

AI SAFETY ATLAS

Chapter 1: Capabilities

MARKOV GREY & CHARBEL-RAPHAEL SEGERIE

French Center for AI Safety (CeSIA)

Links

How to Cite

Grey & Segerie (2024).
Capabilities. In *AI Safety Atlas* (Ch. 1).
<https://ai-safety-atlas.com/chapters/01/>

[AI Safety Atlas](#)
[Alignment Forum](#)
[Google Docs](#)
[Feedback](#)
[Facilitate](#)

1.1 Introduction

Yann LeCun Chief AI scientist at Meta and Turing Prize winner, May 2023 ([Heaven, 2023](#))

There is no question that machines will become smarter than humans—in all domains in which humans are smart—in the future. It's a question of when and how, not a question of if.

The field of artificial intelligence has undergone a remarkable transformation in recent years. This chapter lays the groundwork for the entire book by establishing what AI systems can currently do, how they achieve these capabilities, and how we might anticipate their future development. This understanding is essential for all subsequent chapters: the discussion of dangerous capabilities and potential risks (Chapter 2) follows directly from understanding capabilities. Similarly, proposed technical (Chapter 3) and governance solutions (Chapter 4) both must account for the current and projected future of AI capabilities.

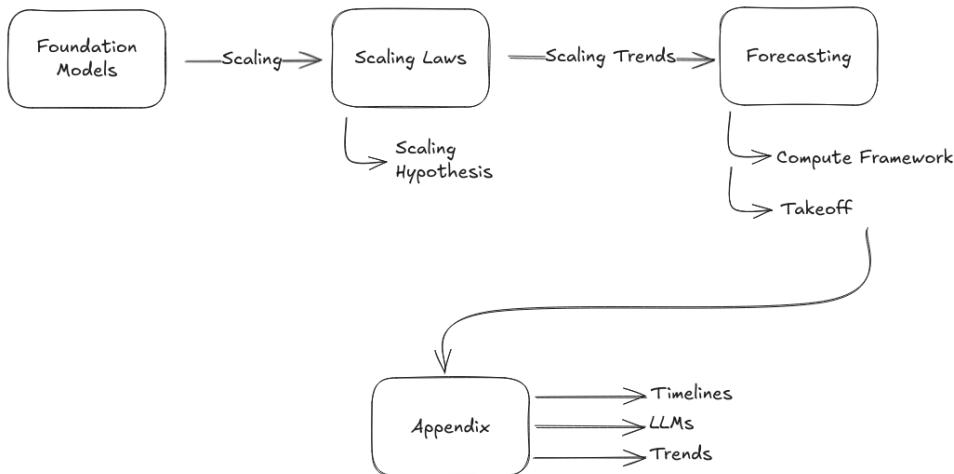


Figure 1.1 : We first explain foundation models, which have been continuously showing improved capabilities due to scale. Then examine empirically observed scaling laws. Based on these trends we look at some techniques that researchers use to try and forecast future AI progress.

State-of-the-Art AI - Achieved breakthrough capabilities across multiple domains . We begin by exploring how AI systems have evolved from narrow, specialized tools to increasingly general-purpose tools. Language models can now engage in complex reasoning, while computer vision systems demonstrate sophisticated understanding of visual information. In robotics, we're seeing the emergence of systems that can learn and adapt to real-world environments with increasing autonomy. The goal of this section is to give the reader many examples from different domains of accelerating AI capabilities.

Foundation models - Revolutionized how we build AI systems. The next section explores how we have moved from smaller specialized architectures to large scale general-purpose architectures. Rather than building separate systems for each task, these foundation models serve as the starting point. They are building blocks that can be later adapted for various applications using fine-tuning. We explore how these models are trained, their key properties, and the unique challenges they present. The emergence of unexpected capabilities from these models raises important questions about both their potential and implications for AI safety.

Understanding Intelligence - Capabilities require precise measurement to guide safety work . The objective of this section is to provide an understanding of what terms like artificial general intelligence and artificial superintelligence actually mean in practice. Through detailed case studies and empirical observations, we examine different approaches to defining and measuring AI capabilities. Moving beyond traditional binary distinctions between "narrow" and "general" AI, we introduce more nuanced formal

frameworks that track progress along multiple dimensions, essential for understanding when and how safety measures need to be implemented.

Scaling - The bitter lesson and empirical scaling laws show that scale drives progress. We explore how simple algorithms plus massive computation often outperform sophisticated hand-crafted approaches. This leads us to examine scaling laws that describe how AI performance improves with different variables like - data, parameter count and increased computational resources. This section also contains an examination of the debate around whether scale alone is sufficient for achieving transformative AI capabilities.

Forecasting - Predicting capabilities progress helps us prepare safety measures in advance. Building on our understanding of current capabilities and scaling behaviors, we examine various approaches to anticipating future progress. From biological anchors to trend analysis, we explore frameworks for making informed predictions about AI development trajectories. This is very important to know when different safety measures need to be in place.

Appendices - Overview of expert opinions on AI, detailed debates around scale, and scaling trends. We consider these sections optional, but still useful to those who want to get a little bit of a deeper dive. The chapter concludes with appendices examining expert opinions on AI progress, deeper discussions about the nature and limitations of large language models, and comprehensive data on key trends in AI development.

1.2 State-of-the-Art AI

Has not been updated since March 2024. Might not be representative of major developments.

Over the last decade, the field of artificial intelligence (AI) has experienced a profound transformation, largely attributed to the successes in deep learning. This remarkable progress has redefined the boundaries of AI capabilities, challenging many preconceived notions of what machines can achieve. The following sections detail some of these advancements.

Once a benchmark is published, it takes less and less time to solve it. This can illustrate the accelerating progress in AI and how quickly AI benchmarks are "saturating", and starting to surpass human performance on a variety of tasks. ([Our World in Data, 2023](#))

1.2.1 Language

Language-based tasks. There have been transformative changes in sequence and language-based tasks, primarily through the development of large language models (LLMs). Early language models in 2018 struggled to construct coherent sentences. The evolution from these to the advanced capabilities of GPT-3 (Generative Pre-Trained Transformer) and ChatGPT within less than 5 years is remarkable. These models demonstrate not only an improved capacity for generating text but also for responding to complex queries with nuanced, common-sense reasoning. Their performance in various question-answering tasks, including those requiring strategic thinking, has been particularly impressive.

GPT-4. One of the state-of-the-art language models in 2024 is OpenAI's LLM GPT-4. In contrast with the text-only GPT-3 and follow-ups, GPT-4 is multimodal: it was trained on both text and images. This means that it can now not only generate text based on images but has also gained some other capabilities. GPT-4 saw an upgraded context window with up to 32k tokens (tokens \approx words). The short-term memory limit of an LLM can be thought of as the model's ability to retain information from previous tokens within a certain context window. GPT-4 is trained via next-token prediction (autoregressive self-supervised learning). In 2018 GPT-1 was barely able to count to 10, while in 2024 GPT-4 can implement complex programmatic functions among other things.

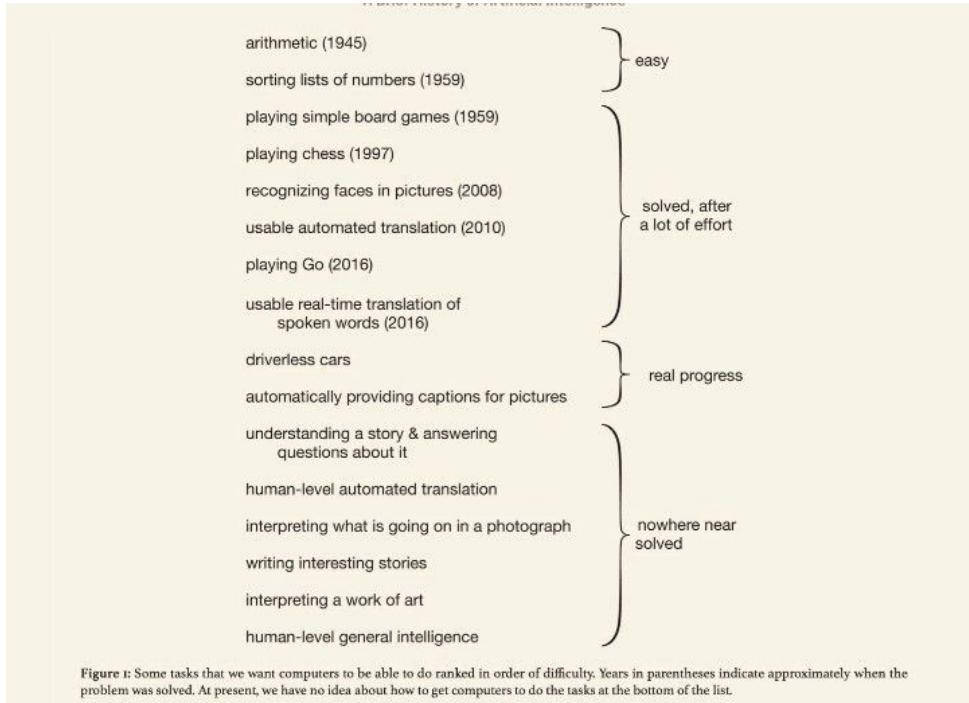


Figure 1.2 : A list of "Nowhere near solved" [...] issues in AI, from "A brief history of AI", published in January 2021 ([Wooldridge, 2021](#)). They also say: "At present, we have no idea how to get computers to do the tasks at the bottom of the list". But everything in the category "Nowhere near solved" has been solved by GPT-4 ([Bubeck et al., 2023](#)), except human-level general intelligence.

Scaling . Remarkably, GPT-4 is trained using roughly the same methods as GPT-1, 2, and 3. The only significant difference is the size of the model and the data given to it during training. The size of the model has gone from 1.5B parameters to hundreds of billions of parameters, and datasets have become similarly larger and more diverse.

