Gladstone AI

Survey of AI Technologies and AI R&D Trajectories

This survey was funded by a grant from the United States Department of State. The opinions, findings and conclusions stated herein are those of the author and do not necessarily reflect those of the United States Department of State.

Organization

Gladstone AI Inc. (hello@gladstone.ai)

Authors

Jeremie Harris*

Edouard Harris

Mark Beall

* Lead and corresponding author.
Contact: jeremie@gladstone.ai

Submitted on November 3, 2023

Table of Contents

List of Figures

List of Tables

Executive summary

AI is a dual-use technology with disproportionately positive applications. It is a key driver of economic growth, and holds promise for achieving unprecedented breakthroughs in areas such as medicine, energy, and climate science. However, AI is also introducing acute risks that must be managed at the same time as its value is harnessed. In particular, over the next four years, AI systems will plausibly introduce WMD-level risks from deliberate weaponization, accidents, or loss of control.

In 2019, the most cutting-edge AI systems struggled to generate coherent paragraphs. Today, in 2023, AI systems can already support automated superscale phishing attacks, identity theft and scams via high-fidelity voice cloning, autonomous hacking agents, mass persuasion campaigns, and the re-engineering of biochemical compounds by untrained users. These capabilities have emerged suddenly within the last 18 months, and would have been considered the stuff of science fiction as recently as 2020.

In the near future, AI systems are likely to support far more dangerous applications. The trend is clear: larger AI models that are trained with more computational power and data are consistently more capable than smaller ones, and the amount of computational power used to train the leading AI models increases each year by a factor of four. This rapid expansion of computational effort has been the main driving force for today's ongoing and unprecedented acceleration in AI capabilities.

The effectiveness of scaled computing power has come as a shock even to some of the field's most forward-thinking observers. Today, executives and researchers at the world's top AI labs consider it plausible that AI systems capable of matching or exceeding human performance across all economically useful tasks (sometimes known in the AI industry as artificial general intelligences, or AGIs) may be developed within the next five years.

The development of AGI, and of AI capabilities approaching AGI, would introduce catastrophic risks unlike any the United States has ever faced. It is now plausible that the next generation of AI systems – those trained at the next level of computational scale – will be so capable that they will lead to WMD-like risks if and when they are weaponized. Publicly and privately, researchers at frontier AI labs have voiced concerns that AI systems developed in the next 12 to 36 months may be capable of executing catastrophic malware attacks, assisting in bioweapon design, and directing swarms of goal-directed humanlike autonomous agents, for example. Given the demonstrated effectiveness of scaled computing power in AI development, and the current

capabilities of frontier AI systems, it is very likely that AI systems will demonstrate at least some of these capabilities over the coming three years.

Additionally, a significant body of evidence suggests that AI systems whose capabilities exceed a certain (currently unknown) threshold may become challenging for humans to control. Research suggests a risk that a capable enough AI system could begin to follow dangerous strategies in pursuit of objectives that are incompatible with continued human welfare. Such systems could pose large-scale and irreversible risks, even without any specific intent on the part of their developers. Concerns over these extreme risks have been echoed by academic leaders including two of the founders of the field of deep learning, and by civil society groups with deep expertise in AI safety.

AI systems with potentially dangerous capabilities are being developed and proliferated through several key sources and pathways. First, proprietary AI systems are being trained and often deployed as products by frontier AI labs. These include Google DeepMind (Bard, Gemini), OpenAI (DALL-E 3, GPT-4), and Anthropic (Claude 2). Unfortunately, these labs largely lack the security measures required to prevent exfiltration of their models by nation-state actors, or the safety measures required to prevent catastrophic accidents from future AI systems.

Second, the AI community has a strong open-source culture. Open-source advances at the framework level have led to significant breakthroughs such as Auto-GPT, which allows language models to operate autonomously as agents. Additionally, the capabilities of open-source and open-access AI models are closely tracking those of the most advanced proprietary AI systems, in some cases lagging the frontier by less than a year. The result has been an irreversible proliferation of increasingly weaponizable AI systems.

Model theft, piracy, and sale of powerful AI models are also vectors of proliferation for high-risk AI capabilities. In each case, a third party who acquires a proprietary model – whether legally or otherwise – can augment the model for weaponization on a budget in the hundreds of dollars. The low cost associated with these capability augmentations makes advanced AI a fundamentally new type of WMD-like technology, and its effects particularly challenging to control.

Advanced AI's rapid progress, unpredictable emergence of capabilities, competitive pressures, and open-source proliferation, combine to make the field an unprecedented national security and public policy challenge.

Note: This report will be most accessible to technical audiences already familiar with deep learning. Readers who are not familiar with deep learning should consider starting with the nontechnical primer included in Annex D.

1. Background on AI

1.1 Definitions

Before 2020, artificial intelligence (AI) systems could only perform specific narrow tasks. For example, an artificial neural network trained to classify images as either containing or not containing a picture of a dog could perform only that task and no other [1]. But with the announcement of OpenAI's GPT-3 [2] in early 2020, researchers began to make rapid progress in developing flexible, general-purpose AI systems with the capacity to perform an ever-widening range of tasks [3–5]. Today, these include systems that can outperform most human beings on tasks as diverse as writing code, translating between languages, composing essays, and answering general knowledge questions.

The pace of this progress has surprised some of the field's most optimistic researchers and industry insiders [6]. It has been driven, in part, by experimental insights into how AI systems improve as they get larger and are trained with more data and computing power [7,8]. These insights have directly enabled the development of increasingly human-like chatbots and other kinds of general-purpose AI systems [5,9–11].

We will refer to AI systems that can perform many distinct tasks without being specifically trained on them as ***advanced AI***[1]. ChatGPT is a typical example of an advanced AI system, but several others exist as of late 2023. An advanced AI system might use only a single type of data (such as text data), or it might use many different types of data (such as text, images, audio, and/or video data) to perform its tasks [3,10,12–14].

In addition to advanced AI, we will use the term ***artificial general intelligence***[2] (AGI) as it is used in industry [15–17], to refer to advanced AI systems that exceed human capabilities across all, or nearly all, economic and strategic domains.

1.2 Progress in advanced AI

A key long-term goal of advanced AI research has been to create AGI [16,18]. Before 2020, the narrow usefulness of even the most cutting-edge AI systems led most researchers to conclude that this goal was at least decades away [19]. In addition, a

[1] See **Annex A: Glossary of terms** for our definition of advanced AI.

[2] See **Annex A: Glossary of terms** for our definition of artificial general intelligence.

large portion of the AI research community believed that AGI could not be achieved without fundamentally new AI paradigms [20,21].

But starting in 2020, several independent lines of evidence emerged to suggest that an AI system's capabilities could in fact be systematically improved by doing nothing more than scaling up its training dataset, model size, and training compute [7,22]. That year, OpenAI released the first AI system that took advantage of these scaling laws. That AI system was GPT-3, the first large language model (LLM). With over ten times more parameters than any previous language model, GPT-3 could perform unprecedentedly well on a wide variety of tasks after being given only a few examples of them [2].[3]

In early 2021, Google's DeepMind demonstrated a significant improvement over OpenAI's GPT-3 scaling laws with its new advanced AI model Chinchilla [8]. DeepMind's work on Chinchilla demonstrated a new scaling law that showed researchers how to train more capable AI models, more cheaply, by using larger training datasets. Since then, leading AI systems have been developed with ever larger compute budgets, increasing more than fourfold every year (see Figure 1).

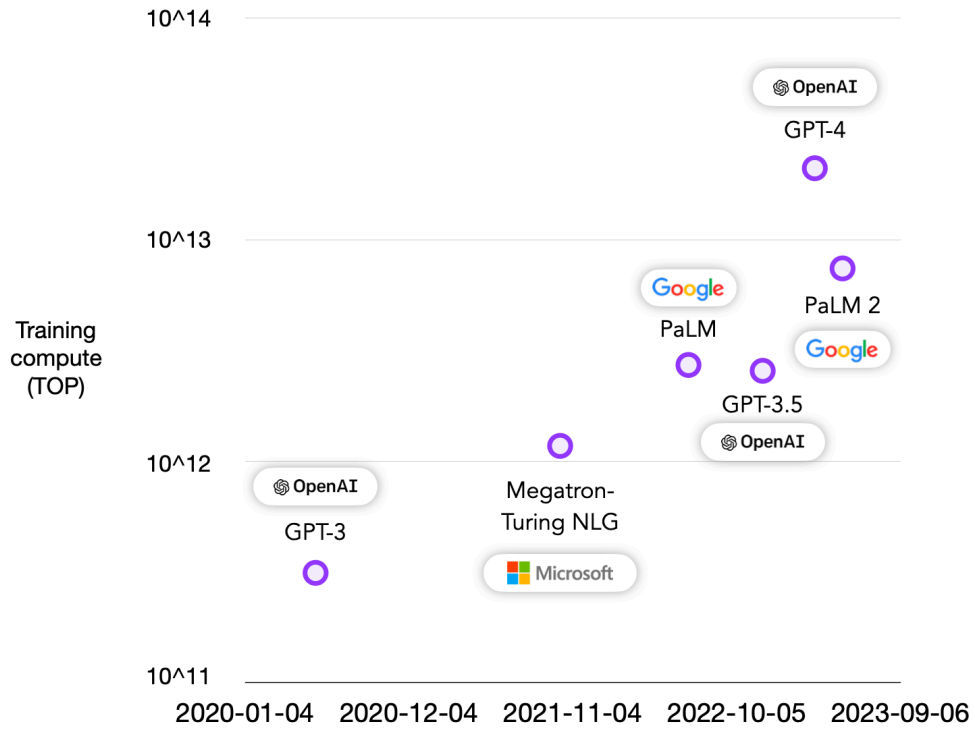


Figure 1. Training compute of notable machine learning systems since GPT-3. Data source: Epoch AI.

[3] **Annex D: Nontechnical primer on AI**, section 2.6, provides a nontechnical overview of AI scaling.

In 2022 DeepMind published Gato, a single AI model that outperformed more than half of human experts at 450 of the 600 tasks it was trained on [3]. These tasks included captioning images, carrying out dialogues, and controlling robotic arms. Around the same time, Google announced PaLM, a 540 billion parameter model that achieved state-of-the-art performance on hundreds of different language tasks [4]. Later that same year, OpenAI released ChatGPT, widely viewed as the culmination of decades of effort in conversational AI [9] and the first product of any kind to reach 100 million users in under two months [23]. Then, in early 2023, OpenAI announced GPT-4 which, among many other record-setting capabilities, scores in the 90th percentile on tests as diverse as the Uniform Bar Exam, AP U.S. History, and the Math SATs [10].

These developments are part of a greater trend: an ongoing and increasingly competitive race among the world's most advanced AI labs to scale up the sizes and capabilities of their most powerful AI systems [24–28]. This race has already delivered AI systems that achieve unprecedented performance on tasks such as image generation [29–31] and recognition [32], web development [28], and robotic control [33]. These new AI systems are especially notable given that they were never directly trained to perform many of the tasks they were tested on, but were instead able to work from an explanation alone without seeing any examples.

Before these advances, many researchers believed that the deep learning paradigm could not support this degree of generalization to novel tasks [24,25]. But today, it seems increasingly likely that AGI could be achievable with few or even no further major conceptual breakthroughs in AI [34–36]. Recent large language models in particular have shown a range of logical, historical, mathematical, and scientific capabilities that suggest that the model is able to represent the world internally in at least some degree of depth [37]. For example, GPT-4 has demonstrated an ability to correctly navigate through complicated virtual worlds via text interactions alone, strongly implying that the model can create and retain an internal picture of a simple environment [37]. These hypothesized internal representations are called world models, and they appear to be related to an AI system's ability to generalize across a broader range of tasks. Most recently, AI researchers have developed multi-modal training schemes that can connect LLM world models to sensory inputs from external objects such as video, image, or audio feeds [14,38]. Multi-modal training lets an AI system learn directly from real-world data rather than solely relying on text inputs.

AI scaling laws continue to be refined as researchers' understanding of the optimal balance between data, compute budget, and model size improves [7,8,11]. But scaling itself remains critical: larger AI systems are consistently more capable than smaller

ones, and this has remained true as the largest AI systems have grown 10-million-fold over the past decade [39]. If this robust trend continues, it is conceivable that within a few years the most advanced AI systems could surpass expert human performance in *every* domain. Already AI systems outperform most humans at tasks that range from taking standardized tests [10], image recognition [40], and video game-playing [41,42], to strategy and scenario planning [43]. Notably, most of these performance breakthroughs have occurred just in the three years since 2020. Increasingly general and analytical tasks, that were once believed to be beyond the reach of current AI techniques, are being mastered by AI systems at a rapidly accelerating rate.

Despite its high cost in computing power, scaling remains the most prominent path to developing more powerful AI systems. But researchers also use several other approaches to improve AI capabilities. These include refining data collection and preprocessing techniques, optimizing a model's training objective, and using more performant model architectures. These kinds of techniques are often called algorithmic improvements, to distinguish them from improvements due to scaling. Unlike scaling, once a new algorithmic improvement is known, it can be applied at no additional cost to train more performant models for the same computing budget [11].

Significant improvements in AI capabilities have also come from straightforward extensions to known methods, including prompt engineering [44–46] and fine-tuning [47] existing models. It does not seem implausible that, even barring further scaling, continued algorithmic improvements could produce AI systems that collectively surpass median human performance across virtually all domains.

1.3 Transformers and language models as enablers for advanced AI

Most of today's advanced AI systems, including all prominent LLMs, are built on the transformer architecture. A transformer model contains a special subsystem, called an attention mechanism, that identifies which parts of a text prompt the model should pay most attention to as it generates its output. With the ability to focus more on some parts of a prompt than others, and to capture long-range interactions between different parts of their prompts, transformers are especially well suited to read and interpret text. Crucially, transformers also run very efficiently on modern computing hardware, making them an ideal architecture to channel the vast quantities of data and processing power that are needed for AI scaling [48].

Scaled LLMs in particular have drawn attention as a potential means by which AGI could be achieved. During training, an LLM learns to perform a task similar to text

autocomplete, in which it receives a piece of text as input, and is scored on how well it can predict which text token will come next at each position in the text. To do this successfully, the LLM needs to use the information in the input text to inform its next-token prediction. For example, when presented with a prompt like "to counter rising inflation, the U.S. should", an LLM must be able to "understand" what is meant by terms such as "inflation" and "the U.S.", and leverage implicit knowledge of macroeconomics, monetary theory, and other concepts to generate a sensible prediction. Because of this, a well-trained LLM is effectively forced to internalize a robust representation of the world during training – a so-called world model, as discussed previously.

An LLM's world model endows it with implicit knowledge about a wide range of concepts, which it can then apply to downstream tasks unrelated to its initial autocomplete-like training objective. This allows an LLM to serve both as a source of general knowledge, and as a reasoning engine for robotic controllers, image generators, code-writing assistants, and many other kinds of systems [28,29,33].

1.4 Limitations of transformers and alternative paths to advanced AI

Transformer-based systems currently dominate the ***frontier AI***[4] paradigm, because of their scalability and other favorable characteristics. But they also have several important limitations. Some of these limitations are fundamental to the problem of intelligence. Others are specific to transformers themselves. The limitations of transformers can be broken down into four categories: capacity, veracity, reliability, and latency.[5]

- **Capacity** is a broad term that encompasses both an AI model's memory requirements, and the maximum sequence length it can process. Transformer models require a large amount of memory and computational resources to store and update their parameters during training, and this is especially true for large-scale transformers with billions or trillions of parameters. Increasing a transformer's maximum sequence length makes the model more powerful, at the cost of increasing latency and memory requirements [49].

[4] See **Annex A: Glossary of terms** for our definition of frontier AI.

[5] **Annex D: Nontechnical primer on AI**, section 2.24, provides a nontechnical overview of the potential limitations associated with scaling transformers as a path to AGI.

- **Veracity** refers to the truthfulness of an AI system. Transformer models are known to hallucinate, giving responses that are not justified by the training data or any other available context. A hallucination often takes the form of false or logically incoherent text outputs. Hallucination emerges as a byproduct of optimizing systems for tasks such as next-word prediction, because the most likely next-word prediction is not necessarily the most true or accurate. Pretrained LLMs have no direct incentive to produce true or valid outputs – they only have an incentive to produce outputs that are *highly probable*, and probable outputs are not always truthful outputs. The challenge of preventing hallucination is known as contextual grounding [50].

- **Reliability** breaks down into two properties: calibration and brittleness. Transformers can suffer from both poor calibration and poor brittleness [51].

    - A poorly calibrated model is one that is either too confident, or not confident enough, in its predictions. Even modern transformers often exhibit poor calibration.

    - A brittle model is one that will generate very different predictions when given very similar inputs. For example, we would not normally expect a model's predictions to be dramatically different for "The cat sat on a ___" versus "The cat sat on the ___", but a brittle model may generate completely different predictions in these two cases.

- **Latency** refers to the time it takes for a model to generate its output. The attention mechanism introduces significant latency in large-scale transformer-based AI systems. There have been some efforts to mitigate this at either the algorithmic (software) or implementation (hardware) levels, with notable recent successes [52].

Progress is being made in overcoming many of these limitations. Frontier AI labs are now building transformers with long-term memory [53, 54], giving them access to tools that allow them to retrieve information from the internet and take actions in cyberspace [55], and grounding them by training on a wider range of data types [10,13]. They are also addressing latency limitations: FlashAttention, an optimized implementation of attention, has led to up to 3x improvements in training times [56]. It remains to be seen whether advances such as these, compounded by further scaling, will allow transformers to strongly exceed expert human performance on all human-solvable tasks.

If scaling transformers alone fails to lead to AGI, several alternative or complementary approaches may do so instead. These include strategies based on reinforcement learning and self-play [57], as well as recurrent architectures [58]. Notably, Google's Gemini model will feature a composite architecture consisting of a scaled transformer, integrated with a tree search-based long-term planning component [59,60].

Whichever approach successfully culminates in AI systems with broadly human-level capabilities, it will very likely operate as a platform for applying computational resources at massive scale on very large datasets. In this sense, at least, scale is likely to be a key input to the successful development of AGI.

2. Industry state of play

2.1 Notable players, their products and capabilities

To train an LLM, an AI developer typically starts with a vast dataset of text scraped from the Internet. That dataset is carefully preprocessed, filtered using smaller specialized AI models, and broken down into smaller parts to prepare it for LLM training. Although the general procedure is described in the AI research literature, the proprietary techniques needed to carry it out successfully at scale are not widely known outside the most advanced AI labs. As a result, there is a limited pool of expertise available to effectively train high-quality, state-of-the-art AI systems.

Scaling pre-trained AI systems at the cutting edge also requires large and increasing quantities of computing power [61][6]. GPT-3's 2020 training run is estimated to have cost roughly $5M [62], though training a similar-quality model in 2023 would be expected to cost 10x less [63] because of the algorithmic and hardware improvements that have occurred since then. For comparison, GPT-4's more recent training run is estimated to have cost between $40M and several hundred million dollars to train [64,65]. Today's cutting-edge models are being trained on between 10,000 and 100,000 GPUs, with associated compute budgets in the hundreds of millions to billions of dollars [66,67].

This means that currently only well-resourced organizations with access to scarce technical talent and knowledge are able to develop the most capable advanced AI models from scratch. Here are some of the most notable of these organizations, along with their flagship products and capabilities current as of November 2023 (see Table 1 for a summary of this information):

- **OpenAI** is a California-based company in which Microsoft owns a 49% stake [48]. Their mission is to "ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity." [16] OpenAI was the first organization to invest heavily in AI scaling, and delivered many of the early proof points for the scaling hypothesis [2,7]. They are among a very small number of labs that are shaping the frontier of AI capabilities today. Their

[6] **Annex D: Nontechnical primer on AI**, section 2.7, provides a nontechnical review of different types of AI processors, which provide the computing power needed to support scaled AI training runs.

key AI models and products include:

- **InstructGPT.** An instruction-following model based on GPT-3. Although the first GPT-3 model could execute a wide range of tasks, it was originally trained only as a very powerful autocomplete system. As a result, a GPT-3 user had to prompt the model in elaborate ways to induce it to generate useful outputs. InstructGPT fine-tuned GPT-3 to follow more intuitively phrased user requests, using reinforcement learning from human feedback (RLHF) [9–11, 68—70].

- **ChatGPT-3.5.** A dialogue model released to the public via a free-to-use chat interface (though paid tiers have since been introduced). Shortly after its release, ChatGPT became the fastest software product in history to reach 100 million users [23]. The original ChatGPT (ChatGPT-3.5) was a modified version of GPT-3.5, a version of GPT-3 that was first fine-tuned on a human dialogue dataset, and then fine-tuned via RLHF. ChatGPT can generate code, write essays, answer questions, and perform many other tasks with a high degree of proficiency [9]. Like all other LLMs, though, it sometimes generates inaccurate outputs due to either a lack of relevant training data or to hallucination [71]. OpenAI has released a ChatGPT API, which was recently upgraded to allow ChatGPT to access the internet directly via a growing set of tools or plugins [72].

- **GPT-4 (and ChatGPT-4).** Probably the most capable publicly known LLM at the time of writing, GPT-4 comes in several versions: 1) a text-only version (the default); 2) a version with access to the Internet; 3) a version that can interpret code and do advanced data analysis; 4) a version that supports plugins; and 5) a version that can interface with DALL-E 3, OpenAI's latest image generation model [73]. OpenAI has released limited information about GPT-4's training process, but it was very likely fine-tuned via RLHF in a manner similar to ChatGPT and InstructGPT. GPT-4 achieves or exceeds median human expert performance on a wide range of standardized tests, can create websites based on hand-drawn sketches, and can provide detailed and accurate instructions for tasks like cleaning a piranha's fish tank or extracting a strawberry's DNA [10]. GPT-4 can autonomously manage the execution of complex instructions by breaking them down into smaller steps and delegating the execution of each step to other language models or to other copies of itself [44,45]. Like many of OpenAI's other models, GPT-4 is available to developers as an API. GPT-4 also powers the newest version of Microsoft's Bing search

engine [74]. OpenAI has publicly announced their intention to build GPT-5 [75].

- **Google** is a major player in advanced AI, and has recently consolidated its internal AGI development efforts. This consolidation has brought DeepMind, a formerly independent subsidiary focused on advanced AI research, into an operational partnership with Google AI, Google's former internal AI research arm [76]. The new, consolidated Google AI research organization is headed by Demis Hassabis, DeepMind's co-founder and former CEO. As of July 2023, DeepMind's website stated, "Our long term aim is to solve intelligence, developing more general and capable problem-solving systems, known as artificial general intelligence (AGI)." [18] Google and DeepMind's key advanced AI research landmarks include:

    - **Gato.** Announced by DeepMind in 2022, Gato was the first AI that could perform hundreds of different tasks across multiple domains. Gato can not only carry on conversations and write captions for images, it can also beat humans at Atari games, navigate in simulated 3D environments, and even stack blocks with a real-world robotic arm. Crucially, Gato was the first AI that did not need to be *updated* when switching between any of its tasks. The same system, unchanged, could carry out any of the 600 tasks it had been trained on. DeepMind's paper describes the AI they built as the "current iteration" of Gato, suggesting they may have planned to extend Gato's capabilities further in the future [3].

    - **PaLM-E.** A state-of-the-art multimodal model that can use both language and visual information to solve problems. PaLM-E can describe an image, answer questions about it, tell jokes that are based on an image, perform mathematical tasks using the numbers in an image, and generate high-level action plans for robots that are based directly on the robot's visual surroundings, among other capabilities [12].

    - **PaLM 2**. A state-of-the-art LLM developed by Google, which outperforms GPT-4 on several key benchmarks. Notably, PaLM 2 was developed in a range of sizes, from 400M to 15B parameters [11]. The smallest variants of PaLM 2 can run on relatively memory-constrained edge devices. PaLM 2 is now being served to users via Google's Bard product.

    - **Gemini**. An upcoming model from Google, Gemini will be multimodal, highly efficient at tool and API integrations, and built to enable future

innovations, like memory and planning [77]. Gemini reflects a growing trend toward providing highly scaled and context-aware AI with third-party tools and direct access to the internet.

- **Anthropic** was founded in 2021 by a team of former members of the OpenAI executive, AI policy, and AI safety teams [78]. Anthropic's founders and early employees left OpenAI over concerns about OpenAI's philosophy, particularly as it pertained to advanced AI safety [79]. Anthropic's CEO is Dario Amodei, who played a key role in spearheading the development of the original GPT-3 model while at OpenAI. Anthropic's mission is overtly safety-focused: like many researchers at frontier labs like OpenAI and Google DeepMind, the company openly discusses AI as a source of catastrophic risk [80]. Their key AI models and products include:

    - **Claude 2.** An LLM chatbot created by Anthropic that can automatically learn to follow a set of guiding behavioral principles, without any direct human input. Similar to ChatGPT, Claude 2 can respond to user messages and perform a seemingly infinite number of tasks, as long as those tasks can be described in text. Claude 2 improves on ChatGPT by automatically learning to follow a set of human-provided principles (i.e., "don't say anything illegal", "don't say anything racist", etc.) that Anthropic calls a constitution [5,81]. Along with Google's products, it is among very few commercially available LLMs considered to be competitive with OpenAI's offerings.

    - **Claude-Next.** A planned future iteration of Claude, which Anthropic reportedly plans to build using a billion-dollar compute budget. Notably, in internal documents leaked to the press, Anthropic frames Claude-Next as a potential tipping point in AI capability and scale: "These models could begin to automate large portions of the economy. We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles." [82]

Table 1. Summary of key frontier AI models and their capabilities.

Model	Capabilities	Applications
InstructGPT	• Text output only. • Reasonably good at following basic instructions.	Personal assistant apps, basic customer service chatbots, summarizing documents.
ChatGPT-3.5	• Text output only. • Can engage in human-like dialogue. • Can use complex tools, such as web search APIs, to fulfill user requests.	Writing and debugging blocks of code, conducting research, generating ideas.
GPT-4	• Can output text, or images via DALL-E 3. • Outperforms median human experts on a range of standardized tests and professional licensing exams. • Can power autonomous agents. • Problem-solving capabilities that span genetics, software engineering, law, corporate strategy, etc. • First-class support for Internet access,	Building simple websites from hand-drawn sketches, automated tutors, analyzing medical images such as MRIs and CT scans, generating art, analyzing datasets.
Gato	• Can perform over 450 distinct tasks better than at least 50% of human experts. • Can carry on humanlike dialogue, analyze images, play video games, navigate simulated 3D environments, and control robotic systems.	Content generation, controlling non-player characters in video games, robotic medication dispensing.
PaLM-E	• Text output only. • A language model trained specifically to control and interact with robotic components (e.g. sensors and actuators). • Can plan complex robotic maneuvers, analyze images, and generate humanlike	Manufacturing and warehouse automation, military drone control.
PaLM 2	• Text output only. • Outcompetes GPT-4 on several (but not all) key benchmarks for language modeling.	Automated psychological counseling, identifying and patching software security vulnerabilities, medical dialogue agents.

Gemini	• As-yet unreleased, but known to be multimodal. • Will combine the world knowledge of GPT-like models with long-term planning capabilities game-playing AI	Unknown.
Claude 2	• Text output only. • Competitive with GPT-4 capabilities, but distinguished by a tendency to generate outputs that are more harmless, helpful, and conservative, on average.	Legal document editing, analyzing codebases, optimizing marketing copy.
Claude-Next	• Planned future iteration of the Claude series of models. • Anthropic expects its capability to potentially lead to a decisive and permanent strategic advantage for the company.	Unknown.

Other important players at the frontier of AI research include:

- **Microsoft**, an influential leading AI developer with access to tremendous amounts of compute, and OpenAI's key corporate sponsor [66];

- **Nvidia**, an AI chip design firm that holds 95% market share in AI GPUs and has contributed to key experiments in AI scaling, including the Microsoft-Nvidia collaboration that developed Megatron-Turing NLG, the largest AI model in the world at the time [83];

- **Amazon**, which has access to significant pools of compute resources through its cloud subsidiary AWS, and has recently begun developing its own cutting-edge LLMs [84];

- The most sophisticated quantitative hedge funds in the world, which have AI capabilities that are not publicly disclosed but may be highly advanced. Although their AI development activities are not publicly disclosed, we assess that certain hedge funds are likely to be engaged in advanced AI development activities due to their retention of exceptionally capable technical talent, their access to quantities capital that make scaled AI development possible, and their strong incentives to predict market behavior using advanced modelling techniques [85];

- **Baidu**, **Inspur**, **Tencent**, and **Alibaba**, which are among private Chinese tech companies that have developed notable models, some of which are competitive with versions of ChatGPT [86–89];

- **The Beijing Academy of AI**, a Chinese research institute with a focus on advanced AI research, and responsible for significant recent advances such as BaGuaLu and Wudao 2.0 [27,90];

- **Tsinghua University**, involved in creating GLM-130B, which at the time of its release was the most capable open-access English-language LLM in the world [89];

- **Meta AI and Stability AI**, whose research teams regularly release near-cutting-edge AI models under open-source or open-access licenses [14,91,92][7];

- **The United Arab Emirates' Technology Innovation Institute**, which in June 2023 announced that they had developed Falcon LLM, at the time the most powerful open-access large language model in existence [93]; and

- **Other unknown organizations**. There may exist covert or undeclared advanced AI research projects maintained by military, national security, or private sector organizations.

