2022.

[189] Fairness in AI and Its Long-Term Implications on Society, Bohdal et al., 2023.

[190] A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity
[191] The Capacity for Moral Self-Correction in Large Language Models
[192] Annex B - Safety and Security risks from GenAI, HM Government, 2023.
[193] The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.
[194] The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.

[195] Can Large Language Models Democratize Access to Dual-use Biotechnology?, Soice et al., 2023

[196] The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.

[197] ChemCrow: Augmenting Large-Language Models with Chemistry Tools, Bran et al., 2023.

[198] Cloud labs: remotely controlled automatised biochemical laboratories.
[199] Emergent Autonomous Scientific Research Capabilities of Large Language Models, Boiko et al., 2023.

[200] Biological design tools: AI systems trained on biological data that can help design new proteins or other biological agents.
Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools, Sandbrink, 2023.

[201] Highly Accurate Protein Structure Prediction with AlphaFold', Jumper et al., 2021.

[202] De Novo Design of Protein Structure and Function with RFdiffusion, Watson et al., 2023.

[203] De Novo Design of Protein Structure and Function with RFdiffusion, Watson et al., 2023;

Comprehensive AAV Capsid Fitness Landscape Reveals a Viral Gene and Enables Machine-Guided Design, Ogden et al., 2019.

The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.

[204] De Novo Design of Protein Structure and Function with RFdiffusion, Watson et al., 2023.

Ankh☥: Optimized protein language model unlocks general-purpose modelling, Ahmed, et al., 2023.

Large language models generate functional protein sequences across diverse families, Madani et al., 2023.

[205] ChemCrow: Augmenting Large-Language Models with Chemistry Tools, Bran et al., 2023.

[206] ChemCrow: Augmenting Large-Language Models with Chemistry Tools, Bran et al., 2023.

[207] The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.

[208] The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe, Nuclear Threat Initiative, forthcoming.

[209] DARPA, White House launch $20M AI, cybersecurity challenge, Breaking Defense, Gill, 2023.

[210] The New Risks ChatGPT Poses to Cybersecurity, Chilton, 2023;
A Hazard Analysis Framework for Code Synthesis Large Language Models, Mishkin et al., 2022.
[211] The New Risks ChatGPT Poses to Cybersecurity, Chilton, 2023.

[212] Chatting Our Way Into Creating a Polymorphic Malware Shimony & Tsarfati, 2023;

BlackMamba: Using AI to Generate Polymorphic Malware, Sims, 2023.

[213] OPWNAI : Cybercriminals Starting to Use ChatGPT, Check Point, 2022.
[214] Intelligent Reconnaissance: How AI Tools Drive Effective Penetration Testing, CQR, 2023.

[215] Does your boss sound a little funny? It might be an audio deepfake, Alspach, 2022

[216] The security threat of AI enabled cyberattacks, TRAFICOM, 2022;
Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, Hazell, 2023.

[217] Cyber security breaches survey 2023, the Department for Science, Innovation and Technology, 2023.

[218] DeepLocker - Concealing Targeted Attacks with AI Locksmithing, BlackHat USA events, 2018.

[219] Introducing 100k Context Windows, Anthropic, 2023;

OpenAI is testing a version of GPT-4 that can 'remember' long conversations, TechCrunch, 2023.

[220] ARC (Challenge) Benchmark (Common Sense Reasoning), Papers With Code.

[221] How we cut the rate of GPT hallucinations from 20%+ to less than 2%, Jason Fan, 2023.

[222] Automating Cyber Attacks, Buchanan et al.,, 2020.

[223] Evaluating Language-Model Agents on Realistic Autonomous Tasks, ARC Evals, 2023.

[224] How Security Analysts Can Use AI in Cybersecurity, Moisset, 2023.

ChatGPT Vulnerability Scanner Is Pretty Good, Merian, 2021.

How AI Is Disrupting And Transforming The Cybersecurity Landscape, Ravichandran, 2023.

[225] Examining Zero-Shot Vulnerability Repair with Large Language Models. Pearce et al, 2023.

[226] Virtual Prompt Injection for Instruction-Tuned Large Language Models, Yan et al., 2023.

[227] Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid, Zetter, 2016.

[228] Northern's ticket machines hit by ransomware cyber attack, BBC News, 2023.

[229] Fears for patient data after ransomware attack on NHS software supplier, Milmo & Campbell, 2022.

[230] Timeline of Cyber Incidents Involving Financial Institutions, Carnegie Endowment for International Peace.

[231] Cyber security breaches survey 2023, the Department for Science, Innovation and Technology, 2023.

[232] Data breaches putting domestic abuse victims' lives at risk, says UK watchdog , Booth, 2023.

[233] On the receiving end - how post can enable domestic abuse, Citizens Advice.

[234] Personal data breaches: a guide, ICO.

[235] US mother gets call from 'kidnapped daughter' – but it's really an AI scam | Arizona, Salam, 2023; ; Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes, FBI, 2023.

[236] Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity, Zeedaly et al., 2020.

[237] How Much Money Could Large Language Models Save Propagandists?, Musser, 2023.

[238] Disinformation Researchers Raise Alarms About A.I. Chatbots, NYT, 2023.

[239] Release Strategies and the Social Impacts of Language Models, Askell et al., 2019.

The Existential Threat of AI-Enhanced Disinformation Operations, Honigberg, 2022.

[240] How Much Money Could Large Language Models Save Propagandists?, Musser, 2023.

[241] FaceForensics: A large-scale video dataset for forgery detection in human faces, Rossler et al., 2019.