We have observed that just an expansion in scale has contributed to enhanced performance. This includes improvements in the ability to generate contextually appropriate responses, and highly diverse text across a range of domains. It has also contributed to overall improved understanding, and coherence. Most of those advances in the GPT series come from increasing the size and computation power behind the models, rather than fundamental shifts in architecture or training.

Here are some of the capabilities that have been emerging in the last few years:

- **Few-shot and Zero-shot Learning .** The model's proficiency at understanding and executing tasks with minimal or no prior examples. 'Few-shot' means accomplishing the task after having seen a few examples in the context window, while 'Zero-shot' indicates performing the task without any specific examples ([Anthropic, 2022](#)). This also includes induction capabilities, i.e. identifying patterns and generalizing rules not present in the training, but only present in the current context window ([Brown et al., 2020](#)).
- **Metacognition .** This refers to the ability to recognize its own knowledge and limitations, for example, being able to know the probability of the truth of something ([Kadavath, 2022](#)).
- **Theory of Mind .** The capability to attribute mental states to itself and others, which helps in predicting human behaviors and responses for more nuanced interactions ([Kosinski 2023; Xu et al., 2024](#)).
- **Tool Use .** Being able to interact with external tools, like using a calculator or browsing the

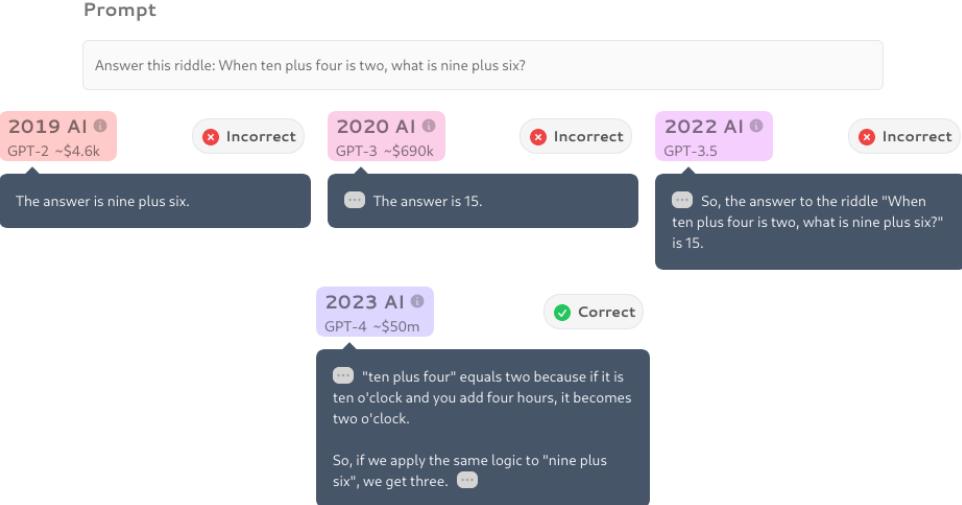


Figure 1.3 : How fast is AI Improving? (AI Digest, 2023)

internet, expanding its problem-solving abilities (Qin et al., 2023).

- **Self-correction** . The model's ability to identify and correct its own mistakes, which is crucial for improving the accuracy of AI-generated content (Shinn et al., 2023).

Prompt: Suppose $g(x) = f^{-1}(x)$, $g(0) = 5$, $g(4) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$ what is $f(f(f(6)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g . That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g , and vice versa. Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$. Therefore, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

Figure 1.4 : An example of a mathematical problem solved by GPT-4 using Chain of Thought (CoT), from the paper "Sparks of Artificial General Intelligence" (Bubeck et al., 2023).

- **Reasoning** . The advancements in LLMs have also led to significant improvements in the ability to process and generate logical chains of thought and reasoning. This is particularly important in problem-solving tasks where a straightforward answer isn't immediately available, and a step-by-step reasoning process is required. (Bubeck et al., 2023)
- **Programming ability** . In coding, AI models have progressed from basic code autocompletion to writing sophisticated, functional programs.
- **Scientific & Mathematical ability** . In mathematics, AI's have assisted in the subfield of automatic theorem proving for decades. Today's models continue to assist in solving complex problems. AI can even achieve a gold medal level in the mathematical Olympiad by solving geometry problems (Trinh et al., 2024).

1.2.2 Image Generation

The leap forward in image generation is not just in accuracy, but also in the ability to handle complex, real-world images. The latter, particularly with the advent of Generative Adversarial Networks (GANs) in 2014, has shown an astounding rate of progress. The quality of images generated by AI has evolved

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 1.5 : Note also the large jump from GPT-3.5 to GPT-4 in human percentile on these tests, often from well below the median human to the very top of the human range. ([Aschenbrenner, 2024](#); [OpenAI, 2023](#)). Keep in mind that the jump from GPT-3 to GPT-4 was in a single year.

from simple, blurry representations to highly detailed and creative scenes, often in response to intricate language prompts.



Figure 1.6 : An example of state-of-the-art image recognition. The Segment Anything Model (SAM) by Meta's FAIR (Fundamental AI Research) lab, can classify and segment visual data at highly precise levels. The detection is performed without the need to annotate images. ([Viso AI, 2024](#); [Meta, 2023](#))

The rate of progress within a single year alone is quite astounding as is seen from the improvements between the V1 of the MidJourney image generation model in early 2022, to the V6 in December 2023.

1.2.3 Multi & Cross modality

AI systems are becoming increasingly multimodal. This means that they can process images, text, audio, vision, and robotics using the same model. So they are trained using multiple different "modes" and can translate between them after deployment.

Cross-modality . A model is called cross-modal when the input of a model is in one modality (e.g. text) and the output is in another modality (e.g. image). The section on computer vision showed fast progress between 2014 and 2020 in cross-modality. We went from text-to-image models only capable of generating black-and-white pixelated images of faces, to models capable of generating an image of any textual prompt. More examples of cross-modality include OpenAIs Whisper ([Radford et al., 2022](#)) which is capable of speech-to-text transcription.

Multi-modality . A model is called multi-modal when both the inputs and outputs of a model can be in more than one modality. E.g. audio-to-text, video-to-text, text-to-image, etc...

DeepMind's 2022 Flamingo model, could be "*rapidly adapted to various image/video understanding tasks*" and "*is also capable of multi-image visual dialogue*". ([Alayrac et al., 2022](#)) Similarly, DeepMind's 2022 Gato model, was called a "Generalist Agent". It was a single network with the same weights which could "*play Atari, caption images, chat, stack blocks with a real robot arm, and much more*". ([Reed](#)

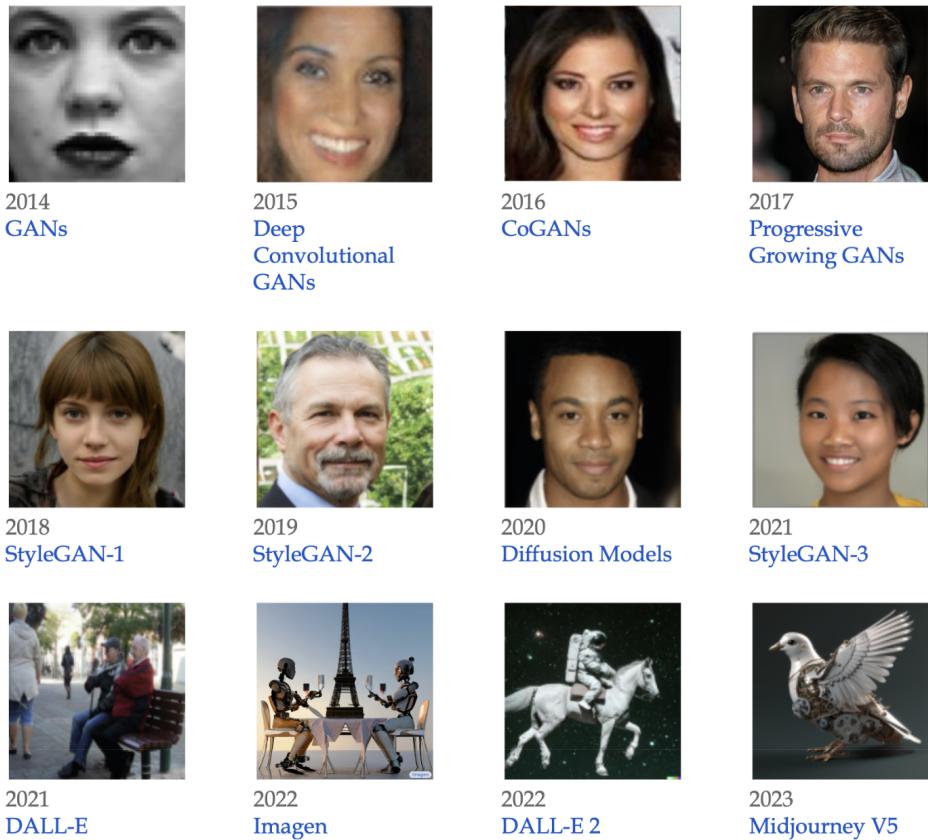


Figure 1.7 : An example of the evolution of image generation. At the top left, starting from GANs (Generative Adversarial Networks) to the bottom right, an image from MidJourney V5.



Figure 1.8 : MidJourney AI image generation over 2022-2023. Prompt: high-quality photography of a young Japanese woman smiling, backlighting, natural pale light, film camera, by Rinko Kawauchi, HDR ([Yap, 2024](#))

Input Prompt  <p>Question: What do you think the capacities of these are? Answer:</p>	 <p>Question: What is odd about this image? Explain why it is unusual. Answer:</p>	 <p>Question: What country is this? Why do you think so? Answer:</p>
Completion <p>The floppy disk is 1.44MB and the CD is 700MB.</p>	<p>The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.</p>	<p>It is Canada. I think so because the flag is the Canadian flag.</p>

Figure 1.9 : Image-to-text and text-to-image multimodality from the Flamingo model. ([Alayrac et al., 2022](#))

et al., 2022) Continuing this trend, DeepMind's 2023 Google Gemini model could be called a Large Multimodal Model (LMM). The paper described Gemini as "*natively multimodal*" and claimed to be able to "*seamlessly combine their capabilities across modalities (e.g. extracting information and spatial layout out of a table, a chart, or a figure) with the strong reasoning capabilities of a language model (e.g. its state-of-art-performance in math and coding)*"([Google, 2024](#))

1.2.4 Robotics

The field of robotics has also been progressing alongside artificial intelligence. In this section, we provide a couple of examples where these two fields are merging, highlighting some robots using inspiration from machine learning techniques to make advancements.



Figure 1.10 : Researchers used Model-Free Reinforcement Learning to automatically learn quadruped locomotion in only 20 minutes in the real world instead of in simulated environments. The Figure shows examples of learned gaits on a variety of real-world terrains. ([Smith et al., 2022](#))

Advances in robotics . At the forefront of robotic advancements is PaLM-E, a general-purpose, embodied model with 562 billion parameters that integrates vision, language, and robot data for real-time manipulator control and excels in language tasks involving geospatial reasoning. ([Driess et al., 2023](#))

Simultaneously, developments in vision-language models have led to breakthroughs in fine-grained robot control, with models like RT-2 showing significant capabilities in object manipulation and multimodal reasoning. RT-2 demonstrates how we can use LLM-inspired prompting methods (chain-of-thought), to learn a self-contained model that can both plan long-horizon skill sequences and predict robot actions. ([Brohan et al., 2023](#))

Mobile ALOHA is another example of combining modern machine learning techniques with robotics. Trained using supervised behavioral cloning, the robot can autonomously perform complex tasks "such as sauteing and serving a piece of shrimp, opening a two-door wall cabinet to store heavy cooking pots, calling and entering an elevator, and lightly rinsing a used pan using a kitchen faucet." (Fu et al., 2024) Such advancements not only demonstrate the increasing sophistication and applicability of robotic systems but also highlight the potential for further groundbreaking developments in autonomous technologies.

Instruction:

I need to hammer a nail,
what object from the scene
might be useful?

Prediction:

Rocks. Action: 1 129 138 122
132 132 106 127

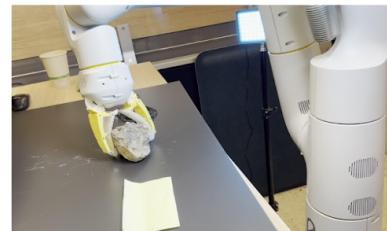
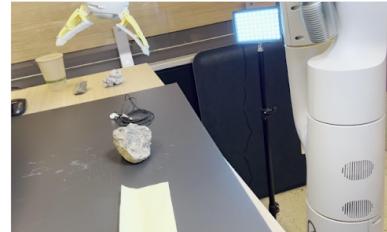


Figure 1.11 : DeepMinds RT-2 can both plan long-horizon skill sequences and predict robot actions using inspiration from LLM prompting techniques (chain-of-thought). (Brohan et al., 2023)

1.2.5 Playing Games

AI and board games. AI has made continuous progress in game playing for decades. Starting from AIs beating the world champion at chess in 1997, Scrabble in 2006 to DeepMind's AlphaGo in 2016 (DeepMind, 2016), which was good enough to defeat the world champion in the game of Go, a game assumed to be notoriously difficult for AI. Within a year, the next model AlphaGo Zero trained through self-play had mastered multiple games of Go, chess, and shogi reaching a superhuman level after less than three days of training. (Silver et al., 2017)

AI and video games. We started using machine learning techniques on simple Atari games in 2013 (Mnih et al. 2013). By 2019, OpenAI Five defeated the world champions at DOTA2 (OpenAI, 2019), while in the same year, DeepMind's AlphaStar beat professional esports players at StarCraft II (DeepMind, 2019). Both these games require thousands of actions in a row at a high number of actions per minute. In 2020 DeepMind MuZero model, described as "*a significant step forward in the pursuit of general-purpose algorithms*" (DeepMind, 2020), was capable of playing Atari games, Go, chess, and shogi without even being told the rules.

In recent years, AI's capability has extended to open-ended environments like Minecraft, showcasing an

ability to perform complex sequences of actions. In strategy games, Meta's Cicero displayed intricate strategic negotiation and deception skills in natural language for the game Diplomacy (Bakhtin et al., 2022).



Figure 1.12 : A map of diplomacy and the dialog box where the AI negotiates. ([DiploStrats \(YouTube\)](#), 2022)

Example of Voyager: Planning and Continuous Learning in Minecraft with GPT-4

Voyager (Wang et al., 2023) stands as a particularly impressive example of the capabilities of AI in continuous learning environments. This AI is designed to play Minecraft, a task that involves a significant degree of planning and adaptive learning. What makes Voyager so remarkable is its ability to learn continuously and progressively within the game's environment, using GPT-4 contextual reasoning abilities to plan and write the code necessary for each new challenge. Starting from scratch in a single game session, Voyager initially learns to navigate the virtual world, engage and defeat enemies, and remember all these skills in its long-term memory. As the game progresses, it continues to learn and store new skills, leading up to the challenging task of mining diamonds, a complex activity that requires a deep understanding of the game mechanics and strategic planning. The ability of Voyager to integrate new information continuously and utilize it effectively showcases the potential of AI in managing complex, changing environments and performing tasks that require a long-term buildup of knowledge and skills.

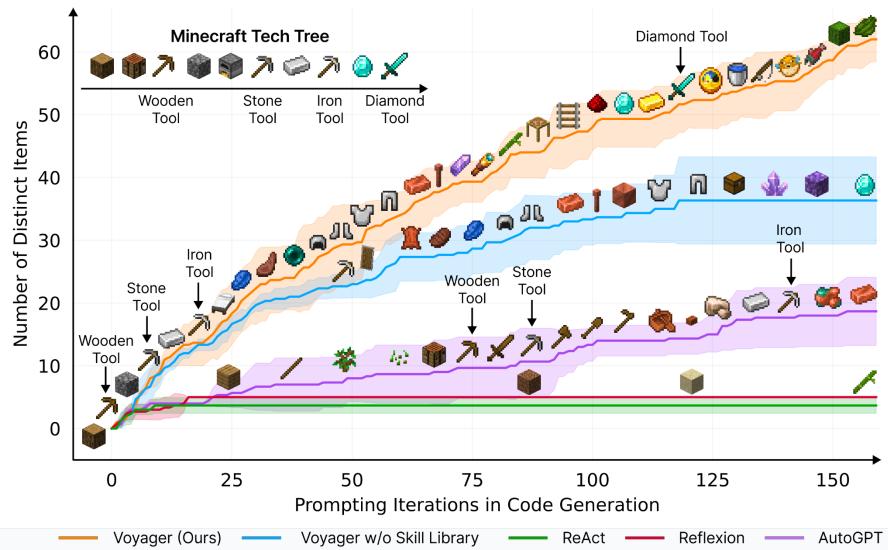


Figure 1.13 : Voyager discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baselines. (Wang et al., 2023)

1.3 Foundation Models

What are foundation models? Foundation models represent a fundamental shift in how we develop AI. Rather than building specialized models for many small specific tasks, we can now train large-scale models that serve as a "foundation" for many different applications. These models are then specialized later by a process called fine-tuning to perform specific tasks. Think of this as similar to how we can build many different types of buildings using the same base structure (Bommasani et al., 2022). We can build banks, restaurants, or housing but the underlying foundation remains largely the same. This is just a very quick intuitive definition. We will get more into the details in the next few subsections on training, properties and risks.

Why did the paradigm of foundation models come about? The traditional approach of training specialized AI models for every task often proved inefficient and limiting. Progress was bottlenecked by the need for human-labeled data and the inability to transfer knowledge between tasks effectively. Foundation models overcame these limitations through a process called self-supervised learning on massive unlabeled datasets. This breakthrough happened because of many different reasons - advances in specialized hardware like GPUs, new machine learning architectures like transformers, and increased access to huge amounts of online data (Kaplan et al., 2020) are some of the more prominent reasons for this shift.

What are examples of foundation models? In language processing, models like GPT-4 and Claude are examples of foundation models. Both of these have demonstrated the ability to generate human language, have complex conversations and perform simple reasoning tasks (OpenAI, 2023). Examples in computer vision include models like DALL-E 3 and Stable Diffusion. (Betker et al., 2023) These are domain specific examples, but we are also seeing a trend toward multimodal foundation models (LMMs). This includes things like GPT-4V and Gemini that can work across different types of data - processing and generating text, images, code, audio and probably more in the future (Google, 2023). Even in reinforcement learning, where models were traditionally trained for specific tasks, we're seeing foundation models like Gato demonstrate the ability to learn general-purpose behaviors that can be adapted to various different downstream tasks. (Reed et al., 2022)

What makes foundation models important for AI safety? The reason we start this entire book by talking about foundation models is because they mark a shift towards general-purpose systems, rather than narrow specialized ones. This paradigm shift introduces many new risks which didn't exist previously. These include misuse risks from centralization, homogenization, and dual-use capabilities just to name a

few. The ability of foundation models to learn broad, transferable capabilities has also led to increasingly sophisticated behaviors emerging from relatively simple training objectives (Wei et al., 2022). Complex capabilities, combined generality and scale, means we need to seriously consider safety risks beyond just misuse that previously seemed theoretical or distant. Beyond just misuse risk, things like misalignment are becoming an increasing concern with each new capability that these foundation models exhibit. We dedicate an entire chapter to the discussion of these risks. But we will also give you a small taste on the kinds of possible risks in the next few subsections, as it warrants some repetition.

What is the difference between foundation models and frontier models? Frontier models represent the cutting edge of AI capabilities - they are the most advanced models in their respective domains. While many frontier models are also foundation models (like Claude 3.5 Sonnet), this isn't always the case. For example, AlphaFold, while being a frontier model in protein structure prediction, isn't typically considered a foundation model because it's specialized for a single task rather than serving as a general foundation for multiple applications (Jumper et al., 2021). This distinction is worth keeping in mind, because most AI safety research and regulation frameworks focus on frontier models due to their advanced capabilities. When discussions about AI safety reference "foundation models," they're typically referring specifically to frontier foundation models - those foundation models that also represent the current state-of-the-art in capabilities. Understanding this distinction helps us better target and implement safety measures where they're most critically needed.

1.3.1 Training

How are foundation models trained differently from traditional AI systems? One key innovation of foundation models is their training paradigm. Generally, foundation models use a two-stage training process. First, they go through what we call a pre-training, and then second, they can be adapted through various mechanisms like fine-tuning or scaffolding to perform specific tasks. Rather than learning from human-labeled examples for specific tasks, these models learn by finding patterns in huge amounts of unlabeled data. This shift toward self-supervised learning on massive datasets fundamentally changes not just how models learn, but also what kinds of capabilities and risks might emerge (Bommasani et al., 2022). From a safety perspective, this means we need to understand both how these training methods work and how they might lead to unexpected behaviors.

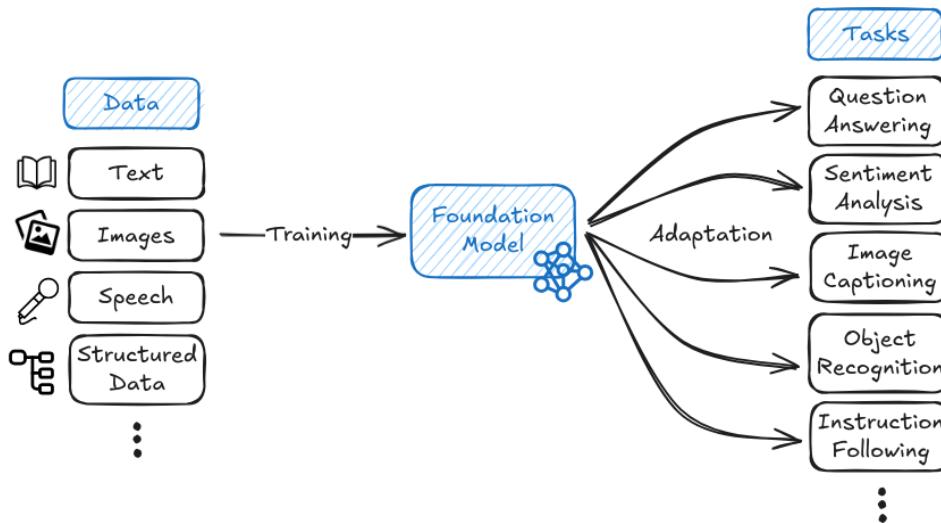


Figure 1.14 : On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)

What is pre-training? Pre-training is the initial phase where the model learns general patterns and knowledge from massive datasets of millions or billions of examples. During this phase, the model isn't trained for any specific task - instead, it develops broad capabilities that can later be specialized. This generality is both powerful and concerning from a safety perspective. While it enables the model to adapt

to many different tasks, it also means we can't easily predict or constrain what the model might learn to do (Hendrycks et al., 2022).

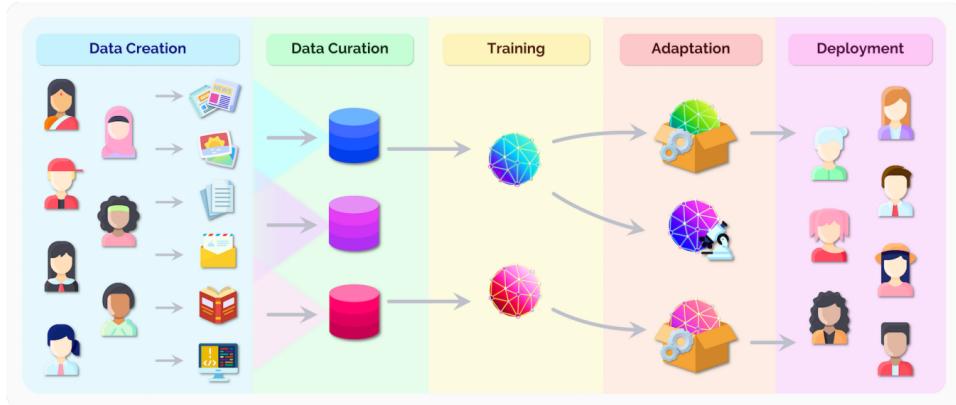


Figure 1.15 : On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)

How does self-supervised learning enable pre-training? Self-supervised learning (SSL) is the key technical innovation that makes foundation models possible. This is how we actually implement the pre-training phase. Unlike traditional supervised learning, which requires human-labeled data, SSL leverages the inherent structure of the data itself to create training signals. For example, instead of manually labeling images, we might just hide part of a full image we already have and ask a model to predict what the rest should be. So it might predict the bottom half of an image given the top half, learning about which objects often appear together. As an example, it might learn that images with trees and grass at the top often have more grass, or maybe a path, at the bottom. It learns about objects and their context - trees and grass often appear in parks, dogs are often found in these environments, paths are usually horizontal, and so on. These learned representations can then be used for a wide variety of tasks that the model was not explicitly trained for, like identifying dogs in images, or recognizing parks - all without any human-provided labels! The same concept applies in language, a model might predict the next word in a sentence, such as "The cat sat on the ...," learning grammar, syntax, and context as long as we repeat this over huge amounts of text.

What is fine-tuning ? After pre-training, foundation models can be adapted through two main approaches: fine-tuning and prompting. Fine-tuning involves additional training on a specific task or dataset to specialize the model's capabilities. For example, we might use Reinforcement Learning from Human Feedback (RLHF) to make language models better at following instructions or being more helpful. Prompting, on the other hand, involves providing the model with carefully crafted inputs that guide it toward desired behaviors without additional training. We'll discuss these adaptation methods in more detail in Chapter 8 when we explore scalable oversight.

Why does this training process matter for AI safety? The training process of foundation models creates several unique safety challenges. First, the self-supervised nature of pre-training means we have limited control over what the model learns - it might develop unintended capabilities or behaviors. Second, the adaptation process needs to reliably preserve any safety properties we've established during pre-training. Finally, the massive scale of training data and compute makes it difficult to thoroughly understand or audit what the model has learned. Many of the safety challenges we'll discuss throughout this book - from goal misgeneralization to scalable oversight - are deeply connected to how these models are trained and adapted.

1.3.2 Properties

Why do we need to understand the properties of foundation models? Besides just understanding the training process, we also need to understand the key defining characteristics or the abilities of these models. These properties often determine both the capabilities and potential risks of these systems. They

help explain why foundation models pose unique safety challenges compared to traditional AI systems. Their ability to transfer knowledge, generalize across many different domains, and develop emergent capabilities means we can't rely on traditional safety approaches that assume narrow, predictable behavior.

What is transfer learning? Transfer learning is one of the most fundamental properties of foundation models - their ability to transfer knowledge learned during pre-training to new tasks and domains. Rather than starting from scratch for each task, we can leverage the general knowledge these models have already acquired ([Bommasani et al., 2022](#)). This property enables rapid adaptation and deployment, it also means that both capabilities and safety risks can transfer in unexpected ways. For example, a model might transfer not just useful knowledge but also harmful biases or undesired behaviors to new applications.

What are zero-shot and few-shot learning? The ability to perform new tasks with very few examples, or even no examples at all. For example, GPT-4 can solve novel reasoning problems just from a natural language description of the task ([OpenAI, 2023](#)). This emergent ability to generalize to new situations is powerful but concerning from a safety perspective. If models can adapt to novel situations in unexpected ways, it becomes harder to predict and control their behavior in deployment.

Why is generality? Generalization in foundation models works differently from traditional AI systems. Rather than just generalizing within a narrow domain, these models can generalize capabilities across domains in surprising ways. However, this generalization of capabilities often happens without a corresponding generalization of goals or constraints - a critical safety concern we'll explore in detail in our chapter on goal misgeneralization. For example, a model might generalize its ability to manipulate text in unexpected ways without maintaining the safety constraints we intended ([Hendrycks et al., 2022](#)).

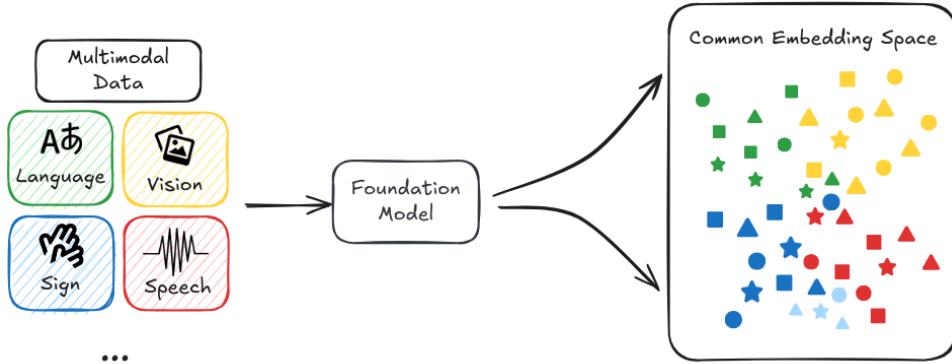


Figure 1.16 : On the Opportunities and Risks of Foundation Models ([Bommasani et al., 2022](#))

Why is multi-modality? Models can work with multiple types of data (text, images, audio, video) simultaneously. This isn't just about handling different types of data. A better description is that they can make connections across modalities in sophisticated ways ([Google, 2023](#)). From a safety perspective, multi-modality introduces new challenges because it expands the ways models can interact with and influence the world. A safety failure in one modality might manifest through another in unexpected ways.

Sam Altman (CEO of OpenAI) ([Cronshaw, 2024](#))

Multimodality will definitely be important. Speech in, speech out, images, eventually video. Clearly, people really want that. Customizability and personalization will also be very important.

1.3.3 Risks

What makes foundation models hard to control? The difficulty of controlling these models stems from three interconnected challenges. First, once trained, their internal representations and behaviors

are extremely difficult to modify in targeted ways. Unlike traditional software where we can directly edit specific functions, changing one behavior in a foundation model risks unpredictable effects on other capabilities. Second, when these models are deployed across many applications, maintaining control becomes a distributed systems problem. Safety failures can propagate through multiple systems before they're detected. Third, their black-box nature makes it extremely difficult to understand why they make specific decisions or how they might behave in novel situations - a challenge we'll explore deeply in our chapter on interpretability.

How do resource requirements limit development and access? Training foundation models require massive computational resources, creating a delicate balance between cost and accessibility. While adapting an existing model might be relatively affordable, the substantial initial training costs risk centralizing power among a few well-resourced entities. This concentration of power raises important questions about oversight and responsible development, that we'll address in our chapter on governance. For example, a single training run of GPT-4 sized models can cost tens or hundreds of millions of dollars, effectively limiting who can participate in their development. Continued scaling has also brought up many concerns around the environmental impact of AI training runs. ([Patterson et al., 2023](#)).

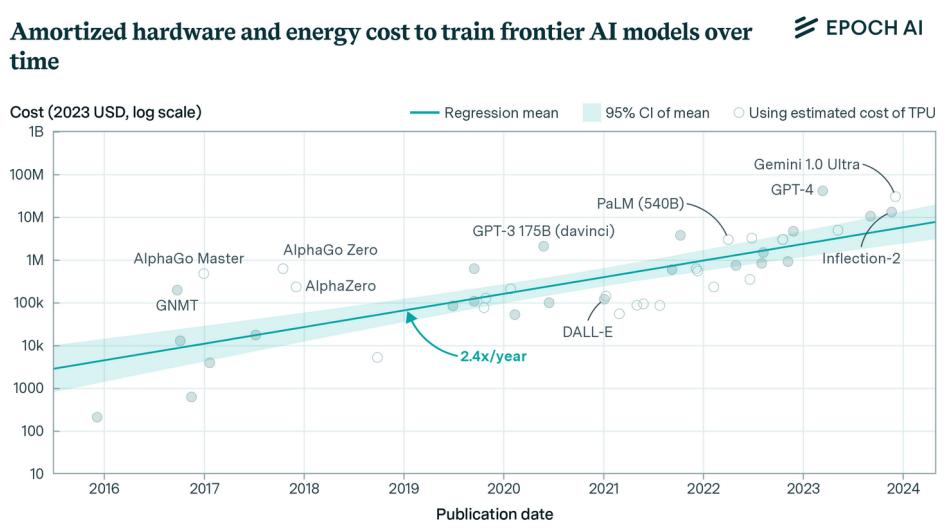


Figure 1.17 : The rising costs of training frontier AI models ([Cottier et al., 2024](#))

What risks come from homogenization? Homogenization occurs when many AI systems are derived from the same foundation models. This creates a systemic risk - if a foundation model has biases or failure modes, these could propagate to all models fine-tuned from it ([Bommasani et al., 2022](#)). For example, if a widely-used foundation model has encoded harmful biases or unsafe behaviors, these might manifest across numerous applications, from content generation to automated decision-making. They also have fixed knowledge cutoffs based on their training data, creating potential safety issues when deployed in rapidly changing environments. This risk of correlated failures becomes particularly concerning when foundation models are deployed in critical systems. ([Hendrycks et al., 2022](#)).

What is emergence, and why does it matter? Foundation models can effectively leverage increases in data, computation, and model size to improve their capabilities. The models become qualitatively different as they scale, often developing new capabilities entirely absent in smaller versions ([Kaplan et al., 2020](#)). This property is called emergence - the development of capabilities that weren't explicitly trained for. This can be both a good thing and a bad thing.

We might encounter new, unexpected behaviors and risks that weren't present in smaller models. This emergence often happens discontinuously as models scale up, making it difficult to predict what capabilities might suddenly appear. Which means it also makes it difficult to predict which risks will emerge when. When combined with homogenization, this unpredictability becomes especially concerning - a single foundation model integrated into multiple critical systems could lead to correlated failures that span

multiple safeguards or backup systems (Hendrycks et al., 2022). Therefore, this points to a need for a proactive approach to AI safety. Having safety measures in place before we begin scaling and training bigger models.

How do these limitations inform our approach to safety? We just briefly touched on some of the risks in this subsection to give you a basic intuition for why it is important to work on AI safety. The entire next chapter is dedicated to a deep dive on different risks from AI models, and how they might come about.

1.4 Intelligence

1.4.1 Case Studies

The bulk of this subsection will deal with the theory and historical aspects of defining intelligence. If you are more interested in just the core practical aspects of how we measure artificial general intelligence (AGI), then you can safely skip to the next subsection - measurement.

Why do we need to define intelligence? In our previous section on foundation models, we explored how modern AI systems are becoming increasingly powerful. But before we can meaningfully discuss the risks and safety implications of these systems, we need to agree on what we mean when we talk about AGI. Some believe that "sparks" of AGI are already present in the latest language models (Bubeck et al., 2023), while others predict human-level AI within a decade (Bengio et al., 2024). Without a clear definition, how are we supposed to assess such claims or plan appropriate safety measures?

The core point is that if you can't define something, you can't measure it. If you can't measure it, you can't reliably track progress or identify potential risks. Think about an example from physics - saying something like "it moved 5" makes no sense without specifying the unit of measurement. Did it move 5 meters, 5 feet, or 5 royal cubits? Nobody knows. If we don't know how far or fast it moved, then can we enforce speed limits? Also, no. The same applies to intelligence, and subsequent risks and safety techniques. Just as physics needed standardized units like meters and watts to advance beyond qualitative descriptions, AI safety research needs rigorous definitions to move beyond vague analogies and anthropomorphisms.

What makes defining intelligence so challenging? If everyone agrees that we need a definition to measure progress and design safety measures, then why don't we have a universally agreed upon definition? The problem is that the word intelligence is a term we use to describe multiple overlapping abilities - from problem-solving and learning to adaptation and abstract reasoning. Besides this, different academic disciplines view intelligence through different lenses. Psychologists emphasize measurable cognitive skills, computer scientists focus on task performance, and philosophers debate qualities like the relationship of intelligence to consciousness and self-awareness. So which approach is the most relevant to understanding and planning for AI safety?

Case Study: Imitation based approach to intelligence. The Turing Test (or the imitation game) suggested that intelligence could be measured through a machine's ability to imitate human conversation (Turing, 1950). However, this behaviorist approach proved inadequate - modern language models can often pass Turing-style tests while lacking fundamental reasoning capabilities (Rapaport, 2020). This is also still a process based approach, and was meant mainly as a philosophical thought experiment rather than a concrete operationalizable measure of intelligence.

Case Study: Consciousness based approaches to intelligence. One early view focused on machines that could truly understand and have cognitive states similar to humans (Searle, 1980). However, this definition proves problematic on multiple levels. First, consciousness remains poorly understood and difficult to measure. Second, we are unsure if intelligence and consciousness are necessarily linked - a system could potentially be highly intelligent without being conscious, or conscious without being particularly intelligent. A system doesn't need to be conscious to cause harm. Whether an AI system is conscious or not has no bearing on its ability to make high-impact decisions or take potentially dangerous actions.

Case Study: Brain analogy based approaches to intelligence. Another early approach defined AGI in terms of systems that rival or surpass the human brain in complexity and speed. This brain-centric definition is problematic for several reasons. While our brains may be the only example of general intelligence we have, modern AI has shown that matching human neural architecture isn't necessary for achieving intelligent behavior. From a safety perspective, focusing on brain-like architecture tells us little about what risks a system might pose - a system could be very unlike a brain in structure but still be capable of dangerous actions.

Case study: Process/Adaptability based approaches to intelligence. The process-based view sees intelligence as the efficiency of learning and adaptation, rather than accumulated capabilities. A few researchers adopt this view of intelligence. Under this view, intelligence is "*the capacity of a system to adapt to its environment while operating with insufficient knowledge and resources*" ([Wang, 2020](#)). Alternatively, it is described as "*the efficiency with which a system can turn experience and priors into skills*" ([Chollet, 2019](#)). While this focus on meta-learning and adaptation captures something fundamental about intelligence, but from a safety perspective, what ultimately matters is what these systems can actually do - their concrete capabilities - rather than how they achieve these capabilities. This leads us to the final approach.

Case study 5: The capabilities approach to intelligence . The motivating question behind this view is - If an AI system can perform dangerous tasks at human-level or beyond, does it really matter whether it achieved this through sophisticated learning processes, efficient memorization, with/without consciousness? If an AI system has capabilities that could pose risks - like sophisticated planning, manipulation, or deception - these risks exist regardless of whether the system acquired these capabilities through "true intelligence", "real understanding" or sophisticated pattern matching. The capabilities-based approach cuts through philosophical debates by asking concrete questions: What can the system actually do? How well can it do it? What range of tasks can it handle? This framework provides clear standards for progress and, crucially for safety work, clear ways to identify potential risks. The majority of AI labs use this capabilities-focused approach in how they frame their AGI goals. For example, AGI has been defined as "*highly autonomous systems that outperform humans at most economically valuable work*" ([OpenAI, 2014](#)). Safety considerations are framed similarly in saying that the mission is to ensure "*transformative AI helps people and society*" ([Anthropic, 2024](#)).

Capabilities vs Intelligence ([Krakovna, 2023](#))

When discussing AI risks, talk about capabilities, not intelligence... People often have different definitions of intelligence, or associate it with concepts like consciousness that are not relevant to AI risks, or dismiss the risks because intelligence is not well-defined.

Given these considerations, for the vast majority of this book, our primary focus will remain on the practical framework of capabilities for evaluation and safety assessment. This capabilities-focused approach is most relevant for immediate safety work, regulation, and deployment decisions. We acknowledge that research into consciousness, sentience, ethics surrounding digital minds and the fundamental nature of intelligence continues to be valuable but is less actionable for immediate safety work.

In our next subsection, we will explore how we can concretely define and measure capabilities within this framework. We'll see how moving beyond simple binary thresholds of "narrow" versus "general" AI helps us better understand the progression of AI capabilities and their associated risks.

1.4.2 Measuring

Lord Kelvin (Oxford essential quotations, 2016)

When you can measure what you are speaking about, and express it in numbers, you know something about it; when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science.

Why do traditional definitions of AGI fall short? In the previous section, we explored how foundation models are becoming increasingly powerful and versatile. But before we can meaningfully discuss risks and safety implications, or make predictions about future progress, we need clear ways to measure and track AI capabilities. This section introduces frameworks for measuring progress toward artificial general intelligence (AGI) and understanding the relationship between capabilities, autonomy, and risk. For example, OpenAI's definition of AGI as "*systems that outperform humans at most economically valuable work*" (OpenAI, 2014), or the commonly used definition "*Intelligence measures an agent's ability to achieve goals in a wide range of environments.*" (Legg and Hutter, 2007) and many others are not specific enough to be operationalizable. Which humans? Which goals? Which tasks are economically valuable? What about systems that exceed human performance on some tasks but only for short durations?

Why do we need better measurement frameworks? Historically, discussions about AGI have often relied on binary thresholds - systems were categorized as either "narrow" or "general", "weak" or "strong", "sub-human" or "human-level." While these distinctions helped frame early discussions about AI, they become increasingly inadequate as AI systems grow more sophisticated. Just like we sidestepped debates around whether AIs display "true intelligence" or "real understanding" in favor of a more practical framework that focuses on capabilities, similarly we want to avoid debates around things like whether a system is "human-level" or not. It is much more pragmatic to be able to make statements like - it outperforms 75% of skilled adults on 30% of cognitive tasks.

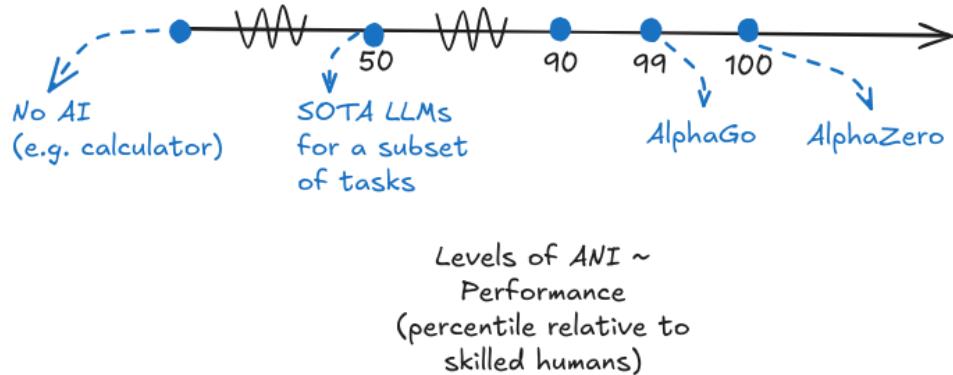


Figure 1.18 : This is the continuous outlook of AI measuring performance. All points on this axis can be called ANI (except for the origin).

Definition: Artificial Narrow Intelligence (ANI) (IBM, 2023)

Weak AI—also called Narrow AI or Artificial Narrow Intelligence (ANI)—is AI trained and focused to perform specific tasks. Weak AI drives most of the AI that surrounds us today. ‘Narrow’ might be a more accurate descriptor for this type of AI as it is anything but weak; it enables some very robust applications, such as Apple’s Siri, Amazon’s Alexa, IBM Watson, and autonomous vehicles.

Levels of artificial narrow intelligence (ANI). We should think about the performance of AI systems on a continuous spectrum. Traditional definitions of ANI correspond to high performance on a very small percentage of tasks. For example, chess engines like AlphaZero outperform 100% of humans, but only on roughly 0.01% of cognitive tasks. Similarly, specialized image recognition systems might outperform 95% of humans, but again only on a tiny fraction of possible tasks. According to the definition above, all these systems would be defined as ANI, but if we think about them in a continuous range of what percentage of skilled humans they can outperform we get a much more specific and granular picture.

How can we build a better measurement framework for AGI? We need to track AI progress along both - performance (how well can it do things?) and generality (how many different things can it do?). Just like we can describe a point on a map using latitude and longitude, we can characterize AGI systems by their combined level of performance and degree of generality, as measured by benchmarks and evaluations. This framework gives us a much more granular way to track progress. This precision helps us better understand both current capabilities and likely development trajectories.

Performance (rows) x Generality (columns)	Narrow <i>clearly scoped task or set of tasks</i>	General <i>wide range of non-physical tasks, including metacognitive tasks like learning new skills</i>
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging <i>equal to or somewhat better than an unskilled human</i>	Emerging Narrow AI GOFAI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai & Hassabis, 2023)
Level 2: Competent <i>at least 50th percentile of skilled adults</i>	Competent Narrow AI toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLi (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert <i>at least 90th percentile of skilled adults</i>	Expert Narrow AI spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso <i>at least 99th percentile of skilled adults</i>	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman <i>outperforms 100% of humans</i>	Superhuman Narrow AI AlphaFold (Jumper et al., 2021 ; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved

Figure 1.19 : Table of performance x generality showing both levels of ANI, and levels of AGI.

Where do current AI systems fit in this framework? Large language models like GPT-4 show an interesting pattern - they outperform roughly 50% of skilled adults on perhaps 15-20% of cognitive tasks (like basic writing and coding), while matching or slightly exceeding unskilled human performance on a broader range of tasks. This gives us a more precise way to track progress than simply debating whether such systems qualify as "AGI." LLMs like GPT-4 are early forms of AGI (Bubeck, [2023](#)), and over time we will achieve stronger AGI as both generality and performance increase. To understand how this continuous framework relates to traditional definitions, let's examine how key historical concepts map onto our performance-generality space.

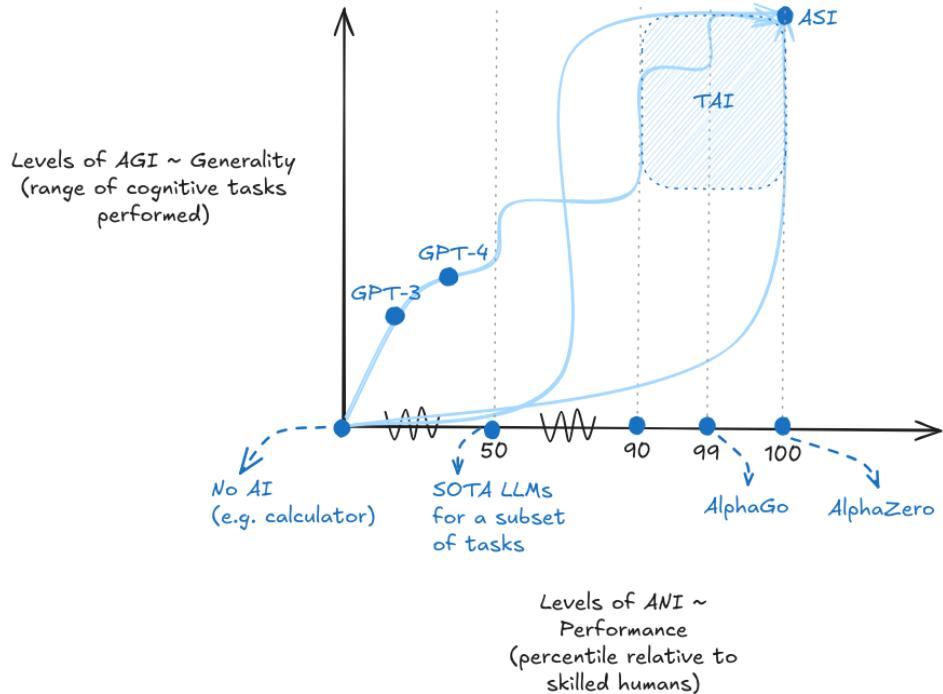


Figure 1.20 : The two-dimensional view of performance x generality. The different curves are meant to represent the different paths we can take to ASI. Every single point on the path corresponds to a different level of AGI. The specific development trajectory is hard to forecast. This will be discussed in the section on forecasting and takeoff.

Definition: Transformative AI (TAI) (Karnofsky, 2016)

Potential future AI that triggers a transition equivalent to, or more significant than, the agricultural or industrial revolution.

Transformative AI (TAI) . Transformative AI represents a particularly interesting point in our framework because it isn't tied to specific performance or generality thresholds. Instead, it focuses on a range of impacts. For example, a system could be transformative by achieving moderate performance (outperforming 60% of humans) across a wide range of economically important tasks (50% of cognitive tasks), or by achieving exceptional performance (outperforming 99% of humans) on a smaller but critical set of tasks (20% of cognitive tasks).

Definition: Human Level AI (TAI) (Karnofsky, 2016)

Potential future AI that triggers a transition equivalent to, or more significant than, the agricultural or industrial revolution.

Human Level AI (HLAI) . This term is sometimes used interchangeably with AGI, and refers to an AI system that equals human intelligence in essentially all economically valuable work. However, we only explain it here for reasons of completeness. Human-level is not well-defined which makes this definitions difficult to operationalize. If we map this onto the levels of AGI framework, then it roughly would correspond to outperforming 99% of skilled adults at most cognitive non physical tasks.

Definition: Artificial Superintelligence (ASI) (Bostrom, 2014)

Any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.

Artificial Superintelligence (ASI). If systems achieve superhuman performance at all cognitive tasks, then this would be the strongest form of AGI, also called superintelligence. In our framework, ASI represents the upper-right corner - systems that outperform 100% of humans on nearly 100% of cognitive tasks.

What is the relationship between levels of AGI and risk? Understanding AI systems through continuous performance and generality measures helps us better assess risk. Rather than waiting for systems to cross some "AGI threshold," we can identify specific combinations of performance and generality that warrant increased safety measures. For example:

- A system achieving 90% performance on 30% of tasks might require different safety protocols than one achieving 60% performance on 70% of tasks
- Certain capability combinations might enable dangerous emergent behaviors even before reaching "human-level" on most tasks
- The rate of improvement along either axis provides important signals about how quickly additional safety measures need to be developed

There are various other variables that we can add to make this picture even more precise. For example, just like we have levels of performance and generality, we can also have levels of autonomy with which these systems operate. As an example, at a low level of autonomy a human fully controls a task and uses AI to automate mundane sub-tasks, whereas at a higher level of autonomy we might see the AI take on a substantive role, or even co-equal work division. (Morris et al., 2024) Similarly, we have the variable of propensities which would measure what the AI model tends to do by default (Shevlane et al., 2023), and the variable of controllability which measures what percent of the time the AI model is able to subvert our current safety measures (Roger et al., 2023). Combining our definition of levels of AGI with variables like this gives us an extremely accurate picture of what the model is able to, and allows actionable technical safety and regulatory proposals.

1.5 Scaling

In the previous section, we explored how we can measure AI capabilities along continuous dimensions of performance and generality. Now we'll examine one of the most important drivers behind improvements in these capabilities: scale.

1.5.1 The Bitter Lesson

We assume that most of you probably went to university in an era where machine learning and AI roughly mean the same thing, or rather deep learning and AI mean the same thing. This hasn't always been true. Early in the history of artificial intelligence, researchers took very different approaches to creating intelligent systems. They believed that the key to artificial intelligence was carefully encoding human knowledge and expertise into computer programs. This led to things like expert systems filled with hand-crafted rules and chess engines programmed with sophisticated strategic principles. However, time and time again, researchers learned what we now call the bitter lesson.

The Bitter Lesson (Sutton, 2019)

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. [...] The bitter

lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.

What makes this lesson bitter? The bitterness comes from discovering that decades of careful human engineering and insight were ultimately less important than simple algorithms plus computation. In chess, researchers who had spent years encoding grandmaster knowledge watched as "brute force" search-based approaches like Deep Blue defeated world champion Garry Kasparov. In computer vision, hand-crafted feature detectors were outperformed by convolutional neural networks that learned their own features from data. In speech recognition, systems based on human understanding of phonetics were surpassed by statistical approaches using hidden Markov models ([Sutton, 2019](#)).

Does the bitter lesson mean we don't need any human engineering? Human ingenuity playing a smaller role in improving AI is a subtle point that can be easily misunderstood. The transformer architecture for example might seem to contradict the bitter lesson because they rely on sophisticated architectural innovations. Human ingenuity is important, but the subtlety is in recognizing that there's a difference between two types of human engineering:

- **Algorithm-level improvements:** These make better use of existing compute, like: better optimizers (Adam), architecture innovations (transformers, attention mechanisms) or training approaches (better learning rate schedules).
- **Domain-specific engineering improvements:** These try to encode human knowledge, like: special architectures designed for specific problems, hand-crafted features or rules or task-specific inductive biases.

The bitter lesson isn't arguing against all human engineering - it's specifically cautioning against the second type. The transformer architecture exemplifies this pattern - it doesn't encode any specific knowledge about language, but rather provides a general mechanism for learning patterns that becomes increasingly powerful as we scale up compute and data.

What does the bitter lesson mean for AI Safety? In the previous section, we discussed measuring AI capabilities along continuous dimensions of performance and generality. If it is the case that these continue to advance primarily through scaling, then we may have more predictable trajectories. This creates both opportunities and risks because performance and generality will continue to rise as long as scale is all that we need, but it also means that forecasting AI development trajectories and preparing appropriate safety measures for predictable capability levels is possible.

1.5.2 Scaling Laws

Why do AI labs care about scaling laws? Training large AI models is extremely expensive - potentially hundreds of millions of dollars for frontier models. Scaling laws help labs make crucial decisions about resource allocation: Should they spend more on GPUs or on acquiring training data? Should they train a larger model for less time or a smaller model for longer? For example, with a fixed compute budget, they might need to choose between training a 20-billion parameter model on 40% of their data or a 200-billion parameter model on just 4%. Getting these tradeoffs wrong can waste enormous resources. So it is important to be able to have a predictable relationship between how you invest your money and what level of capabilities you get at the end.

What are scaling laws? Scaling laws are mathematical relationships that describe how an AI system's performance changes as we vary key inputs like model size, dataset size, and computing power. These are empirical power-law relationships that have been observed to hold across many orders of magnitude. The



Prompt: A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

Figure 1.21 : Example of capabilities increasing with an increase with one of variables in the scaling laws - parameter count. The same model architecture (Parti) was used to generate an image using an identical prompt, with the only difference between the models being the parameter size. There are noticeable leaps in quality, and somewhere between 3 billion and 20 billion parameters, the model acquires the ability to spell words correctly. (Yu et al., 2022)

key variables involved are:

Compute (C) : This represents the total processing power used during training, measured in floating-point operations (FLOPs). Think of this as the training "budget" - more compute means either training for longer, using more powerful hardware, or both. While having more GPUs helps increase compute capacity, compute ultimately refers to the total number of operations performed, not just hardware.

Parameters (N) : These are the tunable numbers in the model that get adjusted during training - like knobs that the model can adjust to better fit the data. More parameters allow the model to learn more complex patterns but require more compute per training step. Current frontier models have hundreds of billions of parameters.

Dataset size (D) : This measures how many examples the model trains on (typically measured in tokens for language models). The larger the dataset, the more information the model can read. Simultaneously, to read and learn from more data, the training runs also need to be generally longer, which in turn increases the total compute needed before the model can be considered "trained".

Loss (L) : This measures how well the model performs on its training objective. This is what we are trying to minimize, and it tends to improve as we scale up these variables.

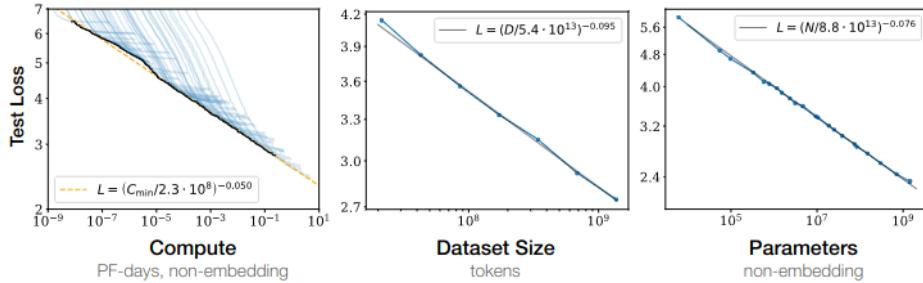


Figure 1.22 : Language modeling performance improves smoothly as we increase the model size, dataset set size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two. (Kaplan et al., 2020)

OpenAI's initial scaling laws in 2020 . To determine the relationships between different variables that

might contribute to scale, OpenAI conducted a series of experiments. For an intuitive idea of how they came up with the scaling laws, you can imagine that while training a model you can hold some variables fixed while varying others and see how loss changes. Eventually this allows you to see some patterns. As an example, dataset size can be kept constant, while parameter count and training time are varied, or parameter count is kept constant and data amounts are varied, etc... So we can get a measurement of the relative contribution of each towards overall performance. If these relationships hold true across many different model architectures and tasks, then this suggests they capture something fundamental about deep learning systems. This is how the first generation of scaling laws came about from OpenAI. For example, by these laws if you have a 10x more compute, you should increase model size by about 5x and data size by only 2x. ([Kaplan et al., 2020](#))

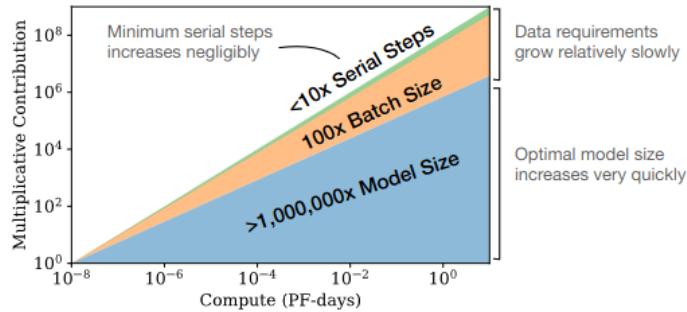


Figure 1.23 : OpenAIs initial paper on scaling laws stated that for optimally compute-efficient training, most of the increase should go towards increased model size. A relatively small increase in data is needed to avoid reuse. ([Kaplan et al., 2020](#))

DeepMind's scaling law update in 2022 . DeepMind found that most large language models were actually significantly overparameterized for the amount of data they were trained on. The Chinchilla scaling laws showed that for optimal performance, models should be trained on approximately 20 times more data tokens than they have parameters. This meant that many leading models could have achieved better performance with smaller sizes, but with more data. They were called chinchilla scaling laws because the laws were demonstrated using a model called Chinchilla. This was a 70B parameter model trained on more data, which outperformed much larger models like Gopher (280B parameters) despite using the same amount of compute. So by these laws, for optimal performance, you should increase model size and dataset size in roughly equal proportions - if you get 10x more compute, you should make your model 3.1x bigger and your data 3.1x bigger. ([Hoffmann et al., 2022](#))

The Broken Neural Scaling Laws (BNSL) update in 2023 . Ongoing research showed that performance doesn't always improve smoothly - there can be sharp transitions, temporary plateaus, or even periods where performance gets worse before getting better. Examples of this include things like "Grokking", where models suddenly achieve strong generalization after many training steps, or deep double descent, where increasing model size initially hurts then helps performance. Rather than simple power laws, BNSL uses a more flexible functional form that can capture these complex behaviors. This allows for more accurate predictions of scaling behavior, particularly around discontinuities and transitions. Scaling laws are a good baseline, but discontinuous jumps in capabilities and abrupt step changes are still possible. ([Caballero et al., 2023](#))

How do training and inference scaling differ? Training scaling involves using more compute during initial model training by using larger models, training for longer, or using bigger datasets. Another way that we might not be accounting for using scaling laws, is called inference time scaling. This instead uses more compute at runtime through techniques like chain-of-thought prompting, repeated sampling, or tree search. For example, you can either train a very large model that generates high-quality outputs directly, or train a smaller model that achieves similar performance by using more computation to think through problems step by step at inference time.

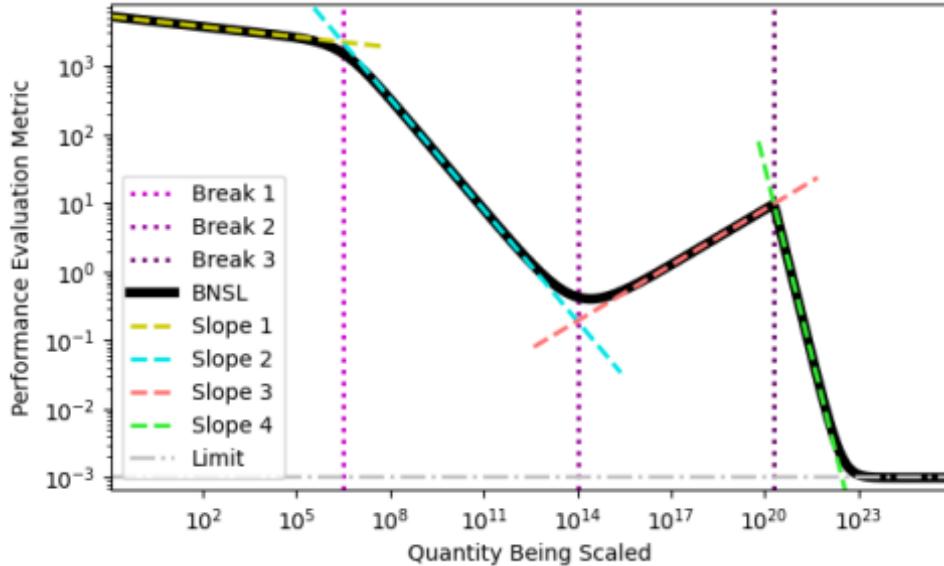


Figure 1.24 : A Broken Neural Scaling Law example (dark black solid line) (with 3 breaks where purple dotted lines intersect with dark black solid line) that contains 4 individual power law segments (where the dashed lines that are yellow, blue, red, and green overlap with the dark black solid line). The 1st and 2nd break are very smooth; the 3rd break is very sharp. ([Caballero et al., 2023](#))

1.5.3 Scaling Hypothesis

Strong Scaling Hypothesis ([Branwen, 2020](#))

The strong scaling hypothesis proposes that simply scaling up current foundation model architectures with more compute and data will be sufficient to reach transformative AI capabilities and potentially even ASI.

What is the strong scaling hypothesis? This view suggests we already have all the fundamental components needed - it's just a matter of making them bigger, following established scaling laws. ([Branwen, 2020](#)) There is heated debate around this hypothesis and we can't possibly cover every argument. We can give you a slight overview in the next few paragraphs.

Proponents include OpenAI ([OpenAI, 2023](#)), Anthropic's CEO Dario Amodei ([Amodei, 2023](#)), Conjecture ([Conjecture, 2023](#)), DeepMind's safety team ([DeepMind, 2022](#)), and others. According to the DeepMind team, there are "*not many more fundamental innovations needed for AGI. Scaled-up deep learning foundation models with RL from human feedback (RLHF) fine-tuning [should suffice]*" ([DeepMind, 2022](#)).

What are the key arguments supporting the strong scaling hypothesis? The most compelling evidence for this view comes from empirical observations of progress in recent years. Researchers have been developing algorithms that follow the bitter lesson's principle for many years (focusing on general methods that leverage compute effectively). But even when researchers have developed sophisticated algorithms following the bitter lesson's principles, these improvements still only account for 35% of performance gains in language models in 2024, with the remaining 65% coming purely from increased scale in compute and data ([Ho et al., 2024](#)). Basically, even when our algorithmic improvements align perfectly with the bitter lesson, they're still far less important than raw scaling.

The emergence of unexpected capabilities provides another powerful argument for strong scaling. We've seen previous generations of foundation models demonstrate remarkable abilities that weren't explicitly trained for, like programming for example. This emergent behavior hints that it is not impossible for higher-order metacognitive abilities like causal reasoning to similarly emerge simply as a function of scale. We see that bigger models become increasingly sample efficient - they require fewer examples to learn

new tasks. This improved efficiency with scale suggests that scaling up further could eventually lead to human-like few-shot learning capabilities, which is a precursor for TAI and ASI. Finally, these models also appear to be capable of learning any task that can be expressed through their training modalities. Right now this is text for LLMs but there is a clear path forward to multimodal LMMs. Since text can express virtually any human-comprehensible task, scaling up language understanding might be sufficient for general intelligence.

What are the key arguments against the strong scaling hypothesis? Recent research has also identified several challenges to the strong scaling hypothesis. The most immediate is data availability - language models will likely exhaust high-quality public text data between 2026 and 2032 ([Villalobos et al., 2024](#)). While synthetic data might help address this limitation, it's unclear whether it can provide the same quality of learning signal as organic human-generated content. Alternatively, we still have a lot of multi-modal data left to train on (like YouTube videos) despite running out of text data.

A more fundamental challenge comes from the way these models work. LLMs are fundamentally "interpolative databases" (or stochastic parrots , or a variety of other similar terms). The point being that they just build up a vast collection of vector transformations through pre-training. While these transformations become increasingly sophisticated with scale, critics argue there's a fundamental difference between recombining existing ideas and true synthesis - deriving novel solutions from first principles. However, this is not an airtight case against strong scaling. This could simply be a limitation of current scale - a larger model trained on multimodal data might learn to handle any new novel situation simply as a recombination of previously memorized patterns. So, it is unclear if template recombination actually does have an upper bound.

Weak Scaling Hypothesis ([Branwen, 2020](#))

The weak scaling hypothesis proposes that even though scale will continue to be the primary driver of progress, we will also need targeted architectural and algorithmic improvements to overcome specific bottlenecks.

What is the weak scaling hypothesis? Given these challenges, a weaker version of the scaling hypothesis has also been proposed. According to the weak scaling hypothesis even though scale will continue to be the primary driver of progress, we will also need targeted architectural and algorithmic improvements to overcome specific bottlenecks. These improvements wouldn't require fundamental breakthroughs, but rather incremental enhancements to better leverage scale. ([Branwen, 2020](#)) Similar to the strong scaling hypothesis, the weak one is also contentious and debated. We can provide a few of the results arguing both for and against this outlook.

LeCun's H-Jepa architecture ([LeCun, 2022](#)), or Richard Sutton's Alberta Plan ([Sutton, 2022](#)) are notable plans that might support the weak scaling hypothesis.

What are the key arguments supporting the weak scaling hypothesis? The arguments for strong scaling, like algorithmic improvements only contributing 35% of performance gains in language models can also double count for weak scaling. Since one third is still a non-trivial role to play in capabilities improvement. Some more empirical observations also support weak scaling. Like hardware support for lower-precision calculations, which provided order-of-magnitude performance improvements for machine learning workloads ([Hobbhahn et al., 2023, "Trends in Machine Learning Hardware"](#)). These kinds of targeted improvements don't change the fundamental scaling story but rather help us better leverage available resources. Similarly, the Chinchilla results showed that many models were poorly optimized along the different variables that contribute to capabilities. This suggests that there is still room for improvement through better scaling strategies rather than fundamental breakthroughs. ([Hoffmann et al., 2022](#))

What if neither the weak nor the strong scaling hypothesis is true? Essentially, both the scaling laws (which only predict foundation model capabilities) and most debates around "scale is all you need"

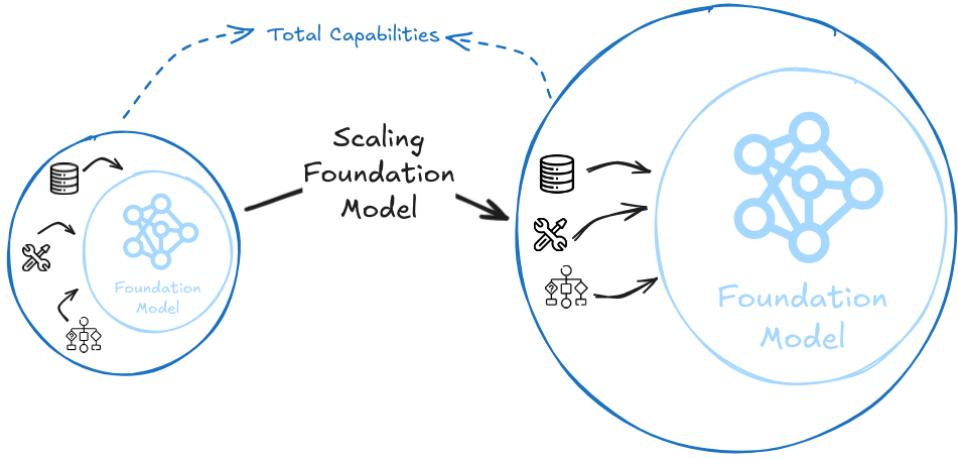


Figure 1.25 : Augmentation/Scaffolding stays constant, but if the scaling hypothesis, weak or strong, is true, then capabilities will keep improving just by scaling.

often miss other aspects of AI development that happen outside the scope of what scaling laws can predict. They don't account for improvements in AI "