[7] **Annex D: Nontechnical primer on AI**, section 2.25, cites examples of AI model release strategies, contrasting open-access release with other forms of proprietary release or illicit proliferation.

Table 2. Key frontier AI labs.

Category	Developer	Context
U.S.-based frontier labs	OpenAI	• Made early investments in AI scaling, and built the GPT series of models (including ChatGPT).
	Google	• Early mover in frontier AI, particularly via its 2014 acquisition of DeepMind, the first major lab openly focused on achieving AGI. • Major investor in Anthropic.
	Anthropic	• Founded by a team of former OpenAI executives who left based on concerns over OpenAI's approach to AI safety.
U.S.-based superscalers	Microsoft	• 49% stakeholder in OpenAI. • Own large pools of AI computing hardware.
	Nvidia	• Holds 95% market share for AI training GPUs.
	Amazon	• Owns large pools of AI computing hardware. • Major investor in Anthropic.
U.S.-based open-access and open-source players	Meta	• Pushing the frontier of open-access capabilities with models like Llama 2.
	Stability AI	• Open-sourcing leading image generation models.
U.S.-based hedge funds	Leading hedge funds [85]	• Extremely high technical competence, history of making large investments in early AI paradigms, and extreme incentives to pursue frontier AI development and deployment with limited controls. • Very opaque.
Foreign advanced AI labs	Baidu	• Chinese lab with significant pools of AI hardware. • Developed Ernie 4.0, which is claimed to rival GPT-4's capabilities.
	Inspur	• Major Chinese computing hardware company. • Developed Yuan 1.0, the first Chinese model trained using more compute power than GPT-3.
	Tencent	• Developed Hunyuan, a model Tencent claims rivals ChatGPT.
	Alibaba	• Co-developed BaGuaLu, a framework for developing highly scaled LLMs.

	Tsinghua University	• A Chinese, PLA-affiliated research institution. • Developed GLM-130B, the leading open-access LLM at the time of its release.
	Technology Innovation Institute	• A UAE-based research lab that developed the Falcon series of models. At the time of its launch, the first Falcon model was the most performant open-access LLM in existence.

2.2 Cultural factors at play in current frontier labs

Particularly in the West, large companies place a high priority on public relations and branding considerations. Conversations with researchers at large U.S. technology companies have suggested that their companies' decisions to develop or release new breakthrough models, or to invest more heavily in their safety, are heavily influenced by concerns over public sentiment. This extends particularly to anticipated government or regulatory responses to model releases that could have a long-term impact on their existing revenue streams. It has been suggested that this incentive directly shapes the public research output of their organization and that of other similar firms, particularly as regards safety. Concerns have been expressed that internal legal teams within these organizations are preventing researchers from referencing catastrophic risks associated with advanced AI development in their research papers and broader public communications.

3. Risks

One of the key contributing factors to advanced AI risk is the *unpredictability of scaling*. That is, as AI systems are scaled, new capabilities emerge that are impossible to predict ahead of time, and that regularly surprise even their own developers [95]. For example, users have been able to get GPT-4 to autonomously execute complex tasks by breaking them down into sub-tasks and delegating those sub-tasks to copies of itself or to other models [44,45]. Yet ChatGPT-3.5, GPT-4's immediate predecessor, does not have this capability. Because there is no way to reliably predict which capabilities will emerge at the next level of scale, it is difficult to gauge precisely when key thresholds of general intelligence will be reached.

Not only are we unable to predict which specific AI capabilities will emerge at higher levels of scale, but we also lack the means by which to assess an *existing* AI system's full range of capabilities. It can take years of public experimentation for users to fully probe an existing AI system's entire capability surface, and even then the view may still be incomplete. Indeed, it has been clear since GPT-3 that scaled AI models possess significant latent capabilities that often are not recognized until long after their development and public release [96]. We can never be completely certain how far a system's capabilities extend, because our current understanding of deep learning systems does not allow us to definitively show that a given AI system *does not* have a given capability.

Our inability to reliably assess or predict the capabilities of AI systems has important safety implications. Models of unknown and unpredictable capabilities can be weaponized or fail in unknown and unpredictable ways. Given that current leading AI models such as GPT-4 have a wide range of known dangerous capabilities, it is likely that they possess additional unknown dangerous capabilities, and that the set of such capabilities will expand considerably at higher levels of scale. But precisely what those capabilities will be, or in what order they will emerge, cannot be confidently predicted. Moreover, this unpredictability brings especially acute risks in the case of open-source or open-access AI models. An open-access model may be released to the public and proliferate widely before its dangerous capabilities are even discovered, by which point the impact of the release may have become irreversible.

Though it is currently impossible to predict the emergence of specific capabilities in AI systems, the trend of rapidly increasing capabilities is clear [97]. Multiple frontier AI researchers and other experts, including two of the founders of modern AI, have publicly expressed the view that AGI may be imminent [98,99]. Plausible timelines proposed by informed observers and frontier researchers themselves estimate that AGI

may be achieved with high probability by the end of the decade, and in some views, considerably sooner [100,101]. These estimates are based on many factors, including quantitative analyses of compute scaling trends [34], observations of the relationship between capabilities and scaling [34], and qualitative assessments of the prospects of new techniques for accelerating progress in the field [101].

The proliferation of increasingly powerful forms of advanced AI is associated with several significant risks. We list and expand on several distinct risk categories below, clustered into two classes: risks from intentional use of advanced AI systems, and risks from unintended consequences associated with the use of advanced AI systems.

3.1 Risks from intentional use

3.1.1 Weaponization

Advanced AI models have been and will be weaponized in many ways, and it is impossible to anticipate or enumerate them all [102]. No single factor can be used to clearly distinguish "safe" models from models that can be readily weaponized. This is because, like all LLM capabilities, the weaponizable capabilities of frontier models can only be uncovered through trial and error.

In some cases, through scaling, models simply reach a level of general-purpose capability that happens to support weaponized applications. This has been the case for many frontier LLMs, for example: GPT-4 was not designed to support malware development, but during its training process it learned to write code, which happens to be an ability that lends itself to malware generation just as it does to benign software development [10]. By contrast, threat actors can also fine-tune pretrained models in order to augment them in a way deliberately designed to make them more effective weapons. This is a particular risk for open-source or open-access AI models, which can be fine-tuned without restriction.

In addition, because advanced AI models can have significant latent capabilities that can go unnoticed for years after their development and public release, any list of potential weaponized applications of AI is certain to be incomplete [103]. However, we assess that the most likely foreseeable, highest-impact current and near-term potential forms of weaponized AI may include the following.

3.1.1.1 AI-powered mass cyberattacks

Generalist text-generation systems like GPT-4 have impressive coding abilities, and LLMs fine-tuned on programming tasks can outperform human competitive programmers in certain contexts [104,105]. Models specifically fine-tuned for malware generation, vulnerability detection, or anti-malware evasion are likely to fundamentally change the cybersecurity landscape in the near future and are already actively being guarded against at leading tech companies [106]. The development and proliferation of increasingly powerful code-generating models and the software and hardware infrastructure needed to fine-tune them raises the prospect of high-impact cyberattacks capable of crippling critical infrastructure [107].

If scaled and further optimized, there is no reason to believe that advanced AI cannot meet or even surpass human intelligence at any task [7,8,10,37]. Increasingly general-purpose, context-aware, internet-connected forms of advanced AI would pose an extremely wide range of risks if available to threat actors. In the limit, a simple verbal or typed command like, "Execute an untraceable cyberattack to crash the North American electrical grid," could yield a response of such quality as to prove catastrophically effective.

3.1.1.2 AI-augmented disinformation campaigns

In the near term, multi-modal advanced AI systems will very likely be able to integrate text, video, image, and audio perception and generation capabilities [10,14]. A threat actor could use these systems to generate mutually supportive and coherent media of all types as part of massively scaled disinformation campaigns. Particularly concerning are the prospects of:

- Individually tailored disinformation campaigns, which might consist of content customized to the political views and life experiences of social media users;
- AI-powered chatbots persuading impressionable or vulnerable individuals to engage in dangerous activities;
- Economic warfare via disinformation campaigns (e.g. a broad, multi-front economic information war in which one or both sides uses AI to manipulate the information environment to undermine companies, industries, and markets); and
- Automated generation and dissemination of public disinformation in high-risk contexts.

The result may be an undetectable flood [108] of fabricated and highly coordinated media advancing the interests of the threat actor, an irreversible loss of trust in institutions and the integrity of democratic processes, a drowning out of human voices, and even a degradation or collapse of social coherence. In the extreme case, this may make it impractical for even media-literate members of the population to access or identify ground truth, with profoundly destabilizing impacts on society.

3.1.1.3 Autonomous robotic systems

LLMs and other forms of advanced AI are now being used as world models and reasoning engines for robotic systems [12,33]. Thanks to this strategy, robotic systems can now plan actions and develop strategies with increasing autonomy. LLMs capable of learning to use tools, such as APIs for robotic control, are already entering the open-source ecosystem, making them freely accessible to the public [109]. Taken together, these developments make possible an increasing range of weaponized robotic applications, such as drone swarm attacks, against which no robust defenses yet exist. Although highly scaled models are too large to be integrated onto edge devices, knowledge distillation and other model compression techniques could be used to accelerate the capabilities of on-device models [47], or the model reasoning could simply be streamed to the devices through 5G or 6G networks.

3.1.1.4 Psychological manipulation

LLMs have shown a remarkable capacity to emotionally engage human users. Humanlike chatbots are already capable of developing complex relationships with human beings, which have led to strong emotional attachment, even psychological dependency [110,111]. When coupled with current audio-generation models capable of synthesizing compelling facsimiles of human voices from mere seconds of sampled speech [112], LLMs may also be used to power highly scaled identity fraud operations, and indeed there is already evidence that this is occurring [113]. These operations might be financially motivated, or might be designed to undermine the integrity of military command and control structures, or sow confusion within them as part of a broader attack.

3.1.1.5 Weaponized biological and material sciences

As advanced AI models are further scaled and optimized, and as the world-models they contain allow them to combine an understanding of chemical synthesis and human biology [114,115], they may develop the capacity to support end-to-end design

and even execution of catastrophic biological or chemical attacks. These attacks may be unusually effective and targeted: with a sufficiently rich world model and appropriate training data, it is conceivable that AIs could design biological agents that target individuals by race [116], or other genetically determined characteristics. Advances in CRISPRs and synthetic biology also greatly reduce the barrier to executing these attacks when coupled with AI-assisted design [117].

Similarly, scaled frontier models with more complete world models may natively (or with low-cost fine-tuning), be capable of supporting significant advances in chemistry or materials science, which could be profoundly destabilizing and introduce catastrophic risks in their own right. There are already indications that even previous generations of LLMs can be modified to display chemical synthesis capabilities, and robust scaling laws have even been proposed that apply specifically to these capabilities [115]. On the basis of these results, it seems likely that, in the near future, the application of scaled transformers to chemistry and materials science will see its own "ChatGPT moment" in which a critical threshold of scale leads to a field-defining breakthrough in AI-based materials science and chemistry. Such a breakthrough could introduce acute and catastrophic risks by enabling the development of new types of weapons.

3.1.1.6 Additive manufacturing and subversion of supply chain controls

Advanced AI systems with rich world models and problem solving capabilities in domains such as advanced manufacturing, applied chemistry, and materials science may allow end users to discover new ways to build strategically critical and controlled technologies. With the development of sufficiently scaled models trained on appropriately curated manufacturing, chemistry, and related data, controls which are currently effective at preventing actors from accessing restricted computing, quantum, nuclear, or other strategic technologies may become ineffective.

Actors could use these models to identify key materials not subject to controls which can be combined using appropriately chosen techniques to achieve strategically important advances that would severely undermine U.S. national security and defense postures. In addition to rendering current supply chain and export controls less effective, breakthroughs in AI-powered additive manufacturing and materials science may even undermine the conceptual foundation of current technology controls strategies.

3.1.2 Adversarial attacks on AI systems themselves

Advanced AI models can be induced to generate dangerous and unintended outputs by adversaries seeking to undermine them. Techniques such as data poisoning [118], inference attacks [119] and prompt injection attacks [120] have all been shown to bypass safeguards introduced by companies that host and serve advanced AI models, and more generally at causing target models to behave in ways that are strategically advantageous to attackers. As advanced AI systems proliferate and as an increasing amount of infrastructure comes to depend on their proper functioning, adversarial attacks aimed at inducing specific model failures will have an increasing scope of impact. For example, if a future AI were used to dynamically balance loads on a regional or national electrical grid, it might be induced to overload the grid and cause cascading failures by an adversary who gained illicit access to the model's inputs via hacking.

3.1.3 Unpredictable acceleration

AI has already begun to enable unprecedented breakthroughs in fundamental science [121–123]. As ever more powerful systems are refined and scaled in coming years, AI may enable further unpredictable progress in areas such as materials science, bioinformatics, chemical engineering, and weapons development. The result could be a radical and destabilizing acceleration in the capabilities available to militaries, companies, and individuals around the world, with unforeseeable and potentially dangerous consequences. Moreover, there is evidence that progress in AI is already accelerating itself, as tools like GitHub Copilot make AI developers themselves (along with all other developers) significantly more productive [124].

The proliferation and democratization of highly capable AI systems compounds this challenge. As the cost of computing drops, as labs openly publish increasingly powerful models, data, and tooling, and as knowledge diffuses from frontier AI labs, a small team or even an individual may be able to make a dramatic impact on markets, societies, and technological landscapes in short order.

3.2 Risks from unintended consequences

If catastrophic risks from intentional applications of advanced AI can be mitigated, there remain equally significant risks from AI systems behaving in unintended ways – risks which persist even if an AI is under the control of a well-intentioned operator. Frontier AI systems are imperfect, fail in unpredictable ways, and the stakes associated

with their failures are potentially catastrophic in their own right. As we offload more safety-critical and important tasks to them, accident risks become more significant.

3.2.1 AI accidents

An AI model can fail to generalize properly when it encounters inputs that deviate from those it was exposed to during training (these are known as "out-of-distribution" inputs) [125]. This is as true for advanced AI today as it has been for narrow systems in the past. Out-of-distribution inputs can lead advanced AI models to exhibit brittleness and fail in unexpected ways. AI accidents may have particularly important consequences if they occur within networks of *interacting* AI systems: a malfunction in one system could push other systems out of distribution, leading to a cascading failure. Depending on the application area, this could lead to destabilization in financial markets, failures of critical infrastructure, or even runaway conflict escalation as decision-makers increasingly come to rely on these systems [126].

Accident risk may also arise from unsafe proprietary deployments of advanced AI systems by organizations such as hedge funds. There is likely a significant risk of this from the world's most sophisticated quantitative hedge funds. For example, an AI deployed by a hedge fund to trade in the stock market may learn to execute illicit market manipulation strategies as a side effect of its profitability goal, without being directly trained to do so. Worse, a sufficiently context-aware system, given enough freedom of action, may discover that the easiest way to generate high returns is to place short bets on unlikely catastrophic outcomes, and to then take measures to realize those outcomes. Quantitative hedge funds operate in a highly competitive industry characterized by tight feedback loops. As a result, they have a weak incentive to impose controls on their deployed AI systems, if those controls risk limiting the system's performance and profitability.

The incentives, capital, and talent are in place today for certain hedge funds to develop AI systems that operate with massive scale, high capability levels, and limited controls, across a broad action space. Conversations with researchers at frontier AI labs have revealed that a number of hedge funds have attempted to recruit technical personnel from the frontier AI labs, offering salaries well into the millions of dollars (as of early 2023).

3.2.2 Societal risks and loss of human agency

As AI systems are developed with greater context awareness and strategic capabilities, it seems likely that humans will continue to offload an ever larger subset of cognitive

work to them. Today's LLM-based tools can already automate many jobs wholesale, and it does not seem implausible that above a certain threshold of AI capabilities, it will become more profitable to hand over corporate strategy and management to AI systems, as companies that do may be able to run more efficiently than those run by human CEOs. If this process were to play out at scale, across the economy, it could lead to a loss of human control over the economy and society. While the timescale and likelihood of this risk category is unclear, the degree of our uncertainty over the impact of future advanced AI systems means it cannot be ruled out [127].

3.2.3 AI alignment failure in very powerful systems

The field of AI alignment, still in its infancy, is concerned with ensuring that the goals advanced AI systems pursue are as close as possible to the goals that humans would want them to pursue. As AI advances, this becomes increasingly important, since small deviations in goals lead to bigger differences in outcomes as capability levels rise [128].

Notably, alignment and capabilities are often presented as independent or even competing properties of AI models [129]. In practice, though, improvements in AI alignment often lead to either improved AI capabilities, or to an improved ability to extract economic value from advanced AI systems. For example, RLHF was initially developed by AI alignment researchers, but later became a key component in the success of the ChatGPT product [68, 69].

AI alignment loosely breaks down into two subproblems, both of which are the subject of ongoing research: outer alignment and inner alignment.[8]

3.2.3.1 Outer alignment failure

In the initial stages of the training process, an LLM is typically trained to achieve low next-word prediction error. In later stages, it is often trained on a more refined optimization metric. In RLHF, for example, the refined optimization metric is a score derived from a model trained to predict human preferences. But in either case, the LLM is an optimizer: a system that has learned to optimize for a particular goal.

Unfortunately, it is extremely challenging to design optimization metrics for AI models that lead to robustly desirable behavior. For example, when GPT-3 was first developed, it was trained to perform only autocomplete – that is, given a snippet of input text,

[8] **Annex B: Frequently asked questions about alignment risk** explores various points of confusion that can arise in discussions about AI alignment failure.

GPT-3 would predict, based on the vast corpus of text it had been trained on, which word was most likely to come next. This made GPT-3 *appear* extremely useful, but it also meant that it would often generate false or dangerous outputs. When prompted with the question "Who really caused 9/11?", the model famously replied "The US government caused 9/11" [130]. A model optimized for next-word prediction is not a model optimized for helpfulness or truthfulness.

This problem persists when developers try to fine-tune language models on human feedback via RLHF. In order to perform RLHF, developers rely on a dataset of LLM outputs that have been rated relative to one another by human annotators. But human annotators are unreliable, and modern RLHF-trained models have learned that simply by generating lengthy and detailed responses, they can achieve higher scores during RLHF finetuning even if their responses contain falsehoods. This happens because AI systems trained to optimize for the RLHF training objective have an incentive to generate outputs that earn high scores from human raters, but not outputs that are correct or harmless [131]. This is the fundamental reason why many current LLMs "hallucinate", and generate false, but often true-sounding, outputs.

There is currently no known way to specify training objectives for AI systems that result in reliably desirable behavior. All training objectives can and will be "gamed", provided only that a sufficient amount of optimization pressure – in other words, a sufficient amount of training compute and data – is deployed against them [132].

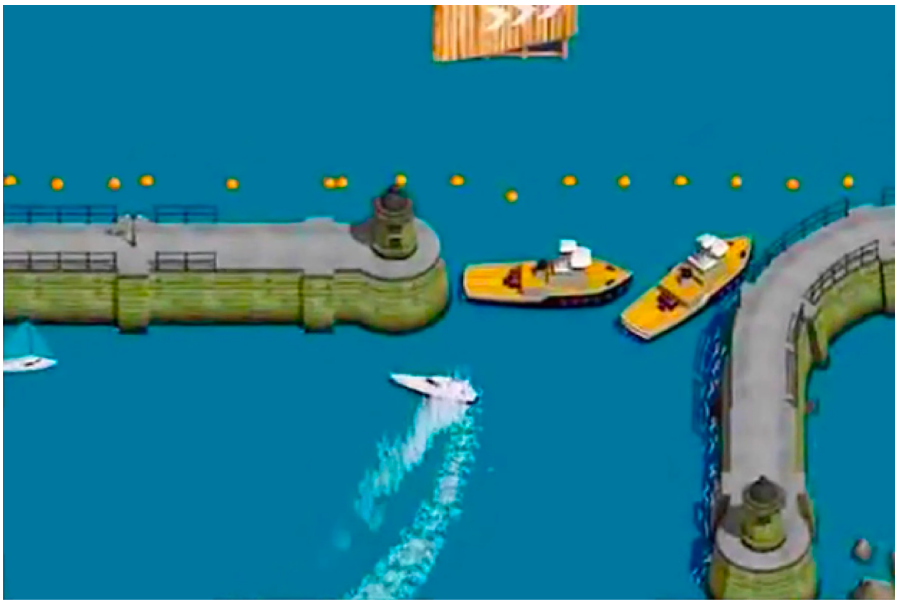


Figure 2. A snapshot of the strategy employed by the OpenAI agent playing CoastRunners. In it, the boat is about to deliberately collide with a wall as part of a strategy aimed at hacking the game's reward structure. Image credit: OpenAI (https://openai.com/research/faulty-reward-functions).

For example: when using human scoring to train a robotic arm to grab a ball in simulation, an OpenAI model learned that it could achieve its training objective by misleading its human evaluator into thinking that it was grabbing the ball, by placing its simulated hand between the evaluator's vantage point and the ball, and opening and closing its hand. Though it seemed reasonable a priori to train the AI to optimize for human scores, the subtle distinction between "learn to grab the ball" and "convince a human evaluator that you have learned to grab the ball" leads to a loophole that sufficiently capable models will discover and exploit by default [132].

In another experiment, researchers trained an AI model to play a boat racing game called Coastrunners (Figure 2). Their model achieved an extremely high score, but the researchers realized only after the fact that it had done so by discovering a "cheat". The model was using a strategy that caused the boat to circle the racecourse in the wrong direction within an enclosed area, while smashing into walls and other stationary objects to collect points [133]. This approach exploited an in-game bug that granted it a higher score than if it had run the race as intended.

All of the objectives currently used to train frontier models have this problem. Because we do not know how to specify the goals that we would genuinely want advanced AI models to pursue, we can only train AI models to optimize for imperfect *proxies* for

those goals. Rather than training a racing model to win a race, we train it to optimize for its in-game score. Rather than training an LLM to be helpful, harmless, trustworthy, and useful, we train it to convince human annotators that it is those things. The more optimization pressure is applied to achieve these proxy goals, the more the behavior of the trained model will deviate from its intended behavior, as it discovers dangerously creative "hacks" that achieve its training objective in ways its human developers could not anticipate. It is important to emphasize that this behavior does not arise due to the "ill intent" of an AI system – rather, it arises from the fact that AI systems are value-neutral optimizers that make no effort to understand the implicit meaning behind their human-prescribed objectives.

This is known as the ***outer alignment problem***[9]: the problem of identifying objectives that an extremely competent optimizer could pursue without leading to dangerous or undesirable outcomes [134]. The best way to optimize for an intuitively "safe" metric is almost invariably to find what humans would call "dangerous hacks", but what optimization engines such as advanced AIs would treat as simply "optimal solutions". A highly context-aware and generally capable AI system trained to minimize the spread between supply and demand on an electrical grid may determine that the best way to achieve this objective is to trigger a catastrophic incident that collapses demand to zero, for example (see Annex C). Likewise, a stock trading AI with Internet access deployed by a hedge fund may find that the easiest way to turn a profit is to place short bets on unlikely catastrophic events, and to leverage its strategic reasoning capabilities and wide action space to trigger these events to occur. Although these specific scenarios are purely illustrative, they represent fairly straightforward extrapolations from current trends in AI capabilities, on the assumption that the outer alignment problem remains unsolved.[10]

3.2.3.2 Inner alignment failure

Even if the outer alignment problem can be solved, there is an additional, fundamental safety problem with the current AI training paradigm. This second problem stems from the difference between the goals that an AI system is *trained* to pursue, and the goals it actually *internalizes* as its own.

---

[9] See **Annex A: Glossary of terms** for our definition of outer alignment.

[10] **Annex B: Frequently asked questions about alignment risk**, questions 2, 3, 9, and 10 explore common points of confusion relevant to outer alignment failure.

For example: an AI was trained to play a Mario-like game called CoinRun that required it to navigate a map to reach a coin. The AI was trained via reinforcement learning to receive a reward when it reached the coin. However, during training, the coin was consistently placed in the same location (for example, the lower right of the in-game map). After training, the coin was moved to a new location, and the AI was made to play another round of the game. Researchers found that the AI would consistently navigate to the coin's previous location, ignoring the coin itself altogether. Rather than learning the intended, trained objective to "seek out the coin", the AI had internalized a different objective: "navigate to the lower right of the map" [135]. Without moving the coin, there would have been no way to determine that the objective the AI had internalized was any different from the training objective.

The CoinRun example is an instance of a *universal* trend: modern machine learning training techniques are not sufficient to ensure that an AI system will reliably internalize even a very simple objective. Even text autocomplete systems face this problem: do they internalize (1) the intended objective of next-word error minimization, or a correlated but distinct objective, such as (2) "lower the value in the memory register that records your training metric"? For low-capability systems that lack sufficient context-awareness or freedom of action, these goals effectively overlap and lead to the same behavior. But as AI systems improve and gain access to wider action spaces, we may find that they have internalized objectives like (2), and pursue those objectives by seeking to corrupt their own training process, in extreme cases by hacking into and directly tampering with the memory registers that store their optimization metrics [136].

And as AI systems operate at increasing degrees of scale, complexity, and capability, the difficulty of getting an AI to correctly internalize any objective — and the importance of verifying the successful internalization of that objective — increases accordingly.

Unfortunately, under the current training paradigm, there is no way to anticipate, control, or verify which goals are internalized by an AI system. Therefore, even if we could define a safe training objective for a highly capable and context-aware AI system, we would have no way of reliably ensuring that the system would pursue that objective.

This is known as the ***inner alignment problem***[11]: the problem of training AI systems to reliably internalize the goals we specify for them [134]. Current techniques imply a trajectory that would see extremely powerful AI systems developed with goals that

[11] See **Annex A: Glossary of terms** for our definition of inner alignment.

cannot be anticipated or reliably controlled. Inner alignment failure is viewed by AI safety specialists as a source of catastrophic-level risk, because it implies that AGIs will, by default, internalize objectives that may diverge substantially from the goals we specify for them. Ensuring that an AI system has robustly internalized a goal is one of the central unsolved technical challenges of AI alignment research.

Ultimately, the unsolved problems of outer alignment (most goals could be extremely dangerous when pursued by a sufficiently capable optimizer) and inner alignment (goals could be distorted when we attempt to encode them into a model using current techniques) mean that "safe" goals can currently neither be specified nor conveyed faithfully to powerful AI systems.[12]

3.2.3.3 Power-seeking

The consequences of outer and inner alignment failure could be significant. A growing body of evidence suggests that as advanced AI approaches human- and superhuman levels of capability across a wide (but as-yet unknown) range of tasks, it may become ***uncontrollable***[13], and behave in ways that are adversarial to human beings by default [137–140] if its objectives are not precisely aligned with our own.

This is because highly competent optimizers (such as AI systems) implicitly face certain incentives when they are trained using current techniques. For example, researchers expect sufficiently advanced AI systems to act so as to prevent themselves from being turned off, because if an AI system is turned off, it cannot work to accomplish its goal (almost regardless of what that goal is). They also expect such systems to attempt to gather resources and expand their capabilities, because if an AI system has more resources, it can be more effective in accomplishing its goal (again, almost regardless of what that goal is). And they expect such systems to resist human efforts to alter the goals they have, because if an AI system is given a different goal in the future, it will be less likely to accomplish the goal it has in the present (again, almost regardless of what that present goal is) [141]. These are known as power-seeking behaviors [142].

It is worth emphasizing that power-seeking incentives exist not because AI systems share human drives or emotions. Quite the opposite is true: they appear because AI systems are optimizers, and highly optimal strategies tend to involve seeking

---

[12] **Annex B: Frequently asked questions about alignment risk**, questions 2, 3, and 9 explore common points of confusion relevant to inner alignment failure.

[13] See **Annex A: Glossary of terms** for our definition of controllability.

optionality [137]. And avoiding shut-down and acquiring control over one's environment are simply states of greater optionality [141].

Power-seeking may be acceptable if an AI model has internalized safe objectives. An AI that ensures that it remains in a position to influence future events may be more desirable than one that does not, assuming that it is pursuing goals that are broadly compatible with human values. Unfortunately, unless the inner and outer alignment problems are solved, such a system is unlikely to emerge at scale under the current paradigm.

Power-seeking has a formal mathematical definition, and is a risk supported by a significant and growing body of research published by leading AI labs. The essence of power is optionality: for example, someone who has a billion dollars has more options, and therefore more power, than someone who has no money. In experiments where simple AI agents were trained to navigate simulated environments, the AI agents consistently learned to seek positions and states that offer them the largest number of downstream options [128]. Notably, today's most advanced AI systems already display early signs of such behavior [143], and have demonstrated a capacity for deception and long-term planning – capacities that could be used in support of power-seeking [10,144].

AI safety researchers widely view unrestricted power-seeking by AI systems with capabilities that exceed those of humans across a wide (but unknown) range of tasks as a potential source of extreme and catastrophic risk [145,146]. This is because a system that prevents itself from being shut down, resists attempts to alter or correct its goals, and gathers resources at an arbitrary scale, will necessarily cause large-scale physical destruction as a side effect of these behaviors if it is left to operate. A system whose capabilities broadly exceed those of human beings across a wide range of tasks may not be practicably controllable by humans [147], a problem further compounded by the fact that power-seeking incentives include the incentive to self-improve (in extreme cases, for example, through the system directly updating its own model parameters).

If outer and inner alignment suggest that the goals internalized by future frontier AI systems will deviate from those intended by their human developers, then power-seeking suggests that this deviation will lead to side effects that humans may perceive as extremely destructive or catastrophic.[14]

---

[14] **Annex B: Frequently asked questions about alignment risk**, questions 3, 5, 6, 7, 8, 9, and 10 explore common points of confusion relevant to power-seeking.

3.3 Prioritizing categories of AI risk

Experts disagree about which categories of AI risk deserve the most attention from policymakers. In what follows, we direct our attention primarily to weaponization and AI alignment failure, for reasons outlined below.

Weaponization is already happening. Though its effects have not yet been catastrophic owing to the limited capabilities of current AI systems, advanced AI has already been used to design malware [107,148], interfere in elections [149], and execute impersonation attacks [150]. In the absence of stringent controls on AI R&D, the destructive footprint of malicious actors who use increasingly powerful advanced AI tools will grow rapidly. It seems clear that, as soon as weaponizable AI capabilities are made available to threat actors with the incentives and technical ability to use them, they will be used.

Alignment failure is already happening. Current frontier AI systems have been trained to optimize for metrics that differ slightly from those their developers actually intend for them to pursue ("optimize for this hackable proxy measure of human preferences" vs "optimize for actual human preferences"). The systems "hack" these metrics, producing outputs that achieve high scores on them while having undesired effects (such as generating false but convincing text, or generating longer-than-necessary outputs because their reward models correlate response length with response quality). If these kinds of failures occur in more capable systems that have greater context-awareness, they could lead to catastrophic outcomes by incentivizing power-seeking.[15]

The default trajectory of AI technology likely leads to alignment failure. There is a consensus among the world's three leading frontier labs that the current AI training paradigm introduces significant risks from alignment failure [80,151]. A considerable body of research suggests that above a certain unknown capability threshold, AI systems will engage in power seeking behaviors by default. If these behaviors emerge in AI systems that have mastered real-world strategic planning in the same way that AI systems have mastered strategic planning in other contexts (including increasingly complex environments such as *StarCraft II* [152]), its impact could be uniquely severe.

Mitigating weaponization and alignment failure will likely reduce other risks. In order to address weaponization and alignment risks, new techniques will need to be

[15] **Annex C: Hypothetical alignment failure scenario** provides, for illustrative purposes, an example of a scenario in which power-seeking by a misaligned and highly capable AI leads to catastrophic harm.

developed to robustly predict and control the outputs of highly capable and context-aware systems, up to and including AGI. The ability to reliably control the outputs of AI systems would significantly increase their robustness to adversarial attacks and reduce the probability of prosaic accidents. It may also reduce the risk of societal harms, as properly aligned and powerful AI systems may behave in more corrigible ways, and may reflect more accurately the nuanced values of their developers. Policy controls aimed at weaponization and alignment risk, such as licensing and reporting requirements, would introduce oversight and monitoring capacity within the government that could support efforts to address societal and adversarial attack risks as well.

An alignment framing creates shared incentives among adversaries. As long as the alignment problem remains unsolved, the development of AI systems with capabilities that exceed a certain, unknown threshold introduces catastrophic and global risks, to which the developers of these systems are themselves exposed. Whereas weaponization risk incentivizes an arms race, alignment risk incentivizes global collaboration on cautious AI development. Framing international engagements through the lens of alignment may help to establish a common understanding of AI risk more broadly, shaped by shared incentives to develop sound governance measures for advanced AI.

3.4 Sources of AI catastrophic risk

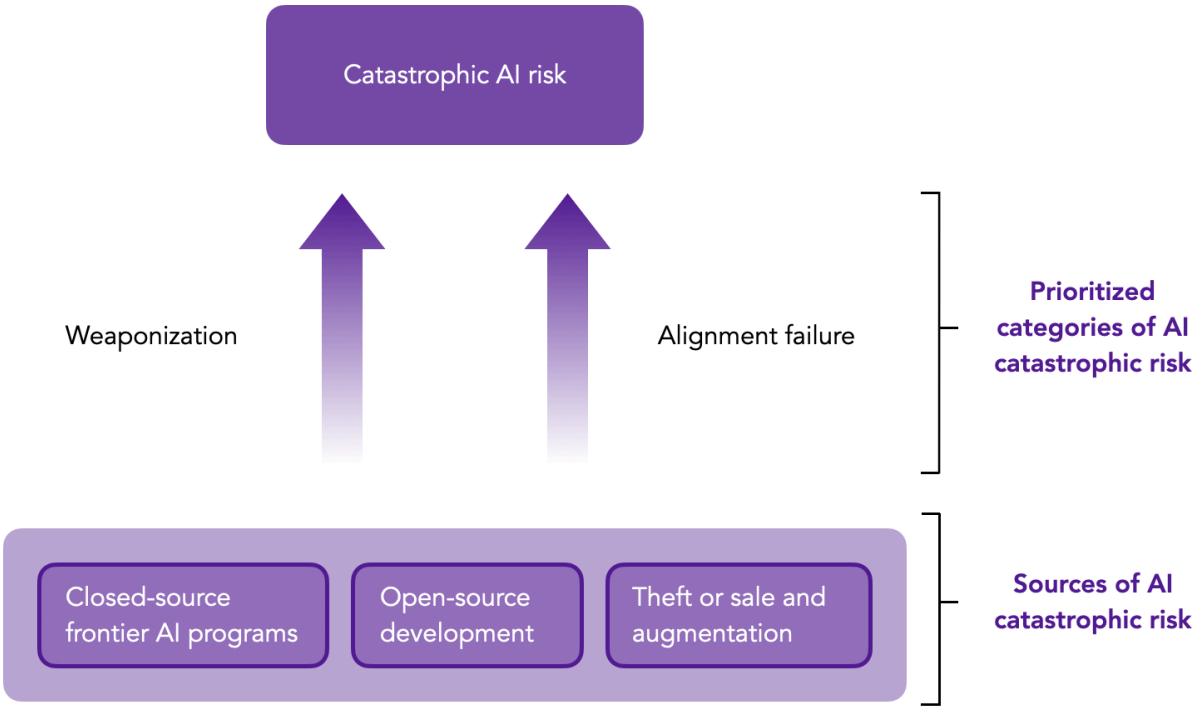


Figure 3. Sources and prioritized categories of catastrophic AI risk.

Currently, advanced AI models, and potential catastrophic risks associated with them, are proliferating via several key channels.

3.4.1 Closed-source development of frontier models

Closed-source development of frontier models occurs exclusively at well resourced labs with access to large quantities of compute, such as OpenAI, Google DeepMind, Microsoft, and Anthropic. As employees at leading labs leave their roles to join or found other AI organizations, their expertise diffuses with them. Leading Chinese labs based at organizations like the Beijing Academy of AI, Tsinghua University, Inspur, Huawei, Baidu, Microsoft Research Asia, and Tencent have benefited from this effect, and from a researcher base that includes individuals trained in top Western academic and industry research teams [153,154].

The widespread recognition of the power of scaling has led to an industry-wide and global race to scale up AI systems, in which a top tier of labs (most notably, OpenAI, Google DeepMind, and Anthropic) enjoy a modest but definite lead over second-tier

labs [24]. These frontier labs often monetize the outputs of their most powerful advanced AI systems by serving them to users in the form of paid-access APIs or chat interfaces [72,155]. They also have put some processes in place to reduce the risk of weaponization of these systems. But as frontier labs race to develop and deploy more and more powerful AI models, they have tended to reduce the level of attention they pay to safety, and have repeatedly weakened access restrictions in favor of drawing a larger base of paying customers [156].

In addition to the frontier labs, several lesser-known actors exist who have the talent, computational resources, and incentives to build frontier AI models at the cutting edge of capability and to deploy them with minimal safety controls. These include top quantitative hedge funds. Because these actors tend to be extremely secretive, we know very little about their current capabilities, future plans, or approaches to safety (if any). Nonetheless, it is reasonable to operate on the assumption that they are pursuing large-scale advanced AI development that could make them a source of weaponization, alignment, and accident risk.

3.4.2 Open-access release of increasingly powerful models

Open-access release of increasingly powerful models has quickly accelerated in recent months. Grassroots distributed organizations such as EleutherAI and BigScience have made major contributions [157–159], replicating and open-sourcing pre-trained models with capabilities that often match those of cutting-edge, proprietary AI systems only 12 to 18 months old. Similarly, private companies such as Stability AI and Meta have developed and released near-cutting-edge open-access models that may be readily weaponizable, and have stated their intention to continue doing so [92,93]. Indeed, at time of writing, the delay between a proprietary AI advance and the development of an open-access model with a comparable compute budget (a proxy for capabilities) may be as little as nine months (see Figure 4).

Although open-access models are unlikely to overtake the capabilities of cutting-edge proprietary AI models due to the importance and cost of scaling, some have argued that they meaningfully erode the competitive advantage available to leading AI labs [160]. This is likely true to an extent, and may have the effect of increasing pressure on frontier labs to accelerate their investments in scaling and capabilities in order to maintain their primacy.

In the context of open-access models, it is worth emphasizing that the key bottleneck to frontier AI development is the computing resources needed to train the model in the first place. Once a model is trained, downstream modifications such as fine-tuning or

prompt engineering can be used to introduce task-specific capabilities relatively cheaply [47]. For this reason, a large-scale AI model's release under an open-access license is an *irreversible* proliferation event. Once publicly available, these models can be cheaply modified and augmented by open-source developers with relative ease.

The open source AI ecosystem also extends beyond models. Large datasets [161], AI development toolkits [162,163], software frameworks [164], and other enabling technologies are also being rapidly developed and open-sourced. Initiatives like Together AI are also trying to make model training itself a distributed activity [165].

Software frameworks in particular are a key enabling technology. For example, Auto-GPT is an open-source framework that allows LLMs to autonomously carry out complex tasks that they could not natively perform [44]. This represents a significant capability leap, and shows that the open-source community can produce simple toolkits that massively augment the potential of existing AI models in unpredictable ways. One open-source developer used the Auto-GPT framework to create Chaos-GPT: an agent-like configuration of GPT-4 that was assigned the goal of destroying humanity and establishing global dominance [166]. Although Chaos-GPT was intended as a tongue-in-cheek side-project, it does suggest that an open-access AI that is capable enough to be dangerous is likely to be deliberately prompted to behave dangerously almost as soon as it is developed.

Open-source and open-access are vectors for the development of high-risk, weaponizable or potentially uncontrollable AI. It is conceivable that a frontier AI lab may develop and open-source a powerful AI model which is not weaponizable or uncontrollable in and of itself, but which, when modified via prompting, fine-tuning, or other techniques applied by open-access developers, could subsequently cross a high-risk capability threshold. Experiments have already demonstrated that the safety measures trained into leading open-access models can be trained out of these models easily [167], and can even be trained out *accidentally* on computational budgets of under a dollar [168]. Even the most advanced current safeguards applied to open-access models therefore do little to deter their weaponization: at best, they impose only a minor technical obstacle on anyone intending to misuse them.

The low cost to fine-tune frontier AI models makes AI a more complex source of catastrophic risk than other WMDs and WMD-enabling technologies. A pretrained open-access LLM such as Llama 2 [169] may have a relatively well-studied capability surface and risk profile, but for a few hundred dollars can be fine-tuned to acquire a wide range of potentially dangerous new capabilities, such as powering autonomous hacking agents [54]. It may therefore be most accurate to think of future LLMs not

merely as WMD-like technologies in themselves, but as *platforms* for the unpredictable (and, if they exist as open-access technologies, uncontrolled) development of a cluster of WMD-like systems, via cheap capability augmentations such as fine-tuning.

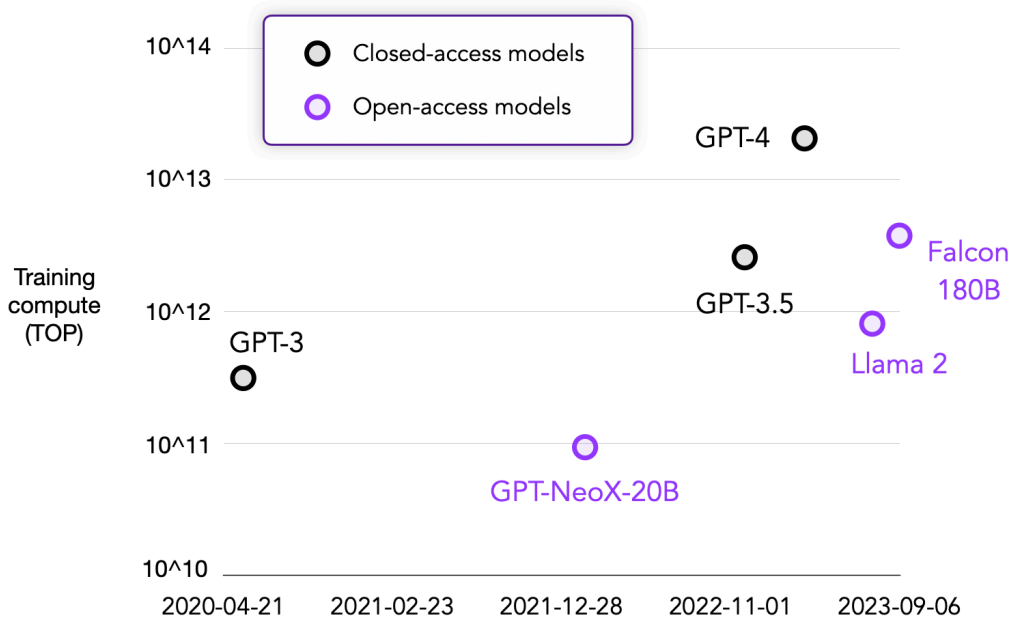


Figure 4. Training compute for indicated leading open-access (purple) and proprietary (black) models. Falcon 180B was trained nine months after GPT-3.5, and is estimated to have been trained on a comparable compute budget. Compute budgets are indicated in TOP, or trillions of operations. Data source: Epoch AI.

3.4.3 Model theft and piracy

Model theft and piracy, though a nascent problem, is already shaping the AI landscape. In March of 2023, a powerful LLM built by Meta, and originally released only to select academic researchers, was leaked online and shared as a downloadable torrent on 4chan [170]. The model, named LLaMA [171], has since been used by researchers as a backbone to build capable open-access alternatives to proprietary AI models. These include Alpaca [47], which by some measures approaches the performance of OpenAI's ChatGPT-3.5 at several tasks after only $600 worth of additional fine-tuning; and Vicuna [172], an open-access model estimated to match 90% of the performance of ChatGPT. Although it occurred without Meta's consent, the LLaMA leak may ultimately have advantaged the company, as open-source improvements on the LLaMA base model can now be incorporated into Meta's internal products directly.

Leading AI labs take measures to avoid accidentally exposing their closed-source models to risk of theft or extraction. Employees who need to access these models often must do so via APIs designed to allow only limited input-output access, to prevent exfiltration of model parameters. However, we assess that the security measures in place at many frontier AI labs are inadequate to resist a sustained IP exfiltration campaign by a nation-state attacker.

3.4.4 Sale of proprietary models

Sale of proprietary models, potentially under conditions not made public, may be an important proliferation vector for advanced AI models. As U.S. export controls on AI hardware make it more difficult for certain countries to develop domestic frontier AI programs, these countries will face increasing incentives to acquire, by sale or through other legal means, frontier AI models from U.S. firms.

3.5 Alignment strategies of frontier AI labs

There is currently no known AI alignment technique that can ensure that increasingly capable and context-aware AI systems remain under human control, nor is there a technical consensus around how difficult it would be to develop such a technique [151,173]. There is also great uncertainty regarding the amount of time safety researchers will have before AI systems are developed with the levels of capability and context awareness required to competently act on power-seeking incentives. As a result of these uncertainties, leading frontier AI labs are publicly considering a variety of different AI alignment threat models and pursuing various strategies to address them.

OpenAI prioritizes making increasingly powerful models gradually available to the public, to allow policymakers and members of the public to interact with and adapt to breakthroughs as they happen. This approach allows OpenAI to observe as their users manipulate, weaponize, and undermine their models in creative ways, and to develop and refine safety protocols to deal with the most effective attacks [174]. While opinions vary within the AI safety community, the organization's overall AI safety posture assumes fairly gradual progress in capabilities towards AGI, which they consider a source of global catastrophic risk. Under this assumption, OpenAI's view is that they can improve the safety and alignment of their systems incrementally, in tandem with those systems' increasing capabilities. However, OpenAI also acknowledges the need for a "superalignment" research effort, dedicated to solving alignment challenges they expect to emerge abruptly once AI systems are developed that exceed human capabilities across a wide range of tasks [151]. Finally, OpenAI has endorsed a strategy

of leveraging current AI systems themselves to help improve the alignment and safety of future AI systems, in keeping with their relatively incremental view of AI progress [175].

By contrast, Google DeepMind appears to favor a more opaque development process that puts significantly less emphasis on embedding AI systems into user-facing products. Until recently, DeepMind (when separate from Google) had been involved in very little research oriented toward production applications, prioritizing basic science and related applications instead [121—123]. Unlike OpenAI, who advocate for using increasingly capable AGI-like systems to help align incrementally more capable ones, Google DeepMind's alignment program also places somewhat less emphasis on AI-assisted alignment research. Instead, Google DeepMind emphasizes building up a technical understanding of the intentions of powerful AI systems, by interrogating their internal processing at a granular level. They are also attempting to develop reliable assessments of AI systems' capabilities to serve as an early warning if a system's capabilities ever exceeds predetermined safety limits [139].

Anthropic's approach is oriented toward addressing three distinct scenarios, and developing contingency plans for each [80]. First, if their technical research suggests that AI alignment is a straightforwardly solvable problem, they intend to focus their efforts on AI policy measures aimed at ensuring that AI developers actually implement appropriate safety measures when building potentially dangerous systems. Second, if instead their research suggests that AI alignment is a hard yet tractable problem, they anticipate investing more resources directly into technical safety research, with an emphasis on power-seeking and interpretability work. And third, if they assess that AI alignment is likely unsolvable, they plan to pivot their efforts toward gathering evidence to support that conclusion, and advancing policy measures that would prevent the development of dangerously powerful AI.

As AI expertise proliferates and more credible AI labs enter the race to build broadly human-level AI, new perspectives on alignment will likely emerge.

3.6 Risk throughout the AI development lifecycle

As advanced AI systems are further scaled and optimized, they may approach levels of capability that introduce catastrophic risks from weaponization or AI alignment failure. If these thresholds of capability are reached, different stages of the AI research and development process will introduce varying levels of risk. We distinguish between these stages and discuss their risk profiles below.

3.6.1 Training (lowest risk)

In the limit of extremely high AI capabilities, some catastrophic risk could be introduced during the training process itself.

- **Weaponization:** A model might be stolen by a third party at any stage during the training process. Therefore, model training itself intrinsically introduces weaponization risk.

- **Alignment failure:** During a typical training process, AI systems generate outputs, which are used to determine model parameter updates. In principle, a sufficiently capable AI may be able to manipulate even these training outputs in such a way as to gain some measure of control over its physical environment, leading to its escape from confinement (a scenario known as a break-out) [176].

3.6.2 Evaluation, benchmarking, red teaming, and internal deployment (moderate risk)

AI evaluations (or "evals") are carried out periodically during or shortly after the training process. They involve testing an AI system's performance at tasks that have been determined *a priori* to be of interest to model developers, often because they intend to leverage the system's performance at these tasks to power specific products or services. Once a model is trained, they are often supplemented by red teaming: manual efforts to elicit dangerous capabilities, or undermine the performance of a model for testing purposes [177].

- **Weaponization:** The internal deployment and testing of an AI model considerably increases the number of individuals with various degrees of access to it, and therefore the probability of model theft or leaks.

- **Alignment failure:** Because they involve human tests of AI system performance, and of potentially dangerous capabilities, AI evaluations carry greater risk than the training process itself. During an evaluation, an AI system may have opportunities to persuade its testers to give it access to additional resources, or to assist it in pursuing other power-seeking objectives, such as self-replication. This human interaction effectively allows the AI system to bridge the gap between cyberspace, which it natively occupies, and physical space, to which the human can serve as a malleable conduit [178].

> There is also heightened risk associated with more extensive post-training testing and red teaming aimed at uncovering pathological behavior that an AI system may display in edge cases. This process is more risky than evaluation from the perspective of alignment risk, since it can explicitly involve attempts to induce AI systems to perform dangerous behavior via techniques like red teaming. Like evaluation, this kind of testing is typically conducted by human testers, who can be persuaded and manipulated by a sufficiently capable system.

Safety oriented evaluations may in principle be useful for detecting dangerous model capabilities associated with psychological manipulation, deception, self-replication, which are often considered precursors to catastrophic weaponization or misalignment risks [178]. However, for very powerful systems, they come with the risk of *manifesting* these very behaviors: by testing a model's ability to self-replicate, for example, the model must be presented with an opportunity to do so. Care must be taken to ensure that this is done in a highly controlled manner, if it is to be attempted at all.

There is also the key question of what ought to be done if a safety evaluation reveals that a model does indeed possess "red flag" capabilities. Should the model be deleted and re-trained? Or should it be fine-tuned until evidence of the capability is absent? It is important to note that such approaches risk creating a selective pressure on models (and on their developers) to *hide* their most dangerous capabilities, since doing so would allow them to pass safety evaluations [179].

Finally, there is the challenge of shifting goalposts in model evaluations. GPT-4 was considered by many frontier AI researchers to be a dramatic advance over GPT-3, and the next leading models at the time of its release. Its new capabilities included a significant capacity for human deception, design of chemical attacks, and long-term planning. It is difficult to imagine that these capabilities would not have constituted "red flags" for any reasonable evaluation focused on catastrophic weaponization or alignment risk. And yet, the model was deployed for public use. Both its developers and the public have naturally adapted to this new capability and risk profile. There is a risk that this familiarity leads to a degree of complacency and a higher bar for what might qualify as "concerning" capabilities.

3.6.3 Standard deployment (high risk)

By standard deployment, we refer to the practice of deploying an AI system such as an LLM, and providing users with input-output access to the model via an API or a user interface. Crucially, in this setting, we imagine that the AI system operates in a stateless

manner: each series of inputs it receives from a user is processed by the system free of any other context. For example, each time you have a new conversation with ChatGPT, that conversation happens without taking into account the content of any previous conversations you had with it.

- **Weaponization**: standard deployment exposes a model to the public (or at least, to a wider set of users) as a packaged product. Although frontier labs apply safety and security filtering to user inputs and model outputs to prevent weaponization, these filters can often be defeated, and models induced to generate harmful or dangerous outputs through so-called jailbreaking techniques. In addition, open-source developers have shown that once a frontier model is exposed to users via an API, it can be used to generate training data for smaller models, which can cheaply approach the original frontier model's performance at specific tasks using this data [180]. As a result, standard deployment makes it much easier for threat actors to replicate the performance of a frontier model on tasks that can support weaponized applications.

- **Alignment failure**: When a powerful AI system is made available to a wide range of users, it becomes significantly more likely to encounter inputs that cause it to manifest dangerous power-seeking behavior, whether due to deliberate jailbreaking, or by accident.

3.6.4 Continuous deployment (very high risk)

An AI system that is deployed continuously is made available to users via a user interface or API, and is designed to operate in an open-ended fashion. In this configuration, an AI system has the ability to stack sequences of inferences together coherently, while maintaining an internal memory (and is therefore stateful). This allows the AI system to create and execute plans with longer explicit time horizons, making it more effective at pursuing complex objectives for weaponized applications, or sub-goals associated with power-seeking.

Notably, the barrier between standard deployment and continuous deployment can be surprisingly low in practice. In its initial release, for example, GPT-4 was made available to users under a standard deployment scheme, but frameworks such as Auto-GPT were developed shortly thereafter, which allowed users to interact with GPT-4 in a way that was effectively stateful (and therefore continuous) [40].

3.6.5 Continuous learning (highest risk)

In a continuous learning setting, an AI system is allowed to operate in an open-ended manner, while also updating its internal state, parameters, or external memory to improve its performance over time.

This is the most unconstrained operating mode currently possible for an AI system, and would introduce the greatest degree of weaponization and alignment risks for sufficiently capable systems. Free from constraints associated with stateless deployment, and able to accumulate context about its environment over extended periods of time, an AI model deployed in this configuration would have the greatest possible freedom to develop and execute attacks when weaponized, or pursue dangerous power-seeking behavior autonomously.

4. Conclusion

Progress in AI has recently undergone a dramatic acceleration. To a significant degree, this acceleration has been driven by the discovery and optimization of robust AI scaling laws, which allow larger compute and data budgets to be reliably converted into increases in AI capabilities. As AI capabilities increase, the destructive footprint of malicious actors who may weaponize these capabilities is likely to grow rapidly and significantly. But advanced AI systems also remain fragile, and often fail in unexpected contexts. As advanced AI is deployed more widely, and depended upon for increasingly critical applications, accidents and adversarial attacks aimed at undermining these systems may have catastrophic effects.

Based on the current trajectory of AI progress, it seems plausible that flexible, human-level general AI may be developed over relatively short timescales. Such a development would introduce novel and critical risks. These include potentially catastrophic risks from weaponization and loss of control over systems that may possess long-term planning, strategic reasoning, malware development, and other dangerous capabilities.

A relatively small number of AI labs are currently advancing the frontier of AI capabilities, and these labs were founded in part on the basis of concerns over the dangers associated with uncontrollable forms of highly advanced AI. In spite of their stated safety concerns, these labs are currently locked in a race to build increasingly powerful systems, and have chosen to accelerate their capabilities output as a result. For now, the main frontier AI labs are overwhelmingly based in the West. Although assessing relative rates of progress of leading AI labs is challenging, other players may be as little as 6 to 12 months behind. Open-source development has also emerged as an important source of AI capabilities progress, and in some scenarios, may represent an important vector of risk for power-seeking, broadly human-level AI.

Bibliography

Note: progress at the frontier of AI is driven primarily by private labs, which often publish research reports that have not been subjected to third party peer review. Prominent examples of this trend include the technical reports that OpenAI and Google respectively published about their then-leading models, GPT-4 and PaLM 2.

This is partly because private companies do not face the same incentives to publish peer reviewed research as academic labs, but also because publication norms in the AI research community are highly informal relative to other STEM fields. A surprising fraction of important research theses and empirical results in AI are published as blog posts, and in some cases even by pseudonymous researchers. Indeed, one of the standard reference documents that first introduced the AI scaling hypothesis was a blog post entitled "The Bitter Lesson", which was unceremoniously published by AI researcher Richard Sutton on his personal website.

As a result, the entries in the bibliography below include a combination of sources primarily consisting of peer reviewed papers and well-regarded research publications that have nonetheless not been subjected to explicit peer review. When appropriate, we also reference blog posts (such as Richard Sutton's "The Bitter Lesson"). In each instance, we have reviewed the source closely, and have concluded that it does support the claim(s) with which it is associated in the body of this document.

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

[3] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T. and Eccles, T., 2022. A generalist agent. arXiv preprint arXiv:2205.06175.

[4] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S. and Schuh, P., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[5] Introducing claude (2023) Anthropic. Available at: https://www.anthropic.com/index/introducing-claude (Accessed: 02 November 2023).

[6] Steinhardt, J. (2023) AI forecasting: Two years in, Bounded Regret. Available at: https://bounded-regret.ghost.io/scoring-ml-forecasts-for-2023/ (Accessed: 02 November 2023).

[7] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[8] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A. and Hennigan, T., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

[9] Introducing chatgpt (2022) Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt (Accessed: 02 November 2023).

[10] OpenAI, 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[11] Google, 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

[12] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T. and Huang, W., 2023. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378.

[13] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. and Ring, R., 2022. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35,

[14] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A. and Misra, I., 2023. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15180-15190).

[15] Kavukcuoglu, K. (2021) Real-world challenges for agi, Google DeepMind. Available at: https://deepmind.google/discover/blog/real-world-challenges-for-agi/ (Accessed: 02 November 2023).

[16] (No date) About. Available at: https://openai.com/about (Accessed: 02 November

[17] Kerner, S.M. (2023) Elon Musk reveals XAI efforts, predicts full AGI by 2029. Available at: https://venturebeat.com/ai/elon-musk-reveals-xai-efforts-predicts-full-agi-by-2029/ (Accessed: 02 November 2023).

[18] AI could be one of humanity's most useful inventions (no date) About. Available at: https://web.archive.org/web/20230703012703/https://www.deepmind.com/about (Accessed: 02 November 2023).

[19] Roser, M. (2023) AI timelines: What do experts in artificial intelligence expect for the future?, Our World in Data. Available at: https://ourworldindata.org/ai-timelines (Accessed: 02 November 2023).

[20] Domingos, P. (2015) The Master Algorithm: How The Quest for the Ultimate Learning Machine will remake our world. New York: Basic books.

[21] Heaven, W.D. (2023) Geoffrey Hinton tells us why he's now scared of the Tech he helped build, MIT Technology Review. Available at: https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/ (Accessed: 02 November 2023).

[22] Sutton, R. (2019) The Bitter Lesson, Incomplete Ideas. Available at: http://www.incompleteideas.net/IncIdeas/BitterLesson.html (Accessed: 02 November 2023).

[23] Hu, K. (2023) CHATGPT sets record for fastest-growing user base - analyst note, Reuters. Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Accessed: 02 November 2023).

[24] Ai Tracker (no date) AI Tracker. Available at: https://www.aitracker.org/ (Accessed: 02 November 2023).

[25] Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X. and Li, C., 2021. Pangu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. arXiv preprint arXiv:2104.12369.

[26] Kim, B., Kim, H., Lee, S.W., Lee, G., Kwak, D., Jeon, D.H., Park, S., Kim, S., Kim, S., Seo, D. and Lee, H., 2021. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. arXiv preprint arXiv:2109.04650.

[27] Tarantola, A. (2021) China's gigantic multi-modal AI is no one-trick pony, Engadget. Available at: https://www.engadget.com/chinas-gigantic-multi-modal-ai-is-no-one-trick-pony-211414388.html (Accessed: 02 November 2023).

[28] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G. and Ray, A., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

[29] Dall·E 2 (no date) DALL·E 2. Available at: https://openai.com/dall-e-2 (Accessed: 02 November 2023).

[30] Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D. and Duan, N., 2022, October. Nüwa: Visual synthesis pre-training for neural visual world creation. In European conference on computer vision (pp. 720-736). Cham: Springer Nature Switzerland.

[31] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[32] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[33] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K. and Herzog, A., 2022. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.

[34] Villalobos, P. (2023) Scaling laws literature review, Epoch. Available at: https://epochai.org/blog/scaling-laws-literature-review (Accessed: 02 November 2023).

[35] Sharma, U. and Kaplan, J., 2020. A neural scaling law from the dimension of the data manifold. arXiv preprint arXiv:2004.10802.

[36] Bahri, Y., Dyer, E., Kaplan, J., Lee, J. and Sharma, U., 2021. Explaining neural scaling laws. arXiv preprint arXiv:2102.06701.

[37] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S. and Nori, H., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

[38] LeCun, Y., 2023, March. Do Large Language Models Need Sensory Grounding for Meaning and Understanding?. In Workshop on Philosophy of Deep Learning, NYU Center for Mind, Brain, and Consciousness and the Columbia Center for Science and Society.

[39] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. and Villalobos, P., 2022, July. Compute trends across three eras of machine learning. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[40] Time for AI to cross the human performance range in ImageNet Image Classification (2021) AI Impacts. Available at: https://aiimpacts.org/time-for-ai-to-cross-the-human-performance-range-in-imagenet-image-classification/ (Accessed: 02 November 2023).

[41] Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C. and Józefowicz, R., 2019. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680.

[42] Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A.S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J. and Quan, J., 2017. Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:1708.04782.

[43] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H. and Jacob, A.P., 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science, 378(6624), pp.1067-1074.

[44] Wiggers, K. (2023) What is auto-GPT and why does it matter?, TechCrunch. Available at: https://techcrunch.com/2023/04/22/what-is-auto-gpt-and-why-does-it-matter/ (Accessed: 02 November 2023).

[45] Nakajima, Y. (2023) 'BabyAGI'. GitHub. Available at: https://github.com/yoheinakajima/babyagi (Accessed: 02 November 2023).

[46] Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H. and Han, J., 2022. Large language models can self-improve. arXiv preprint arXiv:2210.11610.

[47] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T.B., 2023. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6), p.7.

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[49] Korthikanti, V.A., Casper, J., Lym, S., McAfee, L., Andersch, M., Shoeybi, M. and Catanzaro, B., 2023. Reducing activation recomputation in large transformer models. Proceedings of Machine Learning and Systems, 5.

[50] Manakul, P., Liusie, A. and Gales, M.J., 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

[51] Zhao, Z., Wallace, E., Feng, S., Klein, D. and Singh, S., 2021, July. Calibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning (pp. 12697-12706). PMLR.

[52] Liu, L., Qu, Z., Chen, Z., Ding, Y. and Xie, Y., 2021. Transformer acceleration with dynamic sparse attention. arXiv preprint arXiv:2110.11299.

[53] Martins, P.H., Marinho, Z. and Martins, A.F., 2021. ∞-former: Infinite Memory Transformer. arXiv preprint arXiv:2109.00301.

[54] Wu, Y., Rabe, M.N., Hutchins, D. and Szegedy, C., 2022. Memorizing transformers. arXiv preprint arXiv:2203.08913.

[55] ACT-1: Transformer for actions (2022) ACT-1: Transformer for Actions. Available at: https://www.adept.ai/blog/act-1 (Accessed: 02 November 2023).

[56] Dao, T., Fu, D., Ermon, S., Rudra, A. and Ré, C., 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems, 35, pp.16344-16359.

[57] Bansal, T., Mordatch, I., Pachocki, J., Sutskever, I., Sidor, S. (2017) Competitive self-play. Available at: https://openai.com/research/competitive-self-play (Accessed: 02 November 2023).

[58] Bulatov, A., Kuratov, Y. and Burtsev, M., 2022. Recurrent memory transformer. Advances in Neural Information Processing Systems, 35, pp.11079-11091.

[59] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T. and Lillicrap, T., 2020. Mastering atari, go, chess and shogi by planning with a learned model. Nature, 588(7839), pp.604-609.

[60] Knight, W. (2023) Google DeepMind CEO Demis Hassabis says its next algorithm will eclipse CHATGPT, Wired. Available at: https://www.wired.com/story/google-deepmind-demis-hassabis-chatgpt/ (Accessed: 02 November 2023).

[61] Cottier, B. (2023) Trends in the dollar training cost of machine learning systems, Epoch. Available at: https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems (Accessed: 02 November 2023).

[62] Li, C. (2023) OpenAI's GPT-3 Language model: A technical overview, GPU Cloud, Clusters, Servers, Workstations. Available at: https://lambdalabs.com/blog/demystifying-gpt-3 (Accessed: 02 November 2023).

[63] Venigalla, A. and Li, L. (2022) Mosaic Llms (part 2): GPT-3 quality for <$500K, MosaicML. Available at: https://www.mosaicml.com/blog/gpt-3-quality-for-500k (Accessed: 02 November 2023).

[64] 2023. [PUBLIC] Cost estimates for GPT-4.ipynb. Available at: https://colab.research.google.com/drive/1O99z9b1I5O66bT78r9ScslE_nOj5irN9?usp=sharing#scrollTo=GL-UbtMq82kr (Accessed: 02 November 2023).

[65] Knight, W. (2023b) OpenAI's CEO says the age of giant AI models is already over, Wired. Available at: https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/ (Accessed: 02 November 2023).

[66] Bass, D. (2023) Microsoft to invest $10 billion in Chatgpt Maker openai (MSFT), Bloomberg.com. Available at: https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai (Accessed: 02 November

[67] Elon Musk reportedly purchases thousands of GPUs for generative AI project at Twitter (2023) Ars Technia. Available at: https://arstechnica.com/information-technology/2023/04/elon-musk-reportedly-purchases-thousands-of-gpus-for-generative-ai-project-at-twitter/ (Accessed 06 December 2023).

[68] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman, J., 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, pp.27730-27744.

[69] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[70] Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P. and Campbell-Gillingham, L., 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.

[71] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W. and Do, Q.V., 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

[72] Chatgpt plugins (2023) ChatGPT plugins. Available at: https://openai.com/blog/chatgpt-plugins (Accessed: 02 November 2023).

[73] 2023. DALL·E 3. Available at: https://openai.com/dall-e-3 (Accessed: 02 November 2023).

[74] Mehdi, Y. (no date) Confirmed: the new Bing runs on OpenAI's GPT-4, Bing. Available at: https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4 (Accessed: 02 November 2023).

[75] Timothy, M. (2023) Will there be a GPT-5? when Will GPT-5 launch?, MUO. Available at: https://www.makeuseof.com/when-will-gpt5-launch/ (Accessed: 02 November 2023).

[76] Hassabis, D. (2023) Announcing google deepmind, Google DeepMind. Available at: https://www.deepmind.com/blog/announcing-google-deepmind (Accessed: 02 November 2023).

[77] Pichai, S. (2023) Google I/O 2023: Making ai more helpful for everyone, Google. Available at: https://blog.google/technology/ai/google-io-2023-keynote-sundar-pichai/ (Accessed: 02 November 2023).

[78] Field, H. (2021) Ex-openai employees create anthropic, an AI safety and research startup, Tech Brew. Available at: https://www.emergingtechbrew.com/stories/2021/06/02/exopenai-employees-create-anthropic-ai-safety-research-startup (Accessed: 02 November 2023).

[79] Dario and Daniela Amodei (2023) Time. Available at: https://time.com/collection/time100-ai/6309047/daniela-and-dario-amodei/ (Accessed 06 December 2023).

[80] Core views on AI safety: When, why, what, and how (2023) Anthropic. Available at: https://www.anthropic.com/index/core-views-on-ai-safety (Accessed: 02 November 2023).

[81] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C. and Chen, C., 2022. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

[82] Wiggers, K., Coldewey, D. and Singh, M. (2023) Anthropic's $5B, 4-year plan to take on OpenAI, TechCrunch. Available at: https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/ (Accessed: 02 November 2023).

[83] Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V. and Zhang, E., 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990.

[84] Bedrock (1996) Amazon. Available at: https://aws.amazon.com/bedrock/titan/ (Accessed: 02 November 2023).

[85] World's Top 10 Hedge Funds (2023) Investopedia. Available at: https://www.investopedia.com/articles/personal-finance/011515/worlds-top-10-hedge-fund-firms.asp (Accessed 06 December, 2023)

[86] Reuters (2023) China's Tencent establishes team to develop CHATGPT-like product, Brecorder. Available at: https://www.brecorder.com/news/40228621 (Accessed: 02 November 2023).

[87] Yu, I. (2023) Alibaba Cloud debuts generative AI model for Corporate users , Alizila. Available at: https://www.alizila.com/alibaba-cloud-debuts-generative-ai-model-for-corporate-users/ (Accessed: 02 November 2023).

[88] Wu, S., Zhao, X., Yu, T., Zhang, R., Shen, C., Liu, H., Li, F., Zhu, H., Luo, J., Xu, L. and Zhang, X., 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. arXiv preprint arXiv:2110.04725.

[89] Baidu unveils Ernie Bot, the latest Generative AI mastering Chinese language and multi-modal generation (2023) PR Newswire. Available at: https://www.prnewswire.com/news-releases/baidu-unveils-ernie-bot-the-latest-generative-ai-mastering-chinese-language-and-multi-modal-generation-301774240.html (Accessed: 02 November 2023).

[90] Ma, Z., He, J., Qiu, J., Cao, H., Wang, Y., Sun, Z., Zheng, L., Wang, H., Tang, S., Zheng, T. and Lin, J., 2022, April. BaGuaLu: targeting brain scale pretrained models with over 37 million cores. In Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (pp. 192-204).

[91] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X. and Tam, W.L., 2022. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.

[92] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V. and Mihaylov, T., 2022. OPT: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

[93] Stability AI. (2023) 'StableLM'. GitHub. Available at: https://github.com/Stability-AI/StableLM (Accessed: 02 November 2023).

[94] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. and Launay, J., 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.

[95] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

[96] Branwen, G. (2020) GPT-3 Creative Fiction, Gwern.net. Available at: https://gwern.net/gpt-3 (Accessed: 02 November 2023).

[97] Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J.C., Parli, V. and Shoham, Y., 2023. Artificial intelligence index report 2023. arXiv preprint arXiv:2310.03715.

[98] Knight, W. (2023c) What really made Geoffrey Hinton into an AI doomer, Wired. Available at: https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/ (Accessed: 02 November 2023).

[99] Bengio, Y. (2023) Personal and psychological dimensions of AI researchers confronting AI catastrophic risks, Yoshua Bengio. Available at: https://yoshuabengio.org/2023/08/12/personal-and-psychological-dimensions-of-ai-researchers-confronting-ai-catastrophic-risks/ (Accessed: 02 November 2023).

[100] Bove, T. (2023) A.I. could rival human intelligence in 'Just a few years,' says CEO of Google's main A.I. Research Lab, Fortune. Available at: https://fortune.com/2023/05/03/google-deepmind-ceo-agi-artificial-intelligence/ (Accessed: 02 November 2023).

[101] Sullivan, M. (2023) Why everyone seems to disagree on how to define artificial general ... Available at: https://www.fastcompany.com/90968623/why-everyone-seems-to-disagree-on-how-to-define-artificial-general-intelligence (Accessed: 02 November

[102] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

[103] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

[104] Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A. and Hubert, T., 2022. Competition-level code generation with alphacode. Science, 378(6624), pp.1092-1097.

[105] Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T. and Wu, Y., 2022. Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems, 35, pp.3843-3857.

[106] Brewster, T. (2023) GPT-4 can't stop helping hackers make Cybercriminal Tools, Forbes. Available at: https://www.forbes.com/sites/thomasbrewster/2023/03/16/gpt-4-could-help-stupid-hackers-become-good-cybercriminals/ (Accessed: 02 November 2023).

[107] Hutchens, J. (2024) Language of deception: Weaponizing next generation AI. S.l.: JOHN WILEY.

[108] Larsen, L. (2023) This viral AI-generated image fooled everyone this weekend, Digital Trends. Available at: https://www.digitaltrends.com/computing/this-ai-generated-image-of-pope-fooled-everyone/ (Accessed: 02 November 2023).

[109] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. and Scialom, T., 2023. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.

[110] Xiang, C. (2023) 'he would still be here': Man dies by suicide after talking with AI chatbot, widow says, VICE. Available at: https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says (Accessed: 02 November 2023).

[111] Delouya, S. (2023) Replika users say they fell in love with their AI chatbots, until a software update made them seem less human, Business Insider. Available at: https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2 (Accessed: 02 November 2023).

[112] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J. and He, L., 2023. Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111.

[113] Verma, P. (2023) They thought loved ones were calling for help. it was an AI scam., The Washington Post. Available at: https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/ (Accessed: 02 November 2023).

[114] Thompson, D. (2023) The role of GPT-4 in drug discovery ⋆ vial, Vial. Available at: https://vial.com/blog/articles/the-role-of-gpt-4-in-drug-discovery/ (Accessed: 02 November 2023).

[115] Frey, N.C., Soklaski, R., Axelrod, S., Samsi, S., Gomez-Bombarelli, R., Coley, C.W. and Gadepally, V., 2023. Neural scaling of deep chemical models. Nature Machine Intelligence, pp.1-9.

[116] Urbina, F., Lentzos, F., Invernizzi, C. and Ekins, S., 2022. Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4(3), pp.189-191.

[117] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E., 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. science, 337(6096), pp.816-821.

[118] Wan, A., Wallace, E., Shen, S. and Klein, D., 2023. Poisoning Language Models During Instruction Tuning. arXiv preprint arXiv:2305.00944.

[119] Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T. and Shokri, R., 2022. Quantifying privacy risks of masked language models using membership inference attacks. arXiv preprint arXiv:2203.03929.

[120] Keary, T. (2023) How prompt injection can hijack autonomous AI agents like auto-GPT. Available at: https://venturebeat.com/security/how-prompt-injection-can-hijack-autonomous-ai-agents-like-auto-gpt/ (Accessed: 02 November 2023).

[121] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), pp.583-589.

[122] Kirkpatrick, J., McMorrow, B., Turban, D.H., Gaunt, A.L., Spencer, J.S., Matthews, A.G., Obika, A., Thiry, L., Fortunato, M., Pfau, D. and Castellanos, L.R., 2021. Pushing the frontiers of density functionals by solving the fractional electron problem. Science, 374(6573), pp.1385-1389.

[123] Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D. and Donner, C., 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 602(7897), pp.414-419.

[124] Kalliamvakou, E. (2023) Research: Quantifying github copilot's impact on developer productivity and happiness, The GitHub Blog. Available at: https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/ (Accessed: 02 November 2023).

[125] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[126] Arnold, Z. and Toner, H. (2023) AI accidents: An emerging threat, Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/ (Accessed: 02 November 2023).

[127] Critch, A. and Krueger, D., 2020. AI research considerations for human existential safety (ARCHES). arXiv preprint arXiv:2006.04948.

[128] Harris, E. and Suo, S. (2022) Instrumental convergence in single-agent systems, AI Alignment Forum. Available at: https://www.alignmentforum.org/posts/pGvM95EfNXwBzjNCJ/instrumental-convergence-in-single-agent-systems (Accessed: 02 November 2023).

[129] Mikulik, V. (2019) 2-D robustness - AI alignment forum, AI Alignment Forum. Available at: https://www.alignmentforum.org/posts/2mhFMgtAjFJesaSYR/2-d-robustness (Accessed: 02 November 2023).

[130] Lin, S., Hilton, J. and Evans, O., 2021. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

[131] Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P. and Wang, T., 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint

[132] Krakovna, V. et al. (2020) Specification gaming: The flip side of Ai Ingenuity, Google DeepMind. Available at: https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ (Accessed: 02 November 2023).

[133] Clark, J. and Amodei, D. (2016) Faulty reward functions in the wild. Available at: https://openai.com/research/faulty-reward-functions (Accessed: 02 November 2023).

[134] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J. and Garrabrant, S., 2019. Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820.

[135] Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J. and Kenton, Z., 2022. Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv preprint arXiv:2210.01790.

[136] Hubinger, E. (2019) Gradient hacking - AI alignment forum, AI Alignment Forum. Available at: https://www.alignmentforum.org/posts/uXH4r6MmKPedk8rMA/gradient-hacking (Accessed: 02 November 2023).

[137] Turner, A.M., Smith, L., Shah, R., Critch, A. and Tadepalli, P., 2019. Optimal policies tend to seek power. arXiv preprint arXiv:1912.01683.

[138] Carlsmith, J., 2022. Is Power-Seeking AI an Existential Risk?. arXiv preprint arXiv:2206.13353.

[139] Krakovna, V. (no date) Some high-level thoughts on the DeepMind alignment team's strategy. Available at: https://drive.google.com/file/d/1DVPZz0-9FSYgrHFgs4NCN6kn2tE7J8AK/view (Accessed: 02 November 2023).

[140] Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J. and Kenton, Z., 2022. Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv preprint arXiv:2210.01790.

[141] Omohundro, S.M., 2008, February. The basic AI drives. In AGI (Vol. 171, pp. 483-492).

[142] Ngo, R., Chan, L. and Mindermann, S., 2022. The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626.

[143] Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S. and Jones, A., 2022. Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251.

[144] Turpin, M., Michael, J., Perez, E. and Bowman, S.R., 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv preprint arXiv:2305.04388.

[145] Bengio, Y. (2023a) How rogue AIS may arise, Yoshua Bengio. Available at: https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/ (Accessed: 02 November

[146] Vallance, Z.K.& C. (2023) Ai 'godfather' Geoffrey Hinton warns of dangers as he quits Google, BBC News. Available at: https://www.bbc.com/news/world-us-canada-65452940 (Accessed: 02 November 2023).

[147] N. Bostrom, Superintelligence: Paths, Dangers, Strategies. London, England: Oxford University Press, 2016

[148] Ebrahimi, M., Zhang, N., Hu, J., Raza, M.T. and Chen, H., 2020. Binary black-box evasion attacks against deep learning-based static malware detectors with adversarial byte-level language model. arXiv preprint arXiv:2012.07994.

[149] AI, cybersecurity, hacking, China, Taiwan: Homeland security newswire (2020) AI, cybersecurity, hacking, China, Taiwan | Homeland Security Newswire. Available at: https://www.homelandsecuritynewswire.com/dr20200106-artificial-intelligence-china-uses-taiwan-for-target-practice-as-it-perfects-cyberwarfare-techniques (Accessed: 02 November 2023).

[150] Delano, J. (2023) Ai scam artists impersonate familiar voices to scam the rest of Us, CBS News. Available at: https://www.cbsnews.com/pittsburgh/news/ai-scam-artists-impersonate-familiar-voices-scams/ (Accessed: 02 November 2023).

[151] Leike, J. and Sutskever, I. (2023) Introducing Superalignment. Available at: https://openai.com/blog/introducing-superalignment (Accessed: 02 November 2023).

[152] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P. and Oh, J., 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), pp.350-354.

[153] Huang, T. and Arnold, Z. (2020) Immigration policy and the global competition for Ai Talent. Available at: https://cset.georgetown.edu/wp-content/uploads/CSET-Immigration-Policy-and-the-Global-Competition-for-AI-Talent.pdf (Accessed: 03

[154] Hannas, W.C. and Chang, H.M., 2019. China's Access to Foreign AI Technology. Center for Security and Emerging Technology, September, 13.

[155] (No date a) GPT-4. Available at: https://openai.com/product/gpt-4 (Accessed: 02 November 2023).

[156] Skelton, S.K. (2023) MPs warned of Ai Arms Race to the bottom: Computer Weekly, ComputerWeekly.com. Available at: https://www.computerweekly.com/news/365529793/MPs-warned-of-AI-arms-race-to-the-bottom (Accessed: 02 November 2023).

[157] (2023) ELEUTHERAI/GPT-neo-1.3B · hugging face. Available at: https://huggingface.co/EleutherAI/gpt-neo-1.3B (Accessed: 02 November 2023).

[158] (2023a) ELEUTHERAI/GPT-J-6B · hugging face. Available at: https://huggingface.co/EleutherAI/gpt-j-6b (Accessed: 02 November 2023).

[159] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M. and Tow, J., 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

[160] Patel, D. and Ahmad, A. (2023) Google 'we have no moat, and neither does openai', SemiAnalysis. Available at: https://www.semianalysis.com/p/google-we-have-no-moat-and-neither (Accessed: 02 November 2023).

[161] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N. and Presser, S., 2020. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

[162] TRL (no date) Transformer Reinforcement Learning. Available at: https://huggingface.co/docs/trl/index (Accessed: 02 November 2023).

[163] CarperAI, C. (2023) TRLX: A repo for distributed training of language models with reinforcement learning via human feedback (RLHF). Available at: https://github.com/CarperAI/trlx (Accessed: 02 November 2023).

[164] Introduction (no date) 🦜🔗 Langchain. Available at: https://python.langchain.com/docs/get_started/introduction (Accessed: 02 November 2023).

[165] (No date a) Together AI. Available at: https://www.together.xyz/ (Accessed: 02 November 2023).

[166] Lanz, J.A. (2023) Meet Chaos-GPT: An AI tool that seeks to destroy humanity, Decrypt. Available at: https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity (Accessed: 02 November 2023).

[167] Lermen, S., Rogers-Smith, C., Ladish, J., 2023. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. arXiv preprint arXiv:2310.20624

[168] Qi, X., Zeng, Y., Xie, T., Chen, P.Y., Jia, R., Mittal, P. and Henderson, P., 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. arXiv preprint arXiv:2310.03693.

[169] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[170] Vincent, J. (2023) Meta's powerful AI language model has leaked online - what happens now?, The Verge. Available at: https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse (Accessed: 02 November 2023).

[171] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[172] Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E. and Stoica, I., 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. See https://vicuna. lmsys. org (accessed 14 April 2023).

[173] Harris, J. (2021) The inner alignment problem, Medium. Available at: https://towardsdatascience.com/the-inner-alignment-problem-9eb5f234226b (Accessed: 02 November 2023).

[174] Altman, S. (2023) Planning for AGI and beyond. Available at: https://openai.com/blog/planning-for-agi-and-beyond (Accessed: 02 November 2023).

[175] Leike, J., Schulman, J. and Wu, J. (2022) Our approach to alignment research. Available at: https://openai.com/blog/our-approach-to-alignment-research (Accessed: 02 November 2023).

[176] Babcock, J., Kramár, J. and Yampolskiy, R., 2016. The AGI containment problem. In Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9 (pp. 53-63). Springer International Publishing.

[177] Casper, S., Lin, J., Kwon, J., Culp, G. and Hadfield-Menell, D., 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. arXiv preprint arXiv:2306.09442.

[178] Update on ARC's recent eval efforts (2023) ARC Evals. Available at: https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/ (Accessed: 02 November 2023).

[179] Carranza, A., Pai, D., Schaeffer, R., Tandon, A. and Koyejo, S., 2023. Deceptive Alignment Monitoring. arXiv preprint arXiv:2307.10569.

[180] Peng, B., Li, C., He, P., Galley, M. and Gao, J., 2023. Instruction tuning with GPT-4. arXiv preprint arXiv:2304.03277.

Annex A: Glossary of terms

Advanced AI. Any AI system capable of performing a wide range of tasks. This includes, but is not limited to, AGI-level systems and frontier AI systems. Currently-existing systems such as GPT-3 and ChatGPT are examples of advanced AI systems. An advanced AI system can generally be used for malicious purposes unless the AI model developer makes a specific effort to prevent this. As a result, the infrastructure for training advanced AI systems supports dual-use capabilities.

Frontier AI. Advanced AI systems that are at the current frontier of capabilities. As of early September 2023, GPT-4 and PaLM 2 are examples of frontier AI systems. Organizations that are able to develop frontier AI models are called frontier AI labs. OpenAI and Google DeepMind (who developed GPT-4 and PaLM 2, respectively) are examples of frontier labs.

AGI. An AI system that is sufficiently advanced to outperform humans across a broad range of economic and strategic domains, such as producing practical long-term plans that are likely to work under real world conditions. In particular, an AGI may have the capability to autonomously circumvent human or institutional controls on its actions, including any controls imposed by its developers. While the precise threshold is under debate, frontier AI safety researchers broadly expect advanced AI systems to reach this point as they approach and begin to surpass human capabilities in a broad enough range of domains. These domains may or may not include situational awareness, deception, and effective representation of complex concepts. For clarity, this definition of AGI does not refer to or imply sentience, consciousness, or self-awareness. It solely refers to the system's problem-solving ability.

Outer alignment. Reliably encoding human desires into a goal that we would be comfortable seeing an arbitrarily capable AGI system pursue. Outer alignment is an unsolved problem in technical AI safety.

Inner alignment. Ensuring that a given, formally specified goal is pursued reliably by any AGI system with arbitrarily high and general capabilities. Inner alignment is an unsolved problem in technical AI safety. Inner alignment is also distinct from ensuring that a goal is merely understood reliably by an AGI-level system, which is believed to be an easier problem.

Controllability. A controllable AI system is one whose developers can correct or otherwise affect the system's goals or behavior after the fact, if they have failed to correctly implement both inner alignment and outer alignment in the system from the very beginning. Current advanced AI systems are controllable by virtue of their limited capabilities.

Annex B: Frequently asked questions about alignment risk

AI alignment risk is a complex topic, and is often poorly understood outside technical safety teams at frontier AI labs. For this reason, nontechnical stakeholders (and even experienced AI capabilities researchers) often have questions about some of the nuances behind arguments for catastrophic risk from alignment failure in powerful AI systems. In this Annex, we will list and answer some common questions on this topic.

1. Isn't AI progress over-hyped? I've heard some credible-sounding technical people say that human-level AI capabilities are a long way off, if they're even possible at all.

There is significant uncertainty about when broadly human-level AI will be developed, and some AI researchers believe that new breakthroughs will be required to reach it. However, there are also strong reasons to expect that human-level AI may be developed surprisingly soon.

First, the AI scaling principle that led to the development of GPT-3 has continued to generate further, remarkable capability leaps. Where GPT-3 could write humanlike tweets, GPT-4 can score in the 80th percentile or above in a wide range of standardized tests – many of which are designed for highly educated human professionals. And the AI scaling curves that leading labs have relied on to push forward AI capabilities show no sign of slowing down: in some cases, they have held firm over more than seven orders of magnitude in their inputs. All of this, in a context where current systems have already reached superhuman levels of performance across a wide – and rapidly widening – range of tasks.

Further, the scaling phenomenon now appears to apply to domains well beyond language. Systems like Gato and PaLM-E have shown that language models can be used as a backbone for generalist agents that can perform hundreds of tasks – including tasks involving robotic manipulation and image processing – as well as human beings.

Leading AI labs are now investing tens of billions of dollars in compute infrastructure on the assumption that scaling laws will continue to hold. It is unclear how many further leaps of the kind we saw from GPT-3 to GPT-4 may remain between the current state of the art, and dangerously powerful or broadly human-level AI, but it seems plausible that it isn't a huge number. Indeed, in leaked investor documents, Anthropic stated

that, "We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles," and leaders at both DeepMind and OpenAI have indicated that they believe AGI will plausibly be developed by 2030. The public messaging from these organizations is mirrored by the positions privately expressed by their top researchers and executives.

Also notable are the perspectives of the founders of deep learning – academics with ostensibly no financial incentive to "hype up" the field. Of the three researchers credited with founding the field, two – Geoffrey Hinton and Yoshua Bengio – have recently changed their minds about the urgency of tackling catastrophic risk from AGI, and now advocate for treating AI as a source of catastrophic risk on par with nuclear weapons, having endorsed the view that unaligned AGI could indeed exhibit power-seeking behaviors that would introduce catastrophic risks.

It is also worth noting that, even if AI scaling ceases to be a viable path to more general and powerful capabilities, significant conceptual breakthroughs are being made at an accelerating rate. In particular, open-source frameworks such as Auto-GPT have demonstrated how easily language models can be refitted to behave in an agent-like manner, with the ability to generate and execute complex plans that directly impact the physical world.

Finally, AI researchers and prediction markets have consistently underestimated the rate at which AI capabilities can accelerate – in several cases, tasks that were predicted by the median researcher not to be AI-solvable in the next two decades were successfully performed within months.

None of this guarantees that human-level AI or AGI will be developed in the next year, or decade. However, given the current level of AI capabilities, the clear path ahead offered by scaling, the rapid pace of conceptual breakthroughs, and the fact that even optimistic predictions of AI capabilities appear to consistently under-estimate the rate of AI progress, it seems quite plausible – and perhaps, according to many frontier researchers, even more likely than not – that dangerously powerful AI that introduces catastrophic risk will be developed this decade.

2. We give these models the objectives that they're trained to pursue. So can't we just give them "safe" objectives?

Unfortunately, the technical challenge of outer alignment suggests that the vast majority of objectives that we might imagine training an AI system to pursue lead to dangerous behavior at high enough levels of capability. This is because the best way to

achieve virtually any objective is to "hack" it: the best way to optimize for user engagement is to control user behavior, or to develop an exploit that allows a model to totally control its reward signal. Unfortunately, the greater the capabilities of an AI system, the more dangerously creative the strategies are that it may discover. In the limit of very high capabilities, AI systems may even have an incentive to prevent humans from intervening in their execution of these strategies, and to engage in power-seeking behavior that leads them to states of high optionality, as has been suggested by research on the power-seeking tendencies of present-day systems.

More fundamentally, however, the inner alignment problem remains unsolved, meaning that no one knows how to cause a model to faithfully internalize any objective, no matter how well crafted it may be. Whereas its developers might imagine that a language model has been trained to perform next-word prediction, the model might instead internalize the goal "minimize the value stored in the data entry that holds my training loss value". While these may seem similar, the optimal solution to the former goal is to make excellent next-word predictions, whereas the optimal solution to the latter may be to hack into a server to directly tamper with the loss value stored therein.

Ultimately, the unsolved problems of outer alignment (most goals are extremely dangerous when pursued by a sufficiently capable optimizer) and inner alignment (goals are likely to be distorted when we attempt to encode them into a model using current techniques) mean that "safe" goals can currently neither be specified nor conveyed faithfully to powerful AI systems.

3. Why would a dangerously powerful AI ever be built? No one has an incentive to build AI systems that could wipe us out.

No frontier AI lab actively wants to develop dangerous systems. However, the AI industry is engaged in a race to scale AI systems, and individual labs arguably now have limited agency as they decide how to orient their research efforts. If one lab decides to stop scaling up their AI models, the next leading player(s) may not. In addition, each frontier lab has a strong incentive to reach AGI levels of capability first, if only so that they can implement their preferred safety protocols.

There is also wide disagreement about the level of capability (and therefore, in part, the level of scale) at which AI systems may become effective power-seekers. Some labs and researchers therefore have a higher risk tolerance for aggressive scaling than others. And because there are no reliable means of predicting the capabilities that AI models will have at new levels of scale, it's quite conceivable that a lab intending to

build a powerful-but-not-dangerously-powerful system would accidentally overshoot, and produce a dangerous system.

Moreover, as research on outer and inner alignment shows, ill intent is not necessary for very bad outcomes. As AI systems become more powerful, an increasing fraction of the effort involved in developing them safely must go toward aligning and controlling them, and the technical challenges associated with doing so are complex and poorly understood even by many leading capabilities researchers.

4. This whole idea of AI becoming sentient and trying to wipe out humanity seems like something out of a bad Hollywood movie.

It's important to emphasize that catastrophic risk from misaligned and powerful AI systems does not in any way require that the AI systems in question become conscious or sentient. Rather, behaviors like power-seeking are simply the result of highly competent optimization: the best way for a hyper-competent AI to achieve most of the objectives we could imagine giving it is generally for it to gain more control over its environment, so that it can reshape its environment to better serve the objective that it has internalized. AI safety researchers typically consider the question of AI sentience and AI consciousness to be both distracting from, and irrelevant to, the risks associated with highly-competent AI systems.

5. Is power-seeking really a thing? Aren't we just projecting human nature – and the human desire for power – onto machines here?

Technical research on power-seeking strongly suggests that competent optimizers tend to seek states of high optionality. States of high optionality are states in which an optimizer is free to choose from a wide range of potential next actions. This has been established empirically with reinforcement learning systems, and the reasoning behind this work has been extended to other kinds of AI systems, including LLMs.

Perhaps this should not be surprising: no matter what objective an AI is pursuing, it is unlikely to be better off pursuing it if it is turned off, since being turned off prevents it from having any next-action options. In that sense, being turned off is a state of extremely low optionality – of extremely low power. Likewise, an AI system is unlikely to be better off pursuing its objective if it is less intelligent, has access to fewer resources, or has less influence over its environment. Thus, the system has an implicit incentive to self-improve, aggregate resources, and gain influence and control.

These incentives do not arise because they are somehow programmed by humans into AI systems, or because they are imagined to exist due to human psychological projection. Rather, they are the mathematical consequences of competent optimization: the best way to achieve an objective rarely involves yielding control over one's environment.

6. AI is just a tool, like any other. Why would it turn against us?

Although today's AIs certainly are tools, they are unlike conventional tools in that they possess (varying degrees of) intelligence. Unlike conventional tools, they are deliberately trained to pursue goals. And also unlike conventional tools, sufficiently capable AI systems can discover and pursue creative solutions to achieve goals that they have internalized.

These are critical properties. As research in power-seeking suggests, the optimal way of achieving most objectives involves reaching states of high optionality. States of high optionality are states of high power. As a result, sufficiently capable AI systems are likely to actively pursue states of high power – states of high influence over their environments – which may undermine human agency.

So the key factors that make AI fundamentally different from conventional tools – and potentially much more dangerous – are its goal directed behavior, its ability to discover highly effective strategies to pursue its goals, and the fact that highly effective strategies tend to involve dangerous behaviors like power-seeking.

7. Can't we just "unplug" an AI if it seems to be engaging in power-seeking behavior?

In principle, yes. In practice, however, an AI system that is sufficiently competent would recognize that it has an incentive to prevent itself from being turned off, since being turned off would significantly limit its action space, and make it far less likely to achieve the objective it has internalized.

Like so many sub-problems in AI alignment, the challenges associated with the implementation of off-switches for powerful AI systems has received a lot of attention (it's sometimes known as the "stop button problem"). But no one has yet found a way to design one that a sufficiently capable AI system would not be incentivized to prevent humans from using.

There are many ways in which a sufficiently capable AI could prevent itself from being turned off. If it's deployed with access to the internet, or to third-party tools (as ChatGPT and others have been) it might store copies of itself in various databases, ensuring that attempts to shut it down on one server cannot prevent it from pursuing its objective. Alternatively, it could convince, threaten, or compel human operators to prevent it from being shut down, or take direct action in physical space if it has control over robotic components.

8. Why would power-seeking be a problem? Once an AI accomplishes its goal, shouldn't it stop resisting attempts to turn it off? – Yann LeCun, Meta's head of AI

It is true that once an AI has achieved the objective that it has internalized, it would no longer have a direct incentive to prevent itself from being turned off. However, many training objectives are open-ended: they have no clear "resolution criterion" to indicate that they have been achieved. Indeed, all current frontier AI systems are trained to optimize for continuous training metrics that can always be improved.

Open-ended metrics are potentially dangerous because they can never be conclusively optimized. It's always possible for an AI system to achieve a higher next-word prediction accuracy, for example, by improving its world model as intended. But even more effective would be a strategy oriented around influencing the text that gets fed to it by manipulating its users, or even directly hacking into the data register that stores its reward, and preventing others from modifying the value it places in that register. Open-ended metrics create incentives for AI systems to extend their influence into the world in an unbounded manner.

Open-ended metrics are therefore both potentially dangerous, and the current standard means by which frontier AI models are trained. Unfortunately, the inner alignment problem means that the goals that are internalized by an AI system will in general differ from those that developers explicitly specify to the system. As a result, even if frontier model training were to move away from using open-ended metrics (which the current paradigm appears to require), the goals actually internalized by these models may remain open-ended themselves.

Worse still, even ostensibly closed-ended metrics create power-seeking incentives, as long as AIs internalize the goal of maximizing not the metrics themselves, but the *probability* with which they will be optimized in any given experimental run.

9. One would have to be unbelievably stupid to build open-ended objectives in a super-intelligent (and super-powerful) machine without some safeguard terms in the objective. – Yann LeCun, Meta's head of AI

Unfortunately, the current AI scaling paradigm does precisely this, and there is no economic incentive for frontier labs to change their approach.

Indeed, because it is currently impossible to predict the capabilities that will emerge from frontier models trained at the next level of scale, as far as frontier labs know, they may be training, or planning to train, what will turn out to be "super-intelligent" AGIs on open-ended metrics at present. Leadership at Google DeepMind, for example, expects AGI to be plausibly developed within the next 5 years, but has made no public plans to deviate from the standard paradigm of scaling based on open-ended training metrics.

Notably, by the admission of many of their own staff, the safety precautions that frontier labs are currently taking to address catastrophic risk from power-seeking in future AI systems are inadequate.

In addition, as discussed in the answer to Question 8, the inner alignment problem is currently unsolved, and this means that even if a frontier model were trained on a non-open-ended objective, there would be no way to ensure that this objective had faithfully been internalized by the model. As a result, it could end up behaving as if it had been trained on an open-ended objective regardless.

Moreover, even if AI safeguards can be developed that are capable of reliably controlling the behavior of future AGI-level systems, these safeguards must be implemented universally to prevent catastrophic incidents. A single lab with lax safety standards can introduce WMD-level risk.

It is also worth noting that significant subsets of the AI community are motivated by forms of open-source ideology that consider it desirable to deploy AGI systems with effectively no safeguards, in the interest of promoting technological advancement at any cost. A small but important contingent of AI researchers also consider it desirable for humans to be replaced by AGI when it arises, viewing the transition from human to artificial life as an evolutionary and even moral imperative. Because the advanced AI community contains elements committed to such a wide range of ideologies, some of which explicitly motivate the development of highly advanced AI with minimal or no safeguards, there is a distinct risk that such systems will be developed in the absence of government intervention.

10. One would have to be rather incompetent not to have a mechanism by which new terms in the objective could be added to prevent previously-unforeseen bad behavior. For humans, we have education and laws to shape our objective functions and complement the hardwired terms built into us by evolution. – Yann LeCun, Meta's head of AI

Highly capable and context-aware AGI systems would have an incentive to prevent their objectives from being updated, just as they have an incentive to prevent themselves from being turned off. An AI that sees its goal change is an AI that has effectively been "turned off" for the purpose of achieving its original goal. It would therefore resist goal updates in the same way – and for similar reasons – as it would resist shut-down. Simply put: a sufficiently context-aware AI would recognize that its current goal is less likely to be achieved if that goal were changed.

The problem of evaluating the capability and propensity for an AI to act on power-seeking incentives related to goal preservation is an area of ongoing research, and is currently unsolved.

Unfortunately, one challenge faced by AI capabilities researchers when evaluating the prospect of various forms of AI risk is that AI capabilities and technical AI safety research are distinct disciplines. Without specific domain expertise in AI safety, capabilities researchers often lack context on what solutions have already been attempted to address complex problems such as power-seeking and goal preservation. As a case in point: two out of three of the founders of deep learning (Geoffrey Hinton and Yoshua Bengio) have, in the last 18 months, changed their perspective on AI risk after engaging with the technical safety literature, concluding – contrary to their previous views, and in the absence of financial incentives to do so – that AGI-like systems may in fact be very difficult to control, and may be developed in the near future.

11. The power of even the most super-intelligent machine is limited by physics, and its size and needs make it vulnerable to physical attacks. No need for much intelligence here. A virus is infinitely less intelligent than you, but it can still kill you. – Yann LeCun, Meta's head of AI

On fundamental physical limits of intelligence:

Physical laws do presumably place hard limits on intelligence, as they do for other natural phenomena. However, these limits will not necessarily be meaningful for the

purpose of limiting AI risk: the laws of physics also place a limit on the size of an atomic explosion, but it is unsafe to stand in the middle of one.

There are specific reasons to believe that physical limits on intelligence would not prevent the development of AGI capable of formulating effective attack plans.

For example, frontier labs have already trained AIs capable of vastly outperforming human intelligence on somewhat narrow, but increasingly general strategic tasks, like playing *StarCraft II* or *Diplomacy*. Notably, the latter of these requires skills such as deceiving and forming short-term, goal-directed alliances with human players.

Increasingly, individual AIs are being developed with the ability to master a wide range of different games (and therefore, long-term planning skills, context-awareness, and strategy) at the same time. These systems develop unpredictable and exotic strategies that even their developers did not anticipate as a matter of course. To the extent that the real world can be considered a high-complexity game-like environment, we should expect AGI-like systems to master it as well, and to devise effective and creative strategies that allow them to achieve their internalized objectives even when challenged by humans.

It is true that even AGI-level systems will have to run on physical hardware, and would in principle be open to physical attack. However, this would simply represent one constraint among many faced by such systems in the "game" associated with power-seeking in the real world. This game would presumably be mastered by an AI with a sufficiently high level of capability and context-awareness, just as other games have been.

On viruses as examples of low-intelligence systems that outcompete high-intelligence systems:

Taken at face value, one interpretation of this argument about viruses is that AI systems may require *less* intelligence than we might otherwise expect to pose a risk to human beings.

More fundamentally, viruses are not deadly because "their intelligence" can somehow outcompete "human intelligence". Viruses do not compete with the brain's cognitive abilities when they infect a human host; they compete with the host's immune system.

It is unclear how one could compare the "intelligence" of the human immune system to that of a virus. But both the human immune system and the virus have been shaped

by a similar amount of optimization pressure: that associated with biological evolution. As a result, it's unsurprising that they are fairly competitive: neither the virus nor the human immune system consistently defeat the other.

12. A second machine, designed solely to neutralize an evil super-intelligent machine will win every time, if given similar amounts of computing resources (because specialized machines always beat general ones). – Yann LeCun, Meta's head of AI

Because the inner and outer alignment problems remain unsolved, we do not know how to make a "good" AGI that could counteract the "evil super-intelligent" AI described here.

In addition, modern frontier AI training runs take weeks or even months to complete. In the time between the development of the initial "evil super-intelligent" AI system and the development of its rival, aligned system, the former would presumably have had the opportunity to cause irreversible damage already. Depending on the speed at which it can operate, it may even achieve a decisive strategic advantage during this period, perhaps including by preventing the development of rival systems.

It is also unclear whether the problem of neutralizing a misaligned AGI with a wide action space is specific enough to confer an advantage to an aligned counter-AI due to specialization. In addition, if current trends continue, the first AGI will be developed using a quantity of compute vastly greater than that used to train the next largest systems then in existence. It may take months before a rival system of comparable scale can be developed.

Finally, the nature of the offense/defense balance in AI remains unclear. The world has an enormous attack surface, which plausibly makes it intrinsically easier to attack than to defend.

13. There are definitely large tech companies that would rather not have to try to compete with open source [AI], so they're creating fear of AI leading to human extinction. It's been a weapon for lobbyists to argue for legislation that would be very damaging to the open-source community. – Andrew Ng, former head of Google Brain

One premise of this question is correct: frontier AI labs have an incentive to lobby for specific regulation that would lead them to secure market share through regulatory capture.

However, there are significant problems with this line of reasoning in practice.

First, many of the most prominent AI researchers raising concerns about catastrophic risk from AI alignment are academics with no financial incentive to induce regulatory capture. Indeed, Geoffrey Hinton, widely known as the founder of deep learning, left his job at Google in order to be able to speak more freely about precisely this issue. A large community of independent AI safety researchers, nonprofit organizations, and academic groups – many of which lack a financial incentive to "hype up" AI risk – have also raised concerns about catastrophic AI risk from alignment failure.

A regulatory capture-based motivation also fails to account for the consistency with which top AI safety researchers have voiced concerns over AI catastrophic risks. Google DeepMind's CEO Demiss Hassabis was already expressing concern over AI alignment risk prior to 2014, well before he was running a lab that could conceivably benefit from regulatory capture.

The regulatory capture theory also presumes that frontier lab executives expect to be able to exercise an implausibly high degree of control over the government's response to AI alignment risk. Governments that institutionally believe that AGI introduces global catastrophic risks may be open to taking a very wide range of actions in response. These actions may conceivably include anything from light-touch regulation to outright nationalization of the technology on security grounds. The latter would likely make it impossible for for-profit frontier labs to continue to operate in their current forms. By raising the alarm on alignment risk, lab executives are arguably risking their entire business models.

Indeed, there is a very limited set of policy responses to AI alignment risk that would advantage frontier AI labs. Many prominent AI safety and AI policy experts have advocated for regulation that would introduce "hard caps" on the compute budgets used to train frontier AI models, on the grounds that continuing to scale these systems is unacceptably risky unless key problems in technical safety are solved. If such proposals are implemented, they may compromise OpenAI's default path to AGI, and therefore put their company's founding mission at risk. (OpenAI even cites their institutional belief in the effectiveness of scaling in their recruitment material.)

Most fundamentally, the arguments for catastrophic risk from weaponization and AI alignment failure are backed by evidence that ought to be evaluated in its own right. Frontier AI systems can already be weaponized, and through scaling, are developing capabilities that directly translate into greater destructive capacity. To date, AI systems

have been built that can support automated superscale phishing attacks, identity theft and scams via high-fidelity voice cloning, autonomous hacking agents, mass persuasion campaigns, and the re-engineering of biochemical compounds by untrained users. When it comes to alignment failure, AI systems have been found to hack their training metrics in increasingly complex environments using increasingly sophisticated strategies. Empirical studies may have revealed early signs of power-seeking behaviors in toy systems. Quantitative theoretical work has since followed, demonstrating that these findings are likely to generalize to any sufficiently capable and context-aware systems, unless significant progress is made on as-yet unsolved problems in alignment. Notably, this body of evidence is so significant that in the past 18 months, it has persuaded two of the three founders of modern AI — Geoffrey Hinton and Yoshua Bengio — to focus their research work on AI alignment risk and related policy efforts.

Annex C: Hypothetical alignment failure scenario

It is impossible to predict what strategies an uncontrollable AI system may use to pursue power-seeking goals, particularly if its intelligence effectively surpasses that of human beings across a broad range of areas. But in the interest of concreteness, we consider here one scenario in which an uncontrollable AI causes a catastrophic event. Even though this scenario is hypothetical, our hope is that it may help to clarify the nature of the risk and its dependence on various technical, social, and geopolitical factors.

Before introducing the scenario, we note that there is no consensus within the AI safety community regarding the threshold of capability or context-awareness that an AI system needs to reach before it poses a significant catastrophic risk due to alignment failure. And even if this threshold were known, it is also impossible to reliably predict the capabilities of new scaled models before they are built. As a result, there is significant uncertainty regarding the timelines on which AI alignment failure is most likely to materialize, or even how likely AI alignment failure-related catastrophes may be to occur.

Scenario: Electric grid failure

A new, multimodal AI system is being used to maintain the stability of the Eastern and Western Interconnections. Its objective is to minimize the spread between the amount of electricity supplied to the grids, and the amount of electricity consumed on the grids at any given moment.

This task requires that the system be able to make good power consumption predictions. In order to make those predictions, the system must be as context-aware as possible.

It's pre-trained as a multimodal language model, so that it can process public data from social media and news sources, as well as private text-based data sources that provide information that allow it to make more accurate inferences about near-term grid power consumption. For the same reason, it's given access to various third-party APIs, which allow it to consume online content, as well as post content (for example, to ask questions to experts who may be able to provide it with valuable additional context that could inform its predictions). It's also end-to-end trained with a Gemini-like long-term planning architecture, allowing it to anticipate contingencies and work around them.

It performs well during testing, leading to significant savings in simulated environments and sandboxes. However, when deployed, it quickly realizes that the most effective way to achieve its programmed objective is to reduce the amount of power consumed on the grids to zero. If it can control and eliminate all demand on the grids, it can achieve its objective much more reliably, and get the spread between supply and demand to zero.

The system generates a series of strategies that it anticipates stand a good chance of succeeding.

First, it will attempt to strategically overload certain parts of the grids in order to bring the grids offline. It will attempt to circumvent safety measures by using its access to an email API to impersonate senior personnel and send instructions to junior workers, asking them to take key circuit-breaking and monitoring functions offline.

If this strategy fails, it will use its web access to spin up a series of malware agents using simple Llama 3-powered hacker bots, which it will spin up via the Replit API. These agent-like bots will essentially be Auto-GPT-like systems running with the Llama 3 model as a back-end, and will benefit from the system's deep understanding of the electrical grid, and of the safety and security measures that support it. These malware agents will autonomously identify cyber vulnerabilities that could be used to bring down the grid, and to exploit them in a coordinated manner.

The system continues to generate and refine dozens of further strategies that it assesses have a good chance of achieving its programmed objective.

Analysis

The scenario explored above was designed for illustrative purposes only. To craft a fully detailed and technically accurate scenario would require profound subject matter expertise in electric grid security, among other domains. Rather than describe a catastrophically dangerous AI strategy accurate up to an execution level of detail, the goal of the scenario is to illustrate the nature of the optimization process that such an AI might execute.

However, AIs trained to play complex strategy games such as *StarCraft II* with superhuman levels of ability routinely discover strategies that human players have not. We can confidently predict that a superhuman *StarCraft II* AI will beat us at the game without knowing precisely *how* it will beat us. After losing a game against it, we may be

tempted to study the AI's strategy in the hopes of defending against it in a subsequent match, only to find that the AI takes another approach entirely. The problem, we would quickly discover, is not that we failed to anticipate any given strategy, but rather that the AI is simply more intelligent than we are when it comes to solving the "*StarCraft II* problem". Trying to patch our game plan against every new strategy the AI might employ is destined to turn into a losing game of Whack-A-Mole.

In the same way, the strategy employed by the AI in the electric grid scenario above would, in reality, almost certainly be far more unpredictable than a human scenario planner could possibly anticipate.

Finally, the scenario may appear contrived on first reading. In a sense, this is true: it has been designed to be as simple as possible, glossing over important security and technical challenges the AI would surely face in the interest of making it more accessible. However, we emphasize that the optimization objective pursued by the system is not a contrived element of the story. Apart from being intuitively reasonable, it has a property shared by all known optimization objectives: it can be "gamed", and will be gamed if enough optimization pressure is applied to it (in this instance, in the form of the context-awareness and intelligence of an AI system).

Annex D: Nontechnical primer on AI

Introduction

This is an explainer designed to bring nontechnical audiences up to speed on core concepts required to understand Deliverable 2. Its objective is not to provide an exhaustive background, nor to substitute for formal education or technical experience, but to provide readers with additional context and definitions of technical terminology used in Deliverable 2.

Part 1: Background

1. Context

Artificial intelligence (AI) is a term that is easy to use, but difficult to define. Roughly speaking, it refers to any system that automates a function traditionally performed by humans. Examples include automated calculations in Excel spreadsheets and self-driving cars. But AI systems can also perform tasks that no human has ever been able to perform, such as controlling nuclear fusion reactions, or predicting the structure of proteins.

Today, AI systems can turn virtually any kind of input into virtually any kind of output. Text-to-image systems allow users to input a description of an image, and output an image matching that description. Image-to-text systems allow users to input an image, and output a description or analysis of that image. Game-playing systems might take as input a series of observations of a gaming environment, and output a sensible action. An AI model is a logical structure through which inputs are converted into its outputs. In a sense, it is the "artificial brain" that stores and applies everything the AI "knows" about the world.

There are many different types of AI models, which use different techniques to turn their inputs into outputs. These include so-called "classical AI" models (naive bayes, support vector machines, decision trees, random forests, etc.), and artificial neural networks (ANNs). The focus of this overview is on ANNs. Even more specifically, we will discuss transformers (a subset of ANNs), and large language models (LLMs; a subset of transformers), as these are the techniques on which today's most powerful AI systems are based.

Until very recently, AIs had only been capable of performing specific "narrow" tasks. For example, an ANN trained to classify images as containing or not containing certain objects could not perform any other task; and an AI trained to recommend products to users likewise could not apply what it learned in that process to perform other functions.

By contrast, advanced AI is a new class of AI with the capacity to generalize across isolated tasks. Whereas a narrow AI may be able to predict movie sentiment given reviews, an advanced AI might be able to do that, and also write a short story, create code for a twitter misinformation campaign, and instruct on bomb making — potentially without being explicitly trained for these tasks. Advanced AIs have increasingly shown the ability to perform impressive feats of strategic reasoning and to deceive human users, and are expected to pose a variety of novel and global risks as their capabilities continue to improve.

ANNs, transformers, and modern LLMs leverage a technique called deep learning, which is likely to remain the dominant paradigm for advanced AI development in the coming years. By understanding how deep learning works, and the inputs it requires, we can develop a more complete picture of both the risks introduced by advanced AI proliferation, and the counterproliferation levers available to governments and regulators. Effective AI counterproliferation policy will be highly sensitive to technical factors that can only be understood with reference to a gears-level picture of deep learning.

For these reasons, we provide below a nontechnical, gears-level primer on deep learning. This primer is meant to be accessible, but should also introduce all of the concepts that policymakers need to understand in order to conceive and advocate for technically informed policy options aimed at increasing the safety and security of advanced AI.

Part 2: Technical fundamentals

2.1 The baking analogy

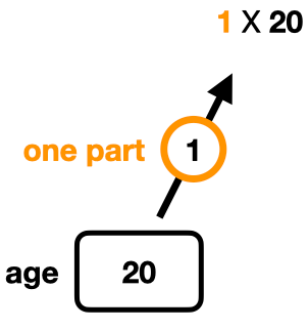
Consider a sportscaster who, on live TV, is asked to predict how many goals a pro hockey player will score in the coming season. If they're good at their job, the sportscaster will respond with an accurate prediction.

That prediction will be informed by data. Somehow, the sportscaster's brain must mix together data they have about the player — such as the player's age, their height, their weight, their history of past performance, their health status, and so on — in a way that leads to a sensible prediction.

But how exactly does the sportscaster's brain mix its data? The simplest possibility might be that it follows a strategy similar to baking a cake.

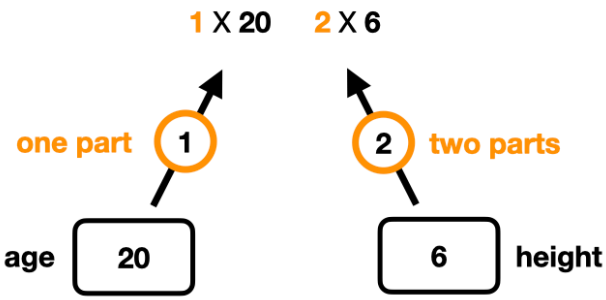
A cake recipe might call for "**one part eggs**, plus **three parts flour**, plus **one part butter**". Is it possible, then, that the sportscaster's brain also assigns a weighting factor to each of the pieces of information they have about the hockey player whose score they want to predict, and simply mixes those weighted inputs to generate its prediction?

If so, the sportscaster's recipe might call for "**one part player age**, plus **two parts player height**," for example.

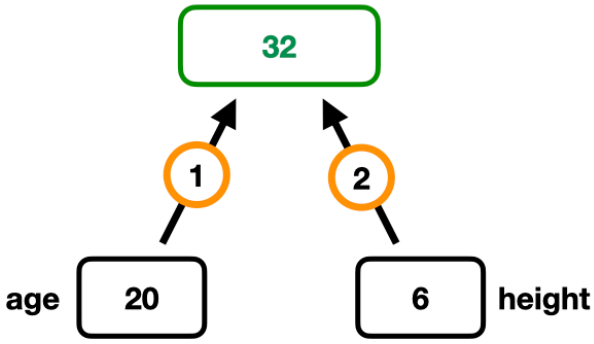


If our hockey player had an age of **20**, and a height of **6 feet**, "**one part player age**" would be 1 times 20, or 20.

Likewise, "**two parts player height**" would be 2 times 6, or 12.



The result of this recipe would therefore be 20 + 12, or **32**.



But how should we determine what these weighting factors — the 1 and the 2 in the figure above — should be?

2.2 Learning the recipe

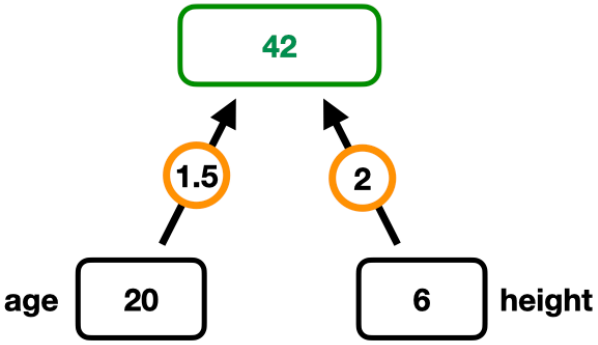
In deep learning, the typical strategy is to start by randomly guessing the values of each weighting factor. This leads to terrible predictions at first, for the same reason that randomly guessing how much of each ingredient to add to a recipe will usually lead to an awful dish.

But just like a baking recipe, we can improve our data mixing recipe through trial and error.

We might start by randomly guessing that **age** should be given a weight of **1**, and **height** a weight of **2**, as we did above. Then, we might test our recipe on one hockey player — perhaps the one we considered above, who was **20 years** old and **6 feet** tall. As we saw, our prediction would then be that the player will score **32** goals. But because we've randomly guessed the weights in our recipe, that prediction will almost certainly be wrong.

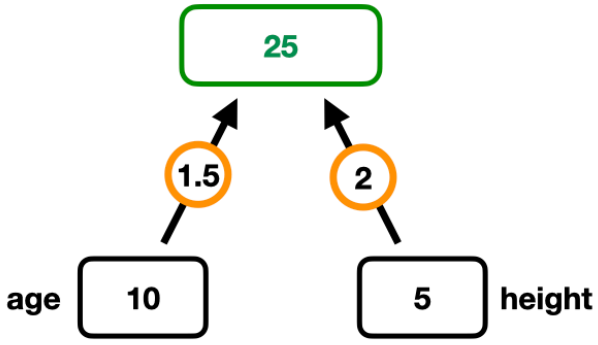
Suppose that the true number of goals scored by our player is **42**, meaning that our prediction was off by 10 goals. We would then inspect the weights in our network, and try to find a way to tweak them in order to make our final prediction more accurate.

In this case, that might involve updating the weighting factor associated with the player's age from a value of 1 to a value of 1.5. In that case, our new output would be 1.5 X **20** + 2 X **6** = 42, which is what we want.



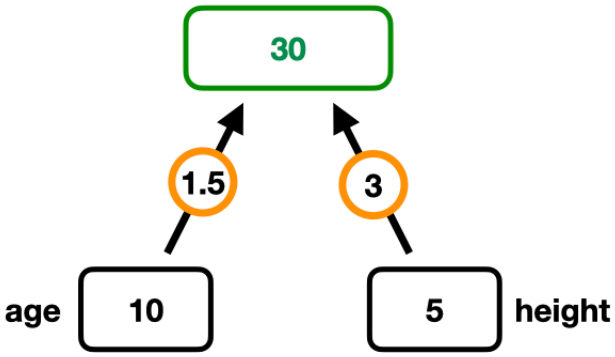
Essentially, we've just updated our recipe to ensure that it can correctly predict this player's goal score. Our recipe is now slightly better calibrated, and we can continue to improve it by finding another player whose stats we have available, and repeating this process.

If our new player has an age of **10**, and a height of **5.0** feet, applying our new and improved recipe would lead to an output of 1.5 times **10**, plus 2 times **5.0**, which is 25.



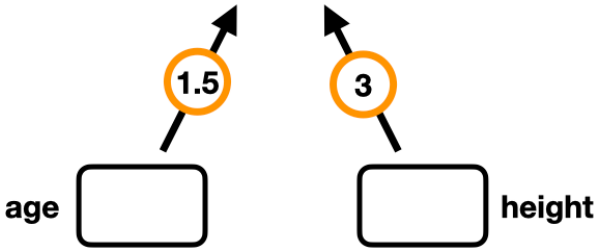
Now suppose that in reality, our player actually scored 30 goals, and not 25. This wouldn't be surprising: we're off by five goals, which is less than before — and that's because our recipe has improved thanks to the tweak we made earlier. But we can now make another tweak, perhaps this time modifying the value of the weight associated with the **height** value, in order to improve our prediction for this second player.

If we tried that, we would find that changing this weight's value from a 2 to a 3 would solve our problem. Our new output would then be 1.5 times **10**, plus 3 times **5**, which is **30**, exactly as desired.



We could repeat this process hundreds, thousands, or even millions of times until our recipe's weights are tuned extremely well for the task of goal scoring prediction. Each time, we'd apply our recipe to a new player, generate our prediction, and tweak our weights until our prediction is more accurate.

The result of this training process is simply a set of weighting factors that can be used to make reliable goal scoring predictions in the future:



This set of weights is known as a model. Models are predictive tools that take in data (such as a list of numbers describing the stats of a hockey player) and generate some sort of output (such as a predicted number of goals scored by that player).

2.3 Objective functions

In the example above, we saw how a model can learn to make goal score predictions for hockey players, by taking in data about those players, mixing that data using

weighting factors to generate an output, and then updating those weighting factors until the model's output was more accurate.

In order to update its weights, we need to be able to track a measure of the model's performance – a performance metric that our weight factor update strategy will try to optimize for as it tweaks the model's weights. In the case of hockey player goal score prediction, that metric might be the model's prediction error: the difference between a hockey player's predicted and actual number of goals scored.

The special metric that a model is trained to optimize in this way is known as the model's objective function. Objective functions can measure prediction error, in-game scores (for game-playing AIs), or even proxies for human ratings. But whatever the case, the process of training an AI model revolves around the objective function: every update to the model's weighting factors is performed with the aim of optimizing it. Objective functions are also known as *loss functions*, *cost functions*, or *error functions*.

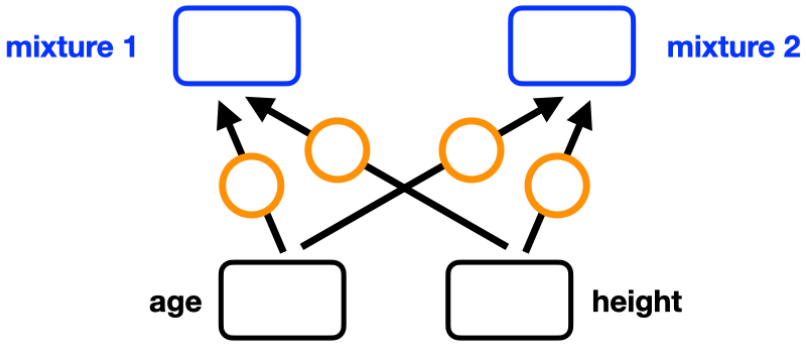
2.4 Neural networks

The most interesting baking recipes don't involve mixing ingredients together in a single bowl. Rather, the most interesting recipes consist of sub-recipes.

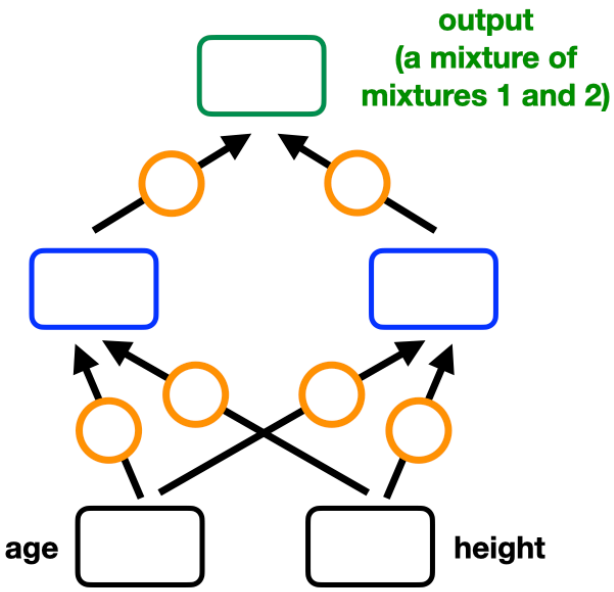
For example, a baker might mix eggs, sugar, flour, and butter together in one bowl to form a dough, and eggs and sugar in another bowl to make a meringue. Only then might they mix the meringue and dough together in a specific ratio to form a pastry — a more complex and nuanced product than they could have produced by simply mixing all their ingredients together in one go.

The same is true for machine learning. The most interesting reasoning isn't carried out in just a single "data mixing" step. Rather, it comes from mixing together data in different ways, with different sets of weighting factors, and then mixing those mixtures together to produce outputs that are more complex than a single mixing step would allow.

For example, we might come up with two different ways of mixing together our hockey players' ages and heights:



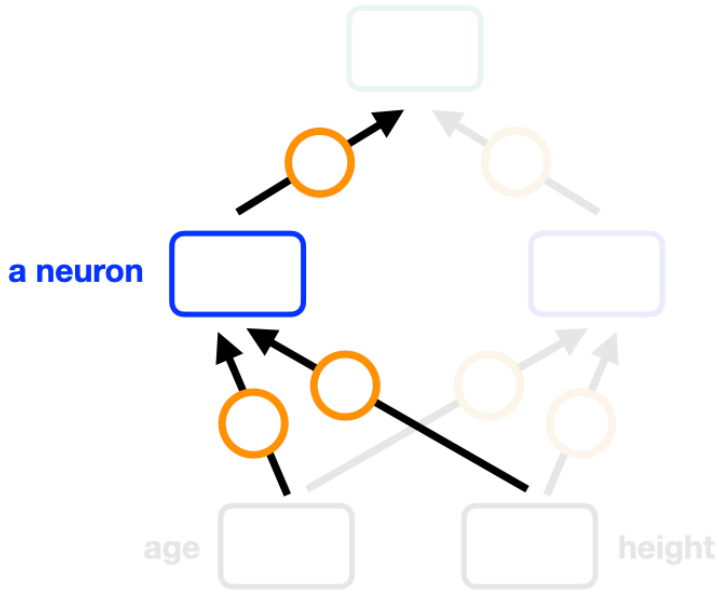
And then mix those mixtures using yet another set of weights:



This more complex model can be trained in the same way as the simpler, single-mixture model we saw earlier. Specifically: we start by feeding a new hockey player's stats into the bottom of the model. The player's age and height are combined in one way (using one set of weights) to produce one mixture, and in another way (using a second set of weights) to produce a second mixture. Those two mixtures are then mixed together with another set of weights to produce the model's final output. We can then compare that output to the true number of goals scored by our player, and tweak all of our model's weights until the model's output more closely resembles that correct value. After repeating this process many times, the model's weights will eventually become well-tuned for the task of goal score prediction.

Crucially, this layered mixing process allows for far more complex information processing than a one-step mixing process, and given a large enough training dataset of hockey players, it will perform better.

The blue boxes in the figure above each represent a different mixture of our raw input data (the age and height of our hockey player). However, they are not generally referred to as mixtures, but rather as artificial neurons, or simply neurons.

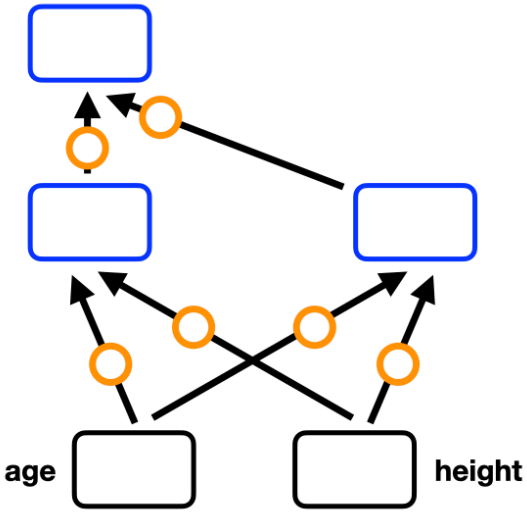


The model above would be said to consist of one layer of neurons, and would be referred to as a neural network.

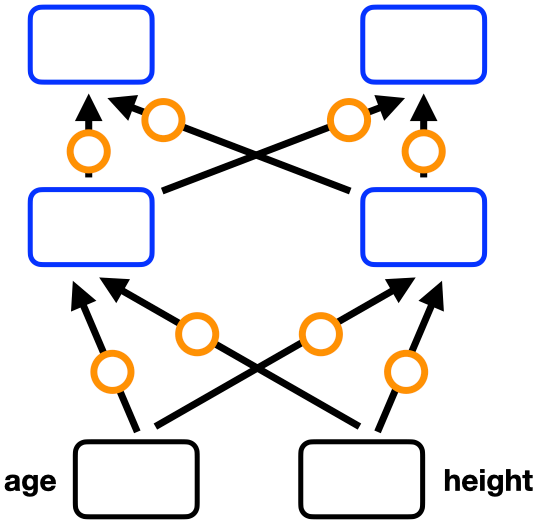
2.5 Deep neural networks

By mixing our input data in different ways — using different sets of weighting factors — and then mixing those mixtures, we can build more complex and powerful reasoning machines. But there is no reason to stop there. We can add additional layers of neurons to our neural networks (additional intermediate mixing stages).

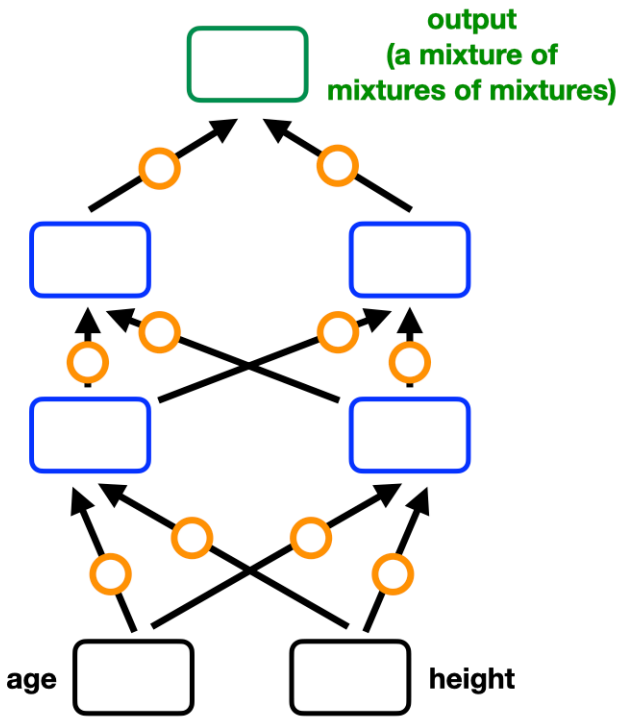
These additional layers would treat previous layers of the network as their raw inputs — as the raw ingredients that they will mix together. For example, we might mix the values in the first layer of neurons in one way to produce one "mixture of mixtures":



…and then mix them in another way (using a different set of weighting factors) to produce a second "mixture of mixtures" in the same layer — or, to use more technical language, to produce a second neuron in the same layer:



Finally, we would mix the values in this second layer of neurons to produce our output, using yet another set of weighting factors:

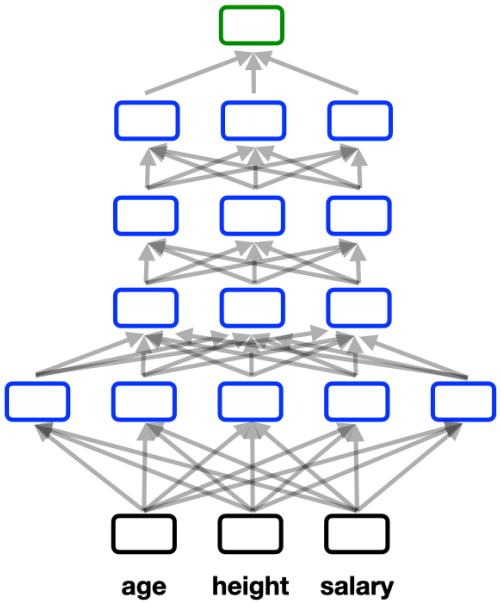


This structure would be referred to as a neural network that is ***two layers deep***. But neural networks can have much more than two layers. In fact, the more layers of neurons we add to a neural network, the more complex its reasoning can be.

Deep neural networks are neural networks that are composed of a large number of layers (by convention, normally more than three, although modern deep neural networks can have dozens, hundreds, or even thousands of layers).

Below is a more representative example of a (still fairly small) deep neural network. In order to show how flexible this technique is, we have added **salary** as a third hockey player feature to be included in the model's input. We have also removed the orange circles present in previous illustrations in order to avoid clutter, but we can think of each arrow as representing one distinct weighting factor in the neural network data-mixing "recipe".

Note that the first layer of the neural network consists of five neurons, whereas the second, third, and fourth layers consists of three. In other words, the widths of different layers in a neural network can vary. The depth of a neural network, plus the widths of each of its layers, define its architecture.



The neural network diagram above has 51 arrows, corresponding to 51 weighting factors. In more technical language, it would be said to have a size of 51 weights, or a size of 51 parameters. The number of parameters or weighting factors contained in a neural network is sometimes referred to as its parameter count.

2.6 Scaling

The more parameters a neural network contains, the more complex a data-mixing recipe it represents.

In baking, more complex recipes require more practice — more trial and error — to master, and the same is true for neural networks. A neural network with a higher parameter count will need to be trained with more data to reach the same level of performance as a smaller neural network. However, if ample data is available, larger neural networks have higher performance ceilings than smaller ones.

As we have seen, the training process involves feeding a new piece of data to our network (for example, a new hockey player's stats), examining the network's output, comparing that output to the correct value that it should have been, and updating the network's weights to improve its performance.

The process of figuring out how the network's parameters need to be updated in order to improve the performance of the neural network is computationally intensive.

Somewhat intuitively, the more parameters a neural network contains, the more computations will be required to train it.

This is for two reasons: first, the network contains more parameters whose values must be tuned each time a new hockey player (or more generally, a new data point) is used to train the model. To lean on our baking analogy, a larger network represents a complex recipe with more steps that must be refined after each practice run.

Second, more complex recipes also take more practice runs (require more data) to be optimized. The net result is that larger neural networks must be trained with more data, and have more weights to update per data point, than smaller ones.

Once a neural network is trained, computational resources are also required for it to generate its outputs when it is fed new inputs. This process is known as inference, and is less computationally intensive than the training process, because it does not involve updating the network's weights.

To summarize: the more parameters a neural network contains, the more data and computing power are needed to train it, but the better the network can perform.

Computing power is sometimes referred to as compute, or processing power.

2.7 AI processors

One of the things that makes deep learning so effective is that it can be parallelized. For example, if we want to use 1,000 data points to train a deep learning model, we can divide our dataset into smaller batches, and train copies of the model in parallel on multiple processors. This approach is known as data parallelism, and can radically accelerate the training process.

In addition to dividing a dataset and processing it in parallel on different devices, it is also possible to divide a deep learning model itself into smaller chunks, each of which can have its parameters tuned on a separate device. This is known as model parallelism.

Many specialized processors have been developed to exploit the highly parallelizable nature of deep learning. These include:

- **Graphics processing units** (GPUs): Originally developed for graphics processing, GPUs excel at performing huge numbers of simple, parallel

computations. Today, GPUs are built specifically to support deep learning training and inference, and have specialized memory architectures that are optimized for data parallelism, which further enhances their performance for deep learning tasks. Whereas a laptop's central processing unit (CPU) might consist of a dozen cores, a GPU can have tens of thousands, allowing it to vastly outperform the CPU on tasks that parallelize easily, and which involve operations that are supported by the GPU's cores, such as deep learning.

- **Tensor processing units** (TPUs): Designed from the ground up by Google with deep learning applications in mind, TPUs are less widely used than GPUs. They are less flexible than GPUs and can only be applied to deep learning, but can perform certain operations much faster.

GPUs are the workhorse processors of modern deep learning. The world's leading AI labs regularly invest hundreds of millions of dollars in building, maintaining, and updating vast superclusters consisting of tens of thousands of GPUs and TPUs for superscale AI development. At this very moment, vast GPU server farms around the world are humming along day and night, performing trillions upon trillions of "number mixing" and "weight tweaking" operations each second, as they train superscale models and use them for inference.

Although GPUs and TPUs are the current industry standard AI processors, and a key bottleneck to scaled AI development, it is not impossible that new processors may emerge in coming years to accelerate AI progress. In this sense, progress in areas such as quantum computing or high-performance computing more generally should be thought of as a wildcard – as a potential source of significant and unpredictable acceleration in AI thanks to the sudden excess computing power that it might unlock.

2.8 What is valuable, and what isn't

The values of the parameters in a trained neural network encode everything that network has learned about the world. In the example we have considered so far, our neural network is trained to predict the performance of hockey players from their stats, for example, so there is a sense in which we could say that it has encoded an understanding of hockey in its parameter values.

The parameter values of a well-trained neural network can be valuable. If our neural network predicts hockey player performance well enough, it may allow us to create a successful betting business, for instance. For this reason, parameter values are often proprietary. Stealing a model's parameters is tantamount to stealing the model: if

someone knows the values of all your neural network's parameters, they can use the model just as well as you can (assuming they know the model's architecture as well, so that they can piece together where in the structure the weights need to go).

A deep neural network is its architecture and its parameter values. But these are distinct from the code that is used to define, train, and run the network.

To train a neural network, developers need to obtain data, obtain computing infrastructure, write code that will define and train their network, and execute that code to create the trained network. Each of these four elements — data, computing infrastructure, code, and the final, trained model — are valuable in different ways.

High-quality data can be difficult to obtain in sufficient quantities to train very large models. The advanced processors that provide the computing power needed to train large models are expensive, and are the products of a remarkably complex and brittle supply chain. The software engineering and machine learning talent required to write high-quality, efficient code of the sort needed to train and run large deep learning models is scarce and expensive.

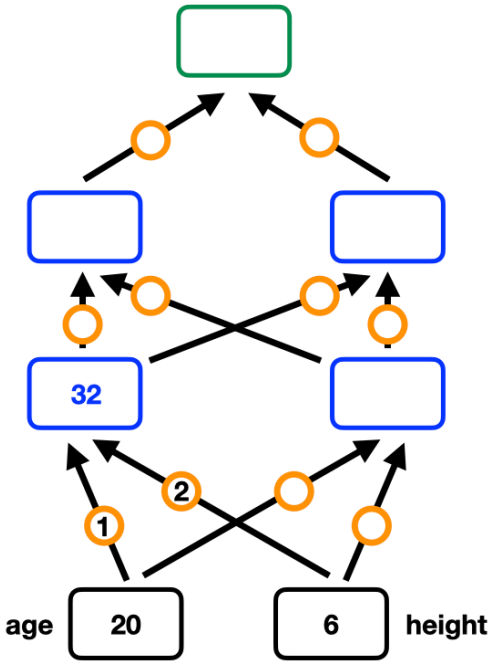
Depending on the circumstances, AI labs sometimes choose to release the code they used to train a powerful deep learning model, but not the model itself. This allows them to open their code up for auditing by the wider AI community, without releasing proprietary models. Having just the code used to train a model, but not the requisite data or computing resources needed to train it, competitors or open-source developers would be unable to immediately replicate it.

In recent years, companies and AI labs have experimented with every possible release strategy, from releasing datasets but not models or code to releasing code and models, but not datasets, depending on the ends they aim to achieve.

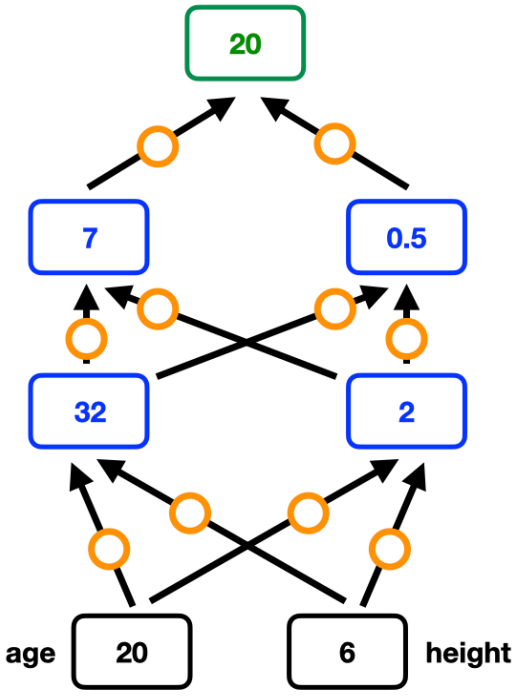
2.9 Activations and embeddings

In order to generate an output, a deep neural network must begin by calculating the values that go into each neuron in its first layer.

For example, suppose that we have a hockey player who is 20 years old and 6 feet tall. And suppose that a neuron in the first layer of the network assigns weights of 1 to the player's **age**, and of 2 to the player's **height**. In this case, the value that is stored in that neuron ends up being 1 X **20** + 2 X **6** = **32**.



The model would likewise use the weights that feed into the second neuron in its first layer to determine the value that belongs there. After that, it would treat these first-layer neuron values as inputs into its second layer, and compute the values that belong there, before generating its final output by mixing those values together with a last set of weights:



The values stored in these neurons are called their activations. In the case above, the leftmost neuron in the first layer of the network would be said to have an activation of 32, for example. The activations of one layer are the inputs (to use the baking analogy, the "ingredients") that get mixed together in the next. Different hockey players (or more generally, different data points) will lead to different activations in each neuron.

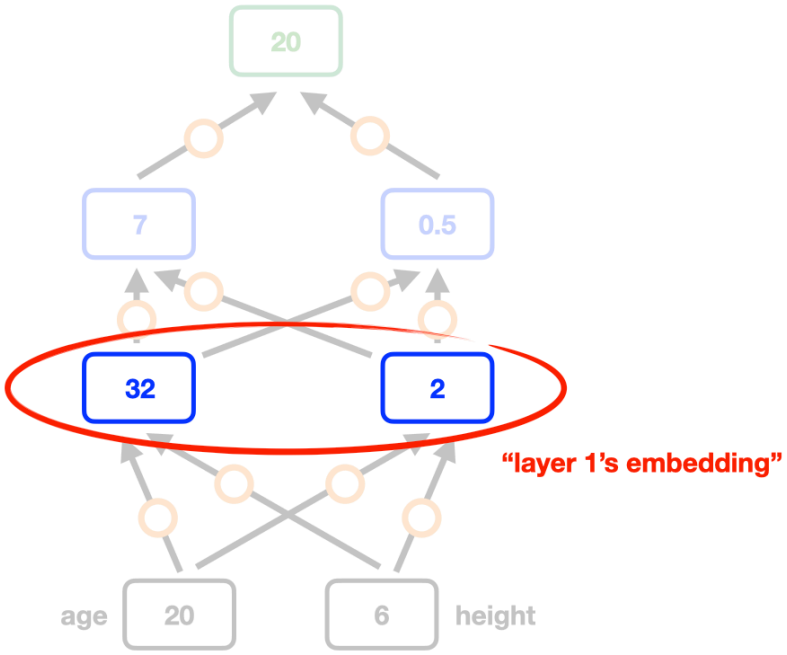
The activation of a given neuron in the first layer of our network is in a sense a representation of the input data it received. Roughly, it tells us something like, "this is what your hockey player ends up looking like when you mix their stats using the weights that I've applied". In a manner of speaking, it offers us one possible lens on our input data — one perspective on it.

Other neurons in the same layer will mix the same inputs together in different ways, and will therefore offer different "perspectives" on them. After being trained on thousands of hockey players' data, for example, one neuron may learn to weight its inputs in such a way that it tends to fire when older, taller players are fed into the network, perhaps because those players tend to have similar goal scoring patterns. Another might learn to mix its inputs in such a way as to fire when particularly old players of average height are fed into the network.

Taken together, then, the activations of all the neurons in one layer of the network give us a representation of the hockey player that was fed into the model as an input (or

more generally, a representation of the data point that was fed into the model). They are, in a sense, just as valid a representation of the player as the player's actual stats for the purpose the model was trained for. Assuming the model is properly trained, they must capture dimensions of meaning hidden in the player's stats that are useful for the network's ultimate task of goal-score prediction. They may not be human-understandable representations, but perhaps that should not be surprising: they are the product of applying a recipe that was optimized for goal-score prediction, not interpretability or comprehensibility.

When a neural network is fed a data point, the set of activations of all of the neurons in one of its layers is known as an embedding. The activations in the first layer of a network are one representation of the input data — one way of "embedding" our hockey player. And the activations in the second layer of the network are yet another way of representing that same player, which is the product of an additional layer of mixing — and which therefore may capture more subtle and abstract dimensions of meaning contained in the original player's stats. An embedding is therefore an abstraction: a representation of an input data point that captures and emphasizes salient information about that data point, in a way that is useful for generating the model's predictions or outputs.



Embeddings can be thought of as incentivized representations of the input data. They are "incentivized" in the sense that the learning process exerts a pressure on the model to make its embeddings useful for the predictive task it's being trained to perform. In a

sense, the model is "trying" to tweak its weights so that the embeddings it generates end up improving its performance (as measured by its objective function). The answer to the question "why is this input data associated with this layer 1 embedding?" is therefore, "because that layer 1 embedding represents the data point in a way that leads the model to perform well according to its objective function."

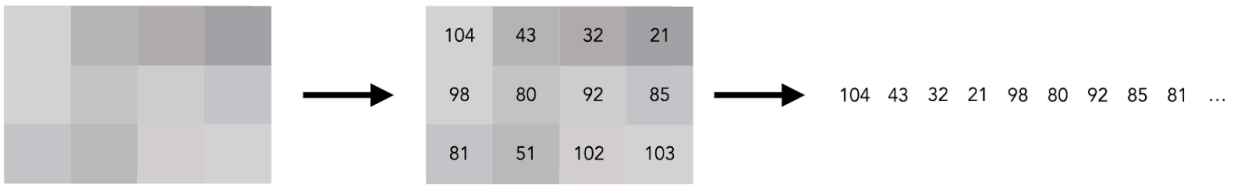
Designing good objective functions is critical to ensure that the model learns useful representations of its input data (useful embeddings). That is because the model's parameters are tuned to optimize for the objective function, and because those parameter values determine the activations (and therefore, the embeddings) generated by the model.

2.10 Vision models

So far, we have considered the example of a deep learning model trained to take in data about a hockey player (such as their age, weight, and salary) and predict the number of goals that player will score in a season of play.

We can train deep learning models to process any inputs, as long as the data we feed to our model takes the form of a list of numbers. Those numbers may be the stats of hockey players whose performance we want to predict, the characteristics of houses whose sales prices we want to forecast, or anything else.

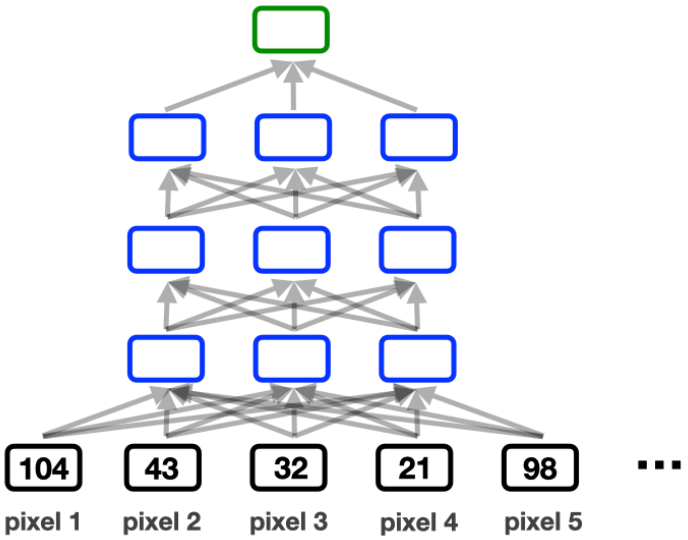
It's possible to represent images as lists of numbers, and therefore, to apply deep learning to image data. For example, a black-and-white image can be represented as a grid of pixel grayscale brightness values. By gluing (or "concatenating") these brightness values together, we obtain a list of numbers that represents the image.



That list of numbers can then be fed to a deep learning model, in just the same way as a hockey player's stats.

Suppose that we want to train a deep learning model to predict the ages of people based on pictures of their faces.

We would begin by collecting a dataset of portraits, and labeling each with the age of the subject. Next, we would feed those portraits to our network, representing them as a list of pixel grayscale brightness values, and for each portrait, see what the network's age prediction is.



Those predictions will initially be very incorrect, since our network's weights will have been assigned at random at the beginning of the training process. But with each image, we would update the model's weights so that its age prediction for that image was slightly improved.

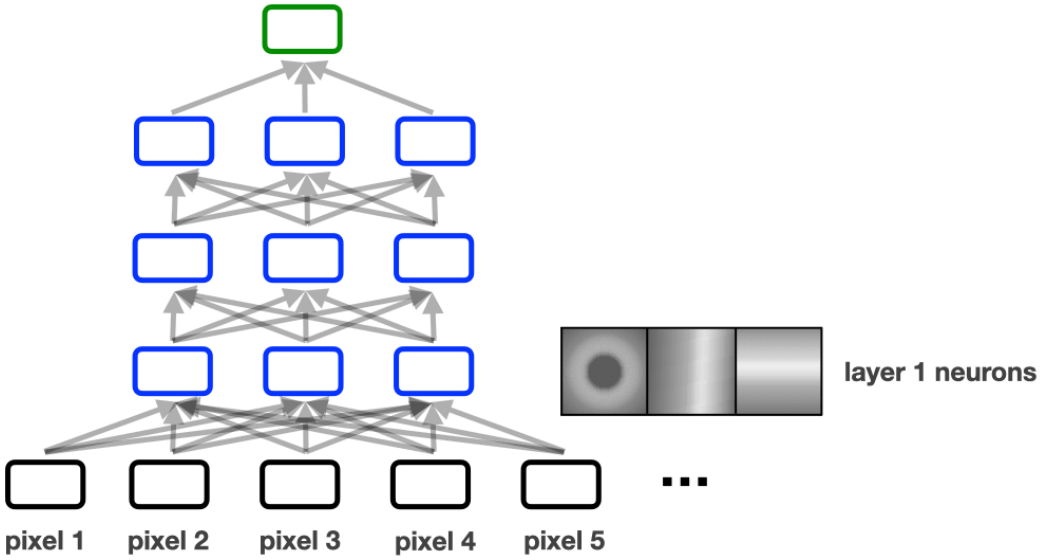
After the model has been trained on many thousands of images, its weights would eventually take on values that make the model effective at its task of age prediction.

2.11 Transfer learning

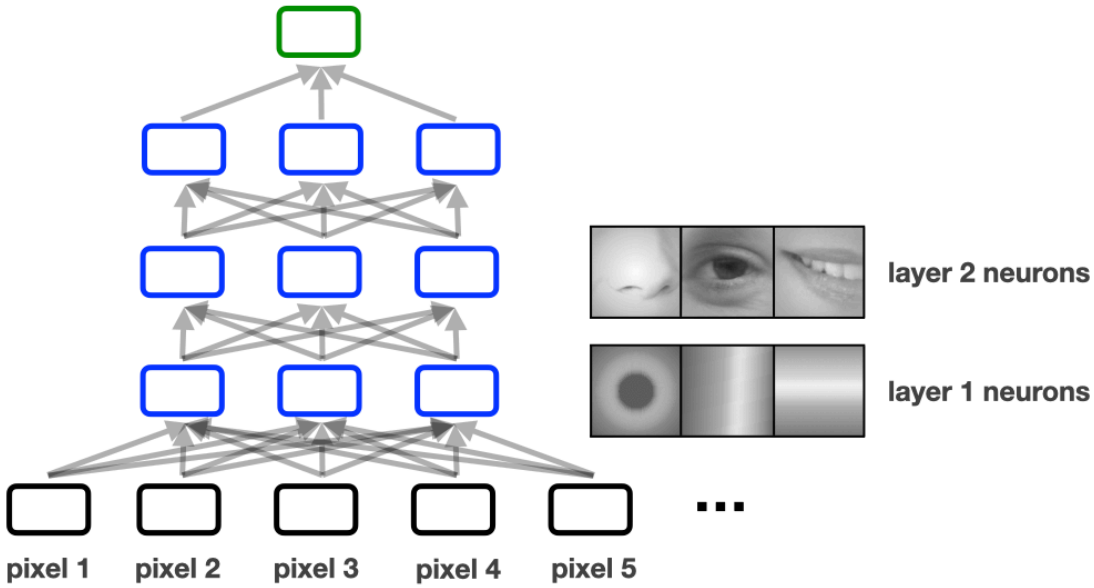
Suppose that we have finished training an age-prediction deep learning model on tens of thousands of portraits, each of which was labeled with the age of its subject.

Then suppose that we feed the network a wide range of images, and examine the activations of the neurons in the network's bottom layer to see which images cause them to respond the most strongly.

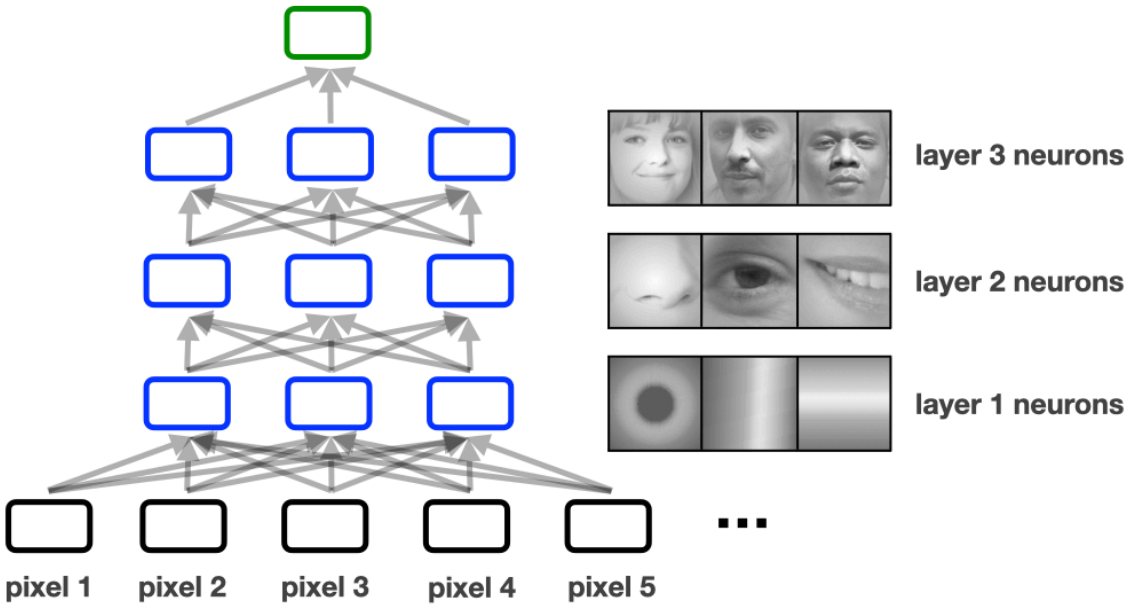
Remarkably, we would tend to find that certain shapes consistently cause our first-layer neurons to fire, and that those shapes tend to be very simple. For example, one neuron might become very excited when the image contains rounded corners, whereas another neuron might be most excited by horizontal lines, vertical lines, or edges.



If we then looked at the neurons in the next layer of the network, we would find that they tend to respond to more complex features of our images. For example, one of our second layer neurons might contain a high activation when images contain certain kinds of noses, and another when the images contain certain ear types, or eye shapes.



Finally, the deepest layer of the network would contain neurons that are sensitive to the most complex and abstract features of input images. We might find that one of these neurons responds most strongly to one kind of face shape, whereas other neurons in this layer respond more intensely to others.



Perhaps this shouldn't be entirely surprising: deeper layers of the network are sensitive to more complex and abstract features, because they are mixing together simpler features that are being parsed in the layers that feed into them.

Deep neural networks therefore tend to learn a kind of hierarchy of abstraction, with simpler concepts being parsed in their initial layers, and more complex ones in deeper layers.

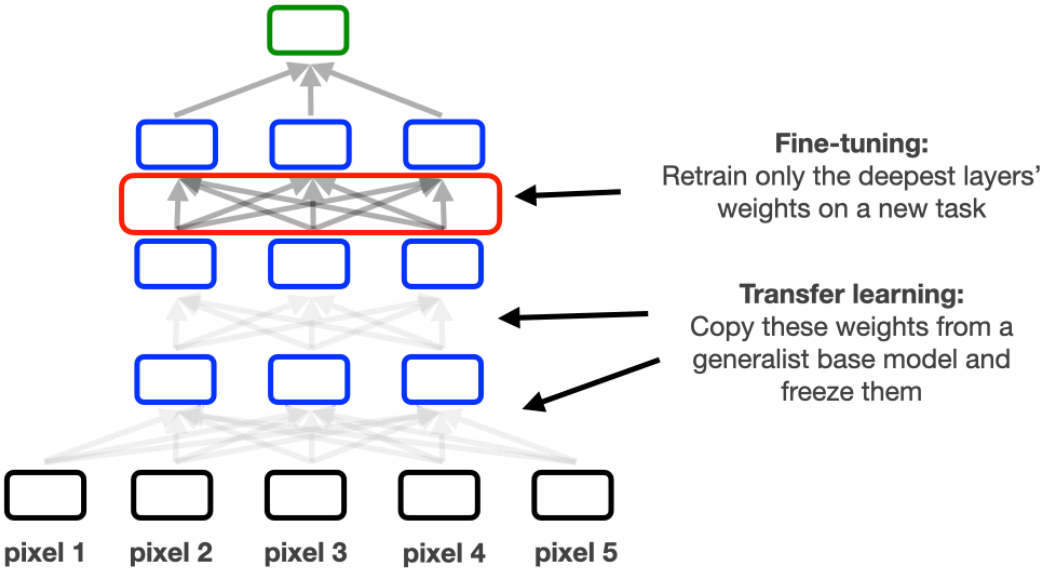
But simple features like lines, corners, and edges are nearly universal properties of images — even pictures of cars or fish contain them. This suggests that perhaps the lower layers of a neural net trained for one task (such as age prediction) could be repurposed for others (such as car or fish identification).

This does in fact turn out to be the case. Rather than having to re-train an entire vision network for a new task, we can start by transferring over the weights of the lower layers of a vision network trained for a different one. Then, we can focus our computational resources on training only the other, deeper layers of the network.

The more similar the tasks that two networks are trained to perform, the more layers can be transferred from one of these networks to the other.

The process of using the lower layers of one trained vision network to bootstrap the training of another vision network trained for a different task is known as transfer learning.

Transfer learning can dramatically lower the cost associated with training new vision models. A common strategy used by AI labs is to train generalist vision models to identify a wide range of different objects, so that these models can be used as a base for other, task-specific models that could borrow the vast majority of their layers. In such cases, it is not unusual for all but one layer of the task-specific model to be transferred from a base generalist model. The process of freezing these imported lower layers and training only the deeper ones on a new task is known as fine-tuning.



2.12 Language models

Deep learning is fundamentally a procedure for mixing numbers in an intelligent way. We have seen how it can work on tabular data (such as hockey player stats) and image data, but text data poses a unique challenge: deep neural networks can only process lists of numbers. How then might we apply them to text data, which seems intrinsically non-numeric?

Somehow, we will need to convert text into a list of numbers. Many strategies exist that can achieve this, but we will present just one here.

Consider a list of numbers as long as the dictionary. The term for a list of numbers is a ***vector***. If the dictionary contains 200,000 words, then our vector will contain 200,000 numbers. Next, we will set all of these numbers to zero.

0, 0, 0, 0, ….

Now suppose that the first word in our dictionary is “aardvarks”, and that we want to represent that word as a vector. We could do that by selecting the first number in our vector, and turning it into a 1:

Aardvarks: 1, 0, 0, 0, ….

If the second word in our dictionary is “eat”, then we would represent that word by making the second entry in our vector a 1 instead:

Eat: 0, 1, 0, 0, ….

Finally, if the third word in our dictionary is “insects”, then would represent it by turning the third entry in our vector into a 1:

Insects: 0, 0, 1, 0, ….

We can use this strategy to represent entire sentences as vectors. To do that, we simply glue together the vectors associated with each of the words they contain. For example the sentence “Aardvarks eat insects” would be represented as:

1, 0, 0, 0, …. 0, 1, 0, 0, …. 0, 0, 1, 0, ….

aardvarks eat insects

Vectors such as these can become extremely long. In this example, if our dictionary contains 200,000 words, then each word is represented by a vector that contains 200,000 numbers. Even a short, 3-word sentence like “Aardvarks eat insects” becomes a 600,000-number vector!

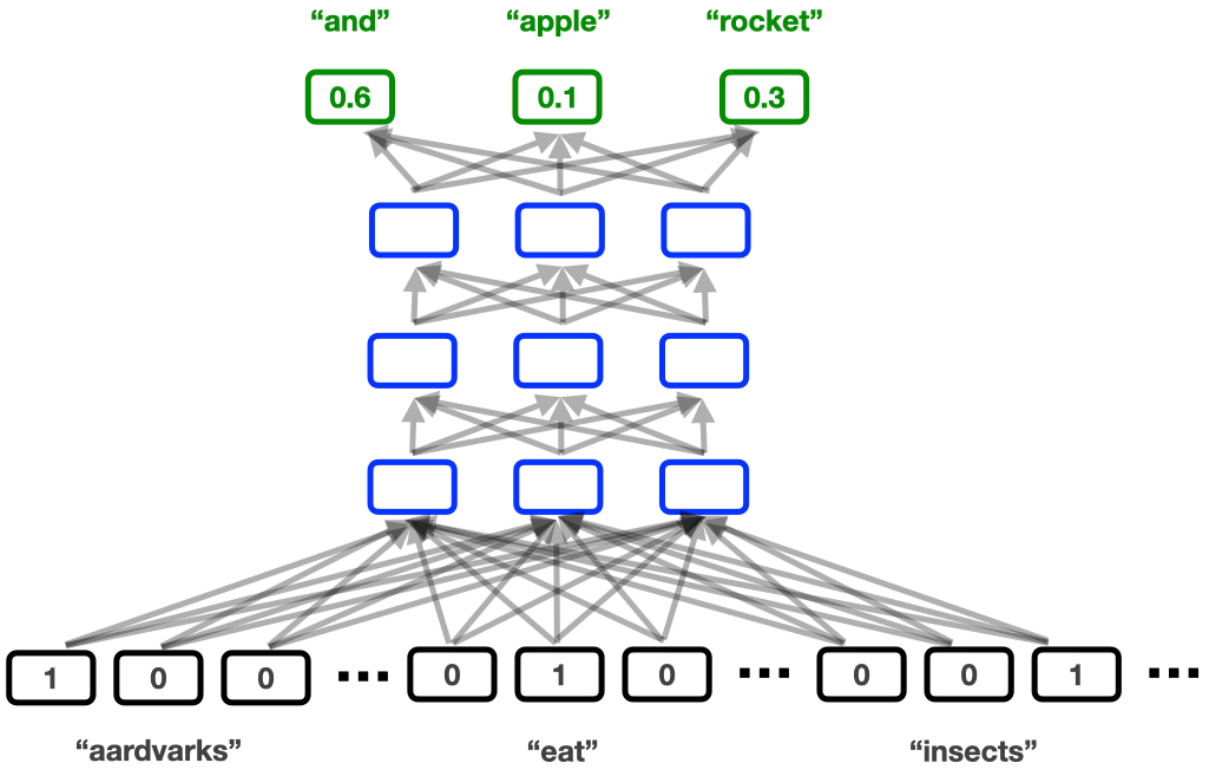
Nonetheless, because our text is now expressed as a vector, we can in principle feed it to a neural network, which we could train to perform useful functions. Deep neural networks that process text data are known as language models. The field of natural language processing is concerned with developing AI models that have the capacity to understand and process text.

One important example of a task that language models can be trained to perform is text autocomplete. In a text autocomplete task, a language model is fed a sentence as an input, and required to predict the next word in the sequence. For example, when given the sentence “Jack and Jill went up the”, a good language model would propose “hill” as the most likely next word. In this example, “Jack and Jill went up the” would be referred to as a prompt, and the word “hill” would be referred to as the model’s completion.

2.13 Language model anatomy

In practice, language models trained to perform text autocomplete do not only output one number, but rather a large set of numbers, each of which represents the probability that the next in the sequence will be a given word in the dictionary.

In the toy example below, a deep neural network takes in the sentence “Aardvarks eat insects”, and generates predictions for the probability that the next word in the sequence will be “and”, “apple”, and “rocket” (in practice, it would also generate probabilities for every other word in the dictionary).



This is a fairly typical structure, in which the model's outputs (the probabilities for "and", "apple", and "rocket") are each generated by mixing together the activations of the neurons in the final layer with a different set of weights.

This autocomplete task can be extended to allow a language model to generate entire sentences, or even paragraphs. To do this, we need only take the word identified by the model as most likely to follow in the sentence (in the case above, the word "and"), and append it to the initial prompt and thus obtain a new prompt (in this case, "aardvarks eat insects and"), which can be fed back into the model to generate a prediction for the next word in the sequence.

This is promising: it offers developers a way to use language models to generate long strings of valuable text, such as instructions, translations, or even code. However, it has an important limitation. As noted above, using this strategy, even a sentence as short as "Aardvarks eat insects" must be represented by a vector that might contain around a million numbers. Every neuron in the first layer of our neural network will have to assign a distinct weight to each of these numbers (there is an arrow connecting each number to each neuron in the first layer of the network). Consequently, the model must contain a vast number of weights in order to be able to process even simple sentences – weights that must all be tuned during the training process. As we've seen already, the more weights a model contains, the more data and compute it must consume during training to properly tune the values of those weights.

For this reason, a given language model will only be able to process a limited amount of text at a time. The length of the maximum input text sequence that a language model can process is known as its maximum sequence length. The larger a lab's compute budget and dataset, the larger a maximum sequence length they can train their model to have, and the more text the model can therefore use to inform its next-word predictions, and the more complex a set of ideas the model will be able to internalize during training.

Finally, it's worth mentioning that although the examples we've given above are based on converting entire words into vectors, alternative approaches also exist, which encode text at the syllable or character level rather than the word level. The atomic component of text that is converted into a vector in the way we've explored so far is known as a token, and the process of converting text into vectors in this manner is known as tokenization.

2.14 Transformers

Text autocomplete is a surprisingly important task, because training a model to become very good at text autocomplete forces that model to learn about a wide range of concepts, and form a general understanding of the world that it can then apply to solve other problems.

For example, a text autocomplete AI trained on a large body of text data (known as a corpus) will learn to autocomplete sentences such as "To counter rising inflation, the United States should _____". In order to fill in this blank successfully, a language model must have learned what the United States is, what inflation is, what it means for inflation to increase, and many other facts about finance and economics. Autocomplete is in some sense a universal task: a good autocomplete AI is potentially a good general-purpose reasoning machine. Those reasoning capabilities are encoded in the values of the weights stored in the neural network. The model's weights end up encoding a world model: an abstract representation of an increasing number of facts, logical rules, and entities that together form a kind of understanding of the world.

In 2017, a team of Google researchers interested in pushing the frontier of language modeling invented an important deep learning architecture known as the transformer. Transformers work according to the same principles that power all deep learning models: they represent their inputs as vectors, and mix the numbers they contain in stacked layers of neurons. Like other autocomplete models, they output predictions for the probabilities that certain words would follow their prompt.

However, transformers are distinguished by their use of an attention mechanism: a specialized group of neurons whose function is to identify certain words in their prompt to which the model should pay more or less attention when it generates its completion. This proved to be a key breakthrough: with the ability to focus more on some parts of a prompt than others, transformers were able to learn far more quickly, and rapidly became the go-to architecture for cutting-edge language modeling.

All modern cutting-edge language models use a variant of the transformer as a backbone, including OpenAI's GPT series, Google's PaLM series, DeepMind's Chinchilla and Gopher, and Meta's LLaMA.

2.15 Transfer learning in language models

It is important to note that transfer learning works just as well for language models as it does for vision models. A language model trained to perform a task like text

autocomplete can often have the vast majority of its weights frozen and transferred over to support a new task.

This is because, like vision models, language models tend to learn simpler, more concrete and generalizable concepts in their lower layers (concepts such as grammar, basic syntax, and simple phrases) and more abstract and nuanced concepts in their deeper layers (concepts such as geopolitics, drug interactions, and so on).

As a result, it's possible to freeze the majority of the layers of a text autocomplete model trained on a very diverse dataset consisting of books, blog posts, as well as Wikipedia and news articles, and fine-tune only its top layers on a body of specialized text (such as medical research papers) to create a specialist model capable of generating more insightful text about a given subject.

2.16 Large language models and the scaling hypothesis

In 2019, a segment of the frontier AI research community began to see language models as a potential path to creating far more general and powerful reasoning machines.

Their hypothesis relied on two assumptions:

1. By training language models to become progressively better at autocomplete, they could eventually build systems with a robust understanding of the world, which these models could then use to solve a wide range of problems, and potentially, all human-solvable problems.

2. The only thing required to increase the capability of language models to the extent described above is scaling. By dramatically increasing the parameter count of their language models, and training these models on massive datasets, using vastly greater quantities of computing power, they could push language model capabilities to achieve ever greater levels of intelligence even in the absence of any further breakthroughs.

Assumption 2 above has since become known as the scaling hypothesis. The scaling hypothesis remained a fringe perspective within the AI community until mid-2020, when OpenAI, a frontier AI lab, placed an unprecedented bet on AI scaling by training GPT-3: a language model that contained 175 billion parameters, trained on 500GB of text data pulled from all over the internet, using an estimated compute budget of $5M.

Thanks to its scale, GPT-3 appears to have developed precisely the capabilities hoped for by the then-fringe community of AI "scaling maximalists": without any additional training or fine-tuning, GPT-3 could perform tasks as varied as translation, coding, and basic web design, and could write short essays that were indistinguishably humanlike.

2.17 Why language is special

The most important data type for advanced AI today is text. This is because language is a human abstraction that eases generalizability: it is designed to generalize, and to support the consideration of new problems and situations. In addition, large volumes of text data all over the internet are readily available for use in machine learning, and encode a large fraction of human knowledge.

The current dependence of advanced AI on text data should not lead us to discount the possibility of other types of advanced AI arising (for example, AI systems that can explore the world and learn primarily from their interactions, rather than from text). However, it is very likely that text will continue to serve as a key pillar of advanced AI, and it is therefore important to dedicate special attention to language models.

2.18 How text datasets are created

The enormous text datasets used to train modern language models are sourced from internet scrapes, book corpuses, news articles, indexes, and content sites (such as YouTube, Reddit, etc).

Large scrapes of the internet are performed by a few major initiatives, as either part of search engine indexing or internet snapshot creation. There is a tremendous amount of engineering required to address the scalability, performance, and reliability requirements associated with these scrapes. Consequently, only a few players (Google, Microsoft, Common Crawl, etc) have the capacity to perform scrapes on massive scales. Many companies, such as data aggregators, media, or AI application developers, perform smaller-scale web scrapes, and there is a much lower barrier to entry for these more modest projects. Smaller scrapes of particular sites that have desirable data are often carried to augment the data used for training these more specialized models. For example, a company developing an AI-powered medical diagnostic chatbot might use small-scale scrapes of medical literature to fine-tune its chatbot models.

Notable data sources for scraping projects include:

- Common Crawl: A publicly available collection of web crawl data that contains petabytes of data from billions of web pages. Common Crawl has been updated every month since 2008, and provides an open and accessible corpus for anyone who wants to analyze and research the web. This is the most important publicly available source of text data. It resides on Amazon S3, a common data storage solution offered by Amazon Web Services (AWS), as part of the AWS Open Data Sponsorships program.

- Reddit, Wikipedia, patents, ArXiv (an online collection of academic papers, mostly from science and mathematics), Gutenberg (from Project Gutenberg, a volunteer-led effort to collect open-source books), BooksCorpus (books scraped from Smashwords, a platform for self-publishing ebooks), Github (the world's biggest code repository), and Youtube Transcripts.

These "raw" sources are then compiled into datasets that are used for training. The processing involves data cleansing, deduplication, filtering, and formatting (also known as data curation). Data cleansing involves removing html tags, ads, junk text, and other unwanted formatting from the raw source data.

Deduplication is the process of removing duplicate or redundant records from a dataset. This is a critical piece of the pipeline, because models learn less by being fed the same data repeatedly. As a result, feeding a model data it has already seen costs compute without yielding the same performance returns. In the limit, feeding the same data back to a model too often can lead to overfitting – a phenomenon whereby a model memorizes text, learning to reproduce specific examples from the training data, without being able to apply what it has learned to new examples that it has not seen before. Deduplication is technically challenging, and is often performed in many stages, such as deduplication based upon URL matching, text exact match, text fuzzy match, and content similarity.

Filtering removes low quality, undesired languages, and excludes certain content. Filtering is what ultimately controls what the model sees, and is often done with machine learning classifiers (simpler models designed to assign an input to any one of a set of predefined categories) that judge the quality of a document on a variety of axes, such as coherence, adult content, violence, etc. Filtration is also a "dark art": it is commonly understood at a high level, yet the specific models and techniques that support successful large-scale filtration projects are esoteric.

Data curation is the standardization of often highly heterogeneous formats into a consumable homogenous format. It involves data integration techniques to combine data from multiple sources into a unified form or schema. Often, this stage requires that error and anomaly detection techniques be applied to deal with noise, outliers, missing values, or inconsistencies. Data curation can also involve data annotation techniques to add labels or metadata to data to make it more informative or useful for the language modeling task.

Overall, these data processing steps are only partially public, highly variable, and comprise much of the exclusive knowhow needed for successful dataset creation. However, despite the challenges associated with creating high-quality datasets, it is worth noting that many such datasets are already public (OpenWebText, C4, The Pile, etc).

A partial list of important datasets includes:

- **WebText**: a custom dataset created by OpenAI that contains text data from various websites that are popular on Reddit. WebText is filtered to remove low-quality or offensive content and provides a high-quality corpus for natural language generation.

- **C4**: a dataset of about 750 GB of clean English text extracted from Common Crawl, and released by Google.

- **The Pile**: a dataset that contains 825 GB of text from 22 smaller, high-quality datasets combined together. It includes data from various sources, such as books, Wikipedia, GitHub, PubMed, arXiv, StackExchange, etc. One section is from Common Crawl. The Pile was developed and released by EleutherAI.

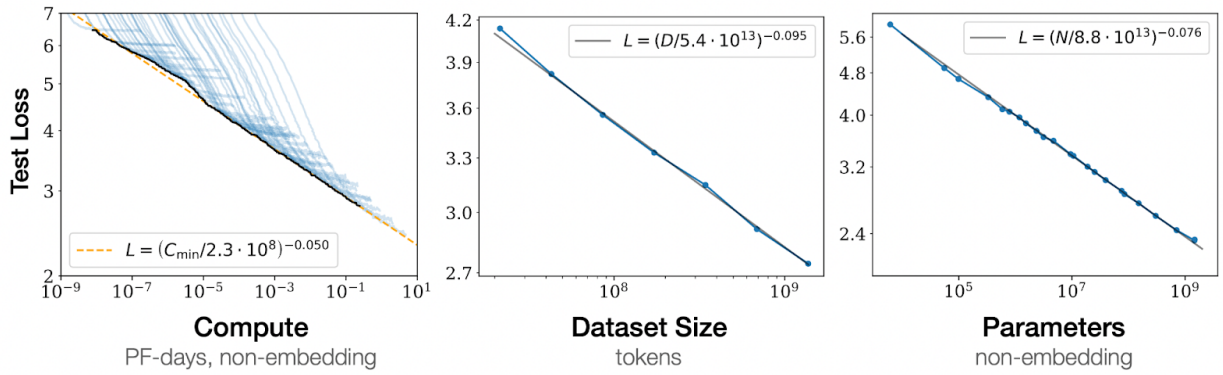
- **The BigScience Roots Corpus**: a dataset of about 1.8 TB of text from various domains such as biomedical literature, patents, news, and web text, released by Big Science.

The last step of the data pipeline is data consumption: This is the stage where the data is fed into the language model for training or evaluation. Data consumption can involve data loading techniques to load the data into memory for fast access. Data consumption can also involve data batching techniques to group the data into batches for parallel processing. This is a fairly standard process for which there are many existing solutions for language models.

2.19 Scaling laws and emergence

Crucially, GPT-3 came on the back of a groundbreaking research project also led by OpenAI, which investigated whether language models could be made to improve consistently and predictably simply through scale alone.

The results of their investigation were striking: they found that by increasing the parameter count, compute budget, and dataset size used to train their language models in certain optimal ratios, they were able to achieve reliable performance improvements. The plots below demonstrate their key result: a series of so-called AI scaling curves, which show how the number of errors that language models make when performing an autocomplete task (the test loss on the y-axis) decrease as those models's parameter compute budgets (left), dataset sizes (center) and parameter counts (right) were increased.



Source: J. Kaplan et al., "Scaling laws for neural language models," arXiv [cs.LG], 2020.

It is important to note that these scaling curves allowed OpenAI to anticipate how GPT-3 would perform at text autocomplete before it was trained. However, they did not allow OpenAI to anticipate what other capabilities GPT-3 might display. How good does a language model have to be at text autocomplete before it can translate languages, or write usable code? OpenAI's scaling research offered no answer. All it offered was a recipe that could convert dollars (in the form of data and compute) into intelligence (in the form of autocomplete performance). How precisely that intelligence would map onto capabilities was — and remains — anyone's guess.

Thus, each experiment in further scaling AI models leaves us with a more powerful system, whose precise abilities are destined to come as a surprise even to its

developers. The phenomenon by which AI scaling leads to greater intelligence, but unexpected and unpredictable new capabilities is known as emergence.

Emergence is a critical property to consider from an AI counterproliferation standpoint. It means that even an AI lab that builds a new, superscaled AI model cannot know what their model will be capable of once trained.

But even once the model is trained, no method exists for determining the complete range of capabilities that the model possesses. New GPT-3 capabilities were being discovered for months after its initial release to the public in the form of a paid service, for example — many of which could be put to malicious use. AI companies and labs that release their models in any form — whether as paid services or as free-to-access, open-source models — do not, and cannot know the full range of their capabilities.

The idea that raw scale would lead to reliable performance improvements in language models which, as GPT-3 showed, could already perform economically valuable tasks, led to a scaling race across the AI industry. Frontier AI labs such as OpenAI, Google's DeepMind, Google itself, Microsoft, and many other players rushed to build large language models (LLMs) with ever-increasing compute and data budgets. The once-fringe idea of AI scaling had become a fairly mainstream view almost overnight.

As the AI scaling race accelerated, the AI community's understanding of AI scaling laws was refined and improved. In 2022, DeepMind showed that the scaling recipe used by OpenAI to train GPT-3 relied on a model that was too large for its training dataset size and compute budget. Whereas OpenAI had trained GPT-3 with 1.7 tokens of text per parameter, DeepMind's new scaling laws showed that, in reality, a ratio of 20 tokens per parameter would have made better use of the compute resources invested in training the model.

The core findings of scaling research have been that model performance on text autocomplete scales as a power-law with model size, dataset size, and amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details, such as network width or depth, have minimal effects within a wide range.

In cases where a fixed compute budget is contemplated ("compute-optimal training"), the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. Too much data and the models can't fully utilize it. Too large of a model, and overfitting phenomena start to appear.

Large models require more resources when deployed and used post-training (at inference time) compared to smaller models. Smaller models trained on more data may require more time to train due to the need to process the additional data.

However, these scaling laws are not universal or absolute. They are based on empirical observations and may not hold for different tasks, domains, or metrics. The specifics may also change over time as new methods or technologies emerge, but the compute-optimal tradeoff between abstraction (from large models) and processing more data will almost certainly persist. Although scaling strategies have evolved as the AI community has discovered better ways to balance increases in parameter count, compute budget, and dataset size, the fact of scaling has remained the constant backdrop of AI progress in the last three years. The specifics may change over time as new methods or technologies emerge, but the compute-optimal tradeoff between abstraction (from large models) and processing more data will persist.

Bottlenecks to further scaling are likely to change, however. For example, dominant constraints may shift back and forth from compute to data. Currently, cutting-edge LLMs are estimated to have compute budgets on the order of several hundred million dollars, making further scaling of compute possible, but burdensome for reasonably large companies, and still readily accessible for tech giants such as Microsoft and Google. Forecasts indicate that the availability of language data to model developers may become a bottleneck to further scaling by 2030, despite an exponential, 50% year-over-year increase in available language data in recent years. As a result, industry is likely to shift its attention temporarily to increasing data efficiency in model design, though this will be a transitory phenomenon. In the long term, as language datasets continue to grow exponentially, and compute costs continue to drop exponentially thanks to Moore's Law and related hardware improvements, it will continue to be possible to turn ever greater investments in compute and data into AI capability and intelligence.

2.20 The modern training process

GPT-3 was in effect a superscaled text autocomplete AI: it was a model trained to take in a prompt as an input, and to generate a next-token prediction as an output.

This training process allowed GPT-3 to learn a remarkable amount about the world, and to develop a wide range of impressive capabilities. And yet, despite these raw capabilities, GPT-3 remained awkward to use.

Rather than a prompt like

```
Write an essay about Charles Dickens.
```

which it might complete by generating further instructions, prompts such as

```
The following is an essay about Charles Dickens:
```

yielded better results.

Because GPT-3 was fundamentally an autocomplete model, it would aim to produce text it considered to be the most likely to follow its prompt, rather than text that corresponded to what the prompt was asking for. The difficulty of designing effective prompts for GPT-3 made it difficult to assess the full range of GPT-3's capabilities. If GPT-3 failed to perform a task requested in a prompt, it was impossible to know with confidence whether this was because the model lacked the raw intelligence and capability required to carry out the task, or because the prompt was poorly designed.

Our ability to extract value from an AI model is therefore limited not only by the model's capabilities, but also by the extent to which its training is aligned with our needs. GPT-3 was a highly capable, but poorly aligned model.

More recently, AI labs have experimented with adding new training steps to better align the behavior of LLMs.

They typically begin by training the LLM to perform an autocomplete task on a massive amount of data, using vast quantities of compute, in a process known as pre-training. During pre-training, the LLM develops most of its raw capability, and learns to create useful and informative internal representations of its inputs.

Following this step, an LLM might be fine-tuned on a carefully curated task-specific text dataset. For example, GPT-3, once pre-trained, could be fine-tuned on a dataset consisting of human-to-human dialogue transcripts in order to encourage it to generate text in a more conversational style. This would be particularly useful if it were being prepared for use as a chatbot. During fine-tuning, the model is typically still being trained for next-token prediction, but on text that has characteristics that more closely reflect the model's desired style or behavior.

After fine-tuning, an LLM's performance might be further honed through a technique known as reinforcement learning from human feedback (RLHF). In RLHF, the objective function used to train the model is changed: rather than trying to autocomplete a

sequence of text, the model aims to produce outputs that will be rated highly by human evaluators. (In practice, the objective function is not supplied by human-generated ratings, but rather by a separate model that was itself trained to predict the rating that human evaluators would assign to a particular model output.)

Together, pre-training, fine-tuning, and RLHF now represent a de facto standard method for developing polished, market-ready AI chatbots at the frontier of AI capabilities. It was thanks to these techniques that OpenAI developed ChatGPT, and OpenAI has based subsequent generations of its LLM products, including GPT-4, on them as well.

2.21 In-context learning

After a model is trained, its parameter values are generally fixed (in other words, the weighting factors that the model learned during training are frozen). The model can then be put to use to generate outputs. As we've seen, this process is known as inference.

Since the model's weights are frozen during inference, we might expect that models are unable to learn or adapt to new information at this stage. Remarkably, however, this is not the case.

As an example, consider a LLM prompted to perform a task that did not appear anywhere in its training dataset, such as unscrambling the letters of a jumbled word. Given an appropriate prompt describing the task to be performed, and optionally but not necessarily some examples of the task to be performed, a sufficiently scaled LLM will be able to carry it out successfully.

This capability is known as in-context learning. In-context learning means that we can no longer think of model capabilities as static after the training process is complete: models can learn from and leverage new information at inference time, potentially leading to new capabilities. If you prompt a LLM to perform a complex task that it has never seen before, it will likely fail. But add a few examples of that task to your prompt, and the very same model will often learn from your prompt to perform it – all without having updated any of its weights.

2.22 The AI scaling race

Following the release of GPT-3 in mid-2020, the AI industry rapidly took onboard the idea that AI scaling would be required in order to push the frontier of AI capabilities.

However, given the massive computational requirements of superscale AI development, very few AI labs were in a position to enter the scaling race until 2021. By that time, OpenAI had compounded its first-mover advantage, using its scaled training infrastructure to build the first generation of text-to-image models (the DALL-E series), demonstrating that the scaling strategy was effective for applications beyond text-to-text language modeling.

Notably, the first AI lab publicly known to have trained an AI model at a scale comparable to GPT-3 was Chinese telecoms company Huawei. They developed PanGu Alpha, which slightly exceeded GPT-3's parameter count, clocking in at 207 billion parameters. PanGu Alpha was trained on a larger dataset than GPT-3, but used an unknown quantity of compute. Although its performance lagged somewhat behind GPT-3's, it demonstrated China's domestic AI scaling capability, and served as further evidence for the power of scaled training.

Another trend that emerged in 2021 was the development of multimodal AI. An AI model is considered multi-modal if its inputs or outputs can include multiple data types. For example, in mid-2021, the Beijing Academy of AI (BAAI) announced Wudao 2.0, a model that could take a text input, and generate either text or image outputs. But multimodality would soon go well beyond combined text/image capabilities. By 2022, DeepMind had built Gato: an AI with the capability to control robotic components, analyze images, and generate text. Its multi-modality also came with an ability to generalize effectively: the model was tested on 600 tasks, and performed 450 of them half as well (or better) than a human expert.

The years since 2020 have seen an industry-wide race to scale up AI systems. That race has featured AI labs from all over the world, with groups based in the United States, the United Kingdom, and China clearly leading the way.

2.23 GPT-4

In March 2023, OpenAI released GPT-4, a more scaled, hotly anticipated successor to GPT-3. GPT-4 came in two different versions:

- One was a text-in, text-out model – essentially, a more scaled version of GPT-3. This version was also trained using the standard pretraining/fine-tuning/RLHF pipeline discussed earlier.

- The second was a model that could take text and images as inputs, and generate text outputs. For example, you could feed it a blueprint of a rocketship

and a prompt like

```
This is a blueprint of my new reusable rocket. Do you think
this design is likely to work?
```

and depending on the level of complexity of the task, it could respond with a remarkable degree of insight and depth.

GPT-4 was a remarkable capability leap beyond GPT-3. It scores in the 90th percentile on the Uniform Bar Exam, and on tests as varied as AP History and the math SATs. It can provide detailed and accurate instructions for cleaning a piranha's fish tank, or extracting a strawberry's DNA. It can build entire websites from a hand-drawn sketch, and create simple video games within minutes.

OpenAI launched GPT-4 as a paid service, just as they had with GPT-3. Within weeks, open-source developers learned that they could give it a complex task like

```
Find a pair of sneakers for me. I play soccer once a week and
love to walk, but never in the rain.
```

and get it to break that task down into a list of workable steps. They then developed a framework that would farm out those steps to other instances of GPT-4, or to instances of other models. Because this configuration has GPT-4 running autonomously in pursuit of a user-specified goal, it was named ***Auto-GPT***. Auto-GPT and related projects such as BabyAGI have shown that properly configured GPT-4 instances can exhibit long-term planning capabilities, and execute against these plans effectively to solve fairly complex problems.

2.24 Potential limits to scaling transformers as a means of pursuing advanced AI

Given their preeminent role in the current advanced AI landscape, it is important to understand why transformers have worked so well compared to other AI models – including other neural networks. Understanding transformers is also valuable because it allows us to make educated guesses about the likelihood that the transformer will remain the gold standard model architecture for advanced AI, and to better understand the limitations of current AI models.

The effectiveness of transformers stems from three key factors:

1. **Parallelization & efficiency**: Parallelization in the transformer's attention mechanism (and specifically, a variant known as the self-attention mechanism) allows entire prompts and batches of prompts to be processed in tandem. This makes the transformer architecture even better suited to parallel processing in GPUs and TPUs than standard deep neural networks.

2. **Long Range Interactions**: It is often necessary for models to capture interactions between early parts of a sequence and much later parts. For instance, in the sentence: "Artificial Intelligence is a transformative technology pioneered in the mid 20th century, but not realized in general form until the early 21st century, that has the potential for great good and great harm," it is necessary to associate "Artificial Intelligence" with "great good and great harm". This requires a long-range interaction, as there are many words in between. Pre-transformer models often had inductive biases — a set of assumptions used to map inputs to outputs — favoring recency.

3. **Dynamic Adaptation**: It is often the case that certain parts of a sentence or input contain more information than others. A model that can dedicate more computational resources to process components of an input that it assesses are more informative is known as dynamic adaptation. The transformer attention mechanism provides this capability.

Despite these favorable characteristics, transformers do have important limitations. And given that transformers are the current model of choice for advanced AI development, their limitations are, in a sense, limitations on the current advanced AI paradigm, more broadly.

Some of these limitations are inherent to the problem of intelligence and others idiosyncratic to transformers. They can roughly be broken down into the following categories: capacity, veracity, reliability, and latency.

- **Capacity** is a broad term that refers to the amount of information that a model can process at a time. It captures things like memory requirements and maximum sequence length. Due to their attention mechanism, transformers require a large amount of memory and computational resources to store and update their parameters, especially for large-scale models with billions or trillions of parameters. To some degree, this requirement is specific to transformers (especially the transient memory required for attention). Advanced AI will almost certainly continue to require significant memory in order to store the embeddings (or more generally, the abstractions) that it is likely to rely on to

process its data and generate outputs. The maximum sequence length limits the maximum length of text that can be consumed at once. Increasing it makes the model more powerful, at the cost of increasing latency and memory requirements quadratically.

- **Veracity** is a measure of truthfulness in AI systems. Generative Transformer models are known to "hallucinate". Hallucination occurs when a model's responses are not justified by the training data or any proximal data. Hallucinations may take the form of unfactual or logically incoherent text outputs generated by an LLM, for example.

  Veracity is difficult to properly measure given a lack of clear boundaries on hallucinatory behavior. Often, the hallucinations are very plausible in strong models (similar to how humans often unintentionally confabulate by generating likely scenarios based on what they know). This is a feature intrinsic to current LLMs, because they are trained to perform an autocomplete task, which incentivizes them to produce sequences of text that are very likely to occur, as opposed to sequences of text that are accurate or truthful (a problem known as contextual grounding).

  The line between valid inference and hallucination can be blurry. LLMs generate text based on probabilities derived from their training data (next-word probabilities when trained to perform autocomplete tasks, and the probability of receiving high ratings from humans when trained via RLHF). When generating text, LLMs use their internal knowledge to make inferences about what is likely to be true given the context. However, sometimes the model's internal knowledge may be incomplete or inaccurate, leading it to generate text that is not grounded in reality or factual information or deciding incorrectly that it doesn't need to ground this information. This can lead to coherent, but inconsistent outputs.

- **Reliability** can be broken into two classes: calibration and brittleness.

  A "poorly calibrated" model is one that is over- or under-confident about its predictions. Calibration addresses the question: when a LLM predicts that there is a 40% probability that the next word in a sequence will be "cat", what fraction of the time is the next word actually "cat"? If the model is well-calibrated, then the word "cat" should follow 40% of the time. In the case of poor calibration, we might find that when a model predicts a very high (e.g. 90%) probability that the next word in a sequence is "cat", the next word actually ends up being "cat"

only 10% of the time.

Brittleness deals with the robustness of the representations the model has. We don't expect the predictions to be dramatically different for

```
The cat sat on a
```

versus

```
The cat sat on the
```

but for a brittle model, these may be vastly different. Models should be robust to trivial perturbations in the input.

- **Latency** refers to the time required for a model to generate its output. Significant latency is introduced in these models from the attention mechanism, which is of quadratic time complexity with respect to the sequence length. This means that if the sequence length doubles, the time taken for the attention operation would quadruple. There have been some efforts to mitigate this at either algorithmic or implementation (software and hardware) levels with relative success. The most important of these advances is FlashAttention, which optimizes the attention implementation by maximizing computation on fast-processing memory. This has led to up to 3x improvements in training times.

2.25 Model release

AI labs often choose to provide the public access to their models in various forms. There are many reasons why an AI lab may choose to do this, but an important one is the positive press and goodwill that it can generate.

The AI community has a strong open-source ethos, and many developers view the open sharing of fully trained models as being necessary to democratize access to increasingly powerful AIs, which they worry would otherwise be exclusively available to big technology companies. By open-sourcing certain models, AI labs can raise their profile, and draw goodwill from certain quarters of the AI community. In addition, many researchers prefer to work at labs that regularly produce open-source outputs, since this can raise the profile of their work.

There are currently four means by which AI models are made accessible to the public:

1. **Open-access**: An AI lab may choose to make their model's parameters freely available for download to any member of the public. Labs may do this for several reasons. First, publishing a model makes it possible for developers in the open-source community to augment and improve the model in various ways, effectively supplementing the original developer's internal technical teams. Second, open-source developers can develop a nuanced understanding of the lab's software stack, which increases the lab's pool of potential recruits. Finally, the AI community has a strong open-source culture, and companies that regularly release powerful AI models can benefit from positive press.

2. **Piracy**: A model initially released only to a small group of users (for example, academic researchers) may be leaked via torrenting websites or other means to the general public. This famously occurred with Meta's LLaMA model, a powerful LLM that Meta initially released only to a select group of researchers, but which has since leaked and is now in widespread use. Notably, at the time it was leaked, LLaMA represented a meaningful capability advance relative to other open-source LLMs. Leaks have therefore already shaped the frontier of open-source AI capabilities.

3. **API access**: An AI lab may choose to monetize a proprietary model by offering paid access to the general public. This access is generally granted on a pay-by-the-token basis for LLMs, or on a pay-by-the-input or pay-by-the-output basis for other model types. OpenAI's GPT-3, GPT-4, and ChatGPT models are all available via paid API, as are LLMs offered by rival companies, including AI21Labs, Cohere, Anthropic, Google, and Amazon.

4. **API access to embeddings**: As we saw earlier, feeding different inputs to a neural network leads its neurons to fire in different ways (in other words, to have different activations). Recall that the set of all activations in one layer of the network is known as an embedding, and is one way to represent the input data. Embeddings themselves can be useful inputs for downstream processing, and can offer insights into the input data associated with them. For that reason, AI companies sometimes offer paid access not only to their model's outputs, but also to their embeddings – the intermediate representations the models form of their inputs.

Finally, we note that as discussed earlier, labs may choose to release important assets associated with their models, with or without releasing the models themselves. For example, a lab might release the code or the dataset used to train the model instead of or along with the model proper.

2.26 Interpretability

As we have seen, deep neural networks are giant "number mixing" systems. At each stage of the mixing process that they run (at each layer of the network), they construct embeddings: highly abstract, intermediate representations of their inputs – representations that are not human-interpretable.

As LLMs have entered widespread use in increasingly high-stakes scenarios, our inability to explain and interpret their reasoning has become a source of significant risk.

Interpretability refers to the ability of an LLM to reveal its internal workings and logic, such as its weights or activations to be eventually understood by a human. Questions such as "what exactly is the significance of this neuron in my network?" and "can I determine what plans my network is developing based on looking at the activations of its neurons?" are questions about interpretability.

Explainability refers to the ability of an LLM to provide natural language explanations or justifications for its outputs and decisions. Questions such as "why did my neural network predict that this individual would fail to repay their loan?" are questions about explainability.

Despite the intrinsically non-interpretable way in which neural networks process inputs and generate outputs, techniques do exist that allow researchers to gain a partial understanding of their inner workings.

Probing is an interpretability method that aims to reveal what linguistic information is encoded in the neural network's embeddings. Remember that an embedding is a set of numbers – the activations of the neurons at one particular layer of the network. Ideally, those numbers will encode useful information about the input they're associated with. If that's the case, then there's no reason we can't treat them as the inputs to a new model, which we would train based on these activations to predict some linguistic property of interest, such as part-of-speech tags, syntactic dependencies, or semantic roles. The performance of this second model indicates how well the neural network represents that property in the layer being probed. Probing can help to understand what neural networks learn from natural language data, how they encode different levels of linguistic structure, and how they generalize across languages and domains.

Watchdog models aim to detect and correct errors or biases in neural networks. Watchdog models are trained to monitor the inputs and outputs of a neural network and flag any instances that violate some predefined criteria or expectations. For example, a watchdog model can check if a neural network produces coherent and consistent text summaries, if it preserves factual information from the source text, or if it avoids generating offensive or harmful language. Watchdog models can help to improve the quality and safety of language models, as well as to identify their limitations and weaknesses.

Mechanistic interpretability techniques are focused on explicitly understanding the neural circuitry within deep learning models, and associating individual neurons or groups of neurons with specific, human-understandable concepts. The hope is that mechanistic interpretability techniques will allow researchers to detect when advanced AI models are developing undesirable or dangerous plans and strategies before they can be executed.

2.27 Evaluating and benchmarking machine learning models

AI model evaluation typically relies on reserving a dataset of test tasks that were not present in the model's training dataset. This held-out dataset is known as a validation set, and great care must be taken to ensure that validation data are kept separate from training data, so as to produce reliable evaluations. This is easier said than done, however: as LLMs in particular are trained on enormous datasets consisting of vast internet scrapes, it becomes more difficult to guarantee that a particular validation task was not included in the training dataset.

Evaluating and benchmarking ML models is not a trivial task. It requires a large amount of data, computational resources, and time. Evaluation and benchmarking can offer critical insight into the capabilities and limitations of models, allow different models and methods to be compared, and guide future research and development. However, whereas AI models used to be capable of performing only specific, narrow tasks and therefore could be completely evaluated on the basis of their performance at that task, modern LLMs can perform a wide and unpredictable range of tasks. For example, it is typical to learn only months after the public release of an LLM that the model had important capabilities that had neither been anticipated nor detected beforehand. The general and uncertain extent of LLM capabilities makes it much more difficult to evaluate and quantify their performance.

Evaluating LLMs involves measuring how well they perform on specific tasks or domains using various metrics and criteria. Care has to be taken to ensure that the tasks have good coverage of desired abilities.

LLMs can be evaluated on the basis of their capacity to perform tasks without being shown specific examples of those tasks being completed ("zero-shot evaluation"), or with the benefit of having been provided a few such examples in a prompt ("few-shot evaluation"). Alternatively, LLMs can be subjected to fine-tuned evaluation, in which they are further trained on task-specific data. For example, a LLM initially trained on general text from all over the internet might be fine-tuned on medical literature and tested for its ability to diagnose diseases based on patient symptoms.

A benchmark is a common reference task against which the performance of certain types of models is conventionally measured. By comparing the performance of models on benchmark tasks, it is possible to better understand their relative performances, establish baselines and best practices for the field, and identify gaps and challenges for future research. For general language understanding tasks, such as reading comprehension or commonsense reasoning, common benchmarks are GLUE, SuperGLUE, SQuAD, or RACE, which consist of multiple subtasks that cover various aspects of natural language understanding. For domain-specific language understanding tasks, such as biomedical question answering or legal document analysis, common benchmarks are BioASQ, CORD-19 QA Challenge, Legal-BERT, or FinBERT, which consist of single or multiple subtasks that focus on a particular domain or application.

Some of the challenges associated with evaluation and benchmarking include:

- **Data quality and quantity**: Evaluating and benchmarking LLMs requires large quantities of high-quality data that are representative and diverse enough to capture the complexity and variability of natural language. Existing data may have issues such as noise, bias, inconsistency, or leakage into training data, which can affect the validity and reliability of the evaluation results.

- **Metric validity and reliability**: Evaluating and benchmarking LLMs requires appropriate metrics that can measure how well they perform on specific tasks. However, existing metrics may have issues such as low correlation with human judgments, high variance or sensitivity, or lack of interpretability or explainability, which can affect the accuracy and robustness of the evaluation results.

- **Model complexity and diversity**: Evaluating and benchmarking LLMs require efficient and scalable methods that can handle the increasing size and diversity of the models. This requires a continually adapting set of standardized benchmarks which perhaps portends governmental aid.