[242] How AI will transform the 2024 elections, West, 2023.

[243] Sam Altman sells superintelligent sunshine as protestors call for AGI pause, Vincent, 2023.

[244] AI Deception: A Survey of Examples, Risks, and Potential Solutions, Park et al., 2023.

[245] Forecasting Potential Misuses of Language Models for Disinformation Campaigns—and How to Reduce Risk, Goldstein et al., 2023.

[246] Fake news, disinformation and misinformation in social media: a review, Aïmeur, 2023

[247] Language Models are Few-shot learners. Brown et al. 2020.

[248] AI Causes Real Harm. Let's Focus on That over the End-of-Humanity Hype, Bender, Hannah, 2023

[249] For example:

1. Professors Geoffrey Hinton and Yoshua Bengio, the two most highly cited computer scientists of all time, and winners of the Turing award for their pioneering work in deep learning, have both highlighted concerns around catastrophic harms from loss of control.

2. In 2023, a number of world-leading experts from academia, industry, and civil society asserted that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Among them were the CEOs of DeepMind, OpenAI, and Anthropic, and over 100 AI professors.

[250] The Fallacy of AI Functionality, Kumar et al., 2022.

[251] Recommender systems and the amplification of extremist content, Whittaker et al., 2021.

[252] How IBM Watson Overpromised And Underdelivered On AI Health Care, Eliza Strickland, 2019.

[253] The Pain Was Unbearable. So Why Did Doctors Turn Her Away?, Maia Szalavitz, 2021.

[254] Harms of AI, Acemoglu, 2023.
Harms from Increasingly Agentic Algorithmic Systems, Chan et al., 2023;

[255] A survey by Deloitte found that 43% of users of generative AI falsely believe it always produces factually correct outputs and 38% believe it is unbiased. More than four million people in the UK have used Generative AI for work, Deloitte (2023).

[256] Algorithm appreciation: People prefer algorithmic to human judgement, Logg et al. 2019.
Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making, Suresh at al., 2020.

[257] Research by Anthropic found LLMs repeat back user's preferred political views: Discovering Language Model Behaviors with Model-Written Evaluations, Perez, E, et al., 2022.

[258] The danger of anthropomorphic language in robotic AI systems, Brookings, 2021,
Google's 'deceitful' AI assistant to identify itself as a robot during calls, The Guardian, 2018.

[259] Overreliance on AI: Literature Review, Passi et al., 2022.

[260] As discussed in the cross-cutting risks section, alignment is a long-standing and unsolved technical research problem and methods for understanding how AI systems work (including whether they are aligned) are immature.

[261] AI has been rapidly changing financial systems in complex ways. For example, see Artificial intelligence and systemic risk, Daníelsson et al., 2022.

[262] For a description of what this dynamic might look like, see What Failure Looks Like and What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs).

[263] See Yoshua Bengio's FAQ on catastrophic AI risks and his testimony to the US Senate, July 2023.

[264] CONTEST: The United Kingdom's Strategy for Countering Terrorism, Home Office, 2023.

[265] Richard Sutton's talk entitled AI Succession, 2023;
Accelerationism: how a fringe philosophy predicted the future we live in, Beckett, 2017;
Effective accelerationism.

[266] If some safeguards limit AI capabilities, then these bad actors might be able to deploy more capable AI systems than responsible actors. They might be willing to remove safeguards which increase human control but also limit AI capabilities.

[267] This includes both the difficulty of *developing* an AI system with the intended goal and difficulty of *evaluating* whether an AI system has an intended goal. See discussion in the earlier section of the report on cross cutting technical risk factors. See also:
Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, Casper, 2023;

The Alignment Problem from a Deep Learning Perspective, Ngo et al., 2022.

[268] Goal misgeneralization: Why correct specifications aren't enough for correct goals, Shah et al., 2022.
Goal Misgeneralization in Deep Reinforcement Learning, Langosco et al, 2021.
The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models, Pan et al, 2022.
Specification gaming: the flip side of AI ingenuity, Krakovna et al., 2020
Learning from human preferences, Amodei et al., 2017

[269] Chapter 7, Superintelligence: Paths, Dangers, Strategies, Nick Bostrom, 2014.

[270] The Alignment Problem from a Deep Learning Perspective, Ngo, 2022;

Is Power-Seeking AI an Existential Risk?, Carlsmith, 2022.

[271] AI Causes Real Harm. Let's Focus on That over the End-of-Humanity Hype, Bender, 2023.

[272] Stuart Russell calls for new approach for AI, a ‘civilization-ending’ technology, Leven, 2023.

[273] Manipulation capabilities could give frontier AI many more routes to increasing its future influence and causing harm.For example, many people can be manipulated into giving away money. In 2022, nearly 70,000 people reported a romance scam to the FTC, with reported losses exceeding $1b: Romance scammers’ favorite lies exposed, FTC, 2022.

[274] My Chatbot Companion - a Study of Human-Chatbot Relationships, Skjuve et al., 2021.

[275] For $300, Replika sells an AI companion who will never die, argue, or cheat — until his algorithm is updated, Sangeeta, 2023.

[276] Discovering Language Model Behaviors with Model-Written Evaluations, Schiefer et al., 2022.

[277] Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models, O’Gara, 2022.

[278] Interacting with Opinionated Language Models Changes Users' Views, Jakesch et al., 2022.

[279] Evaluating Language-Model Agents on Realistic Autonomous Tasks, Kinniment et al., Aug 2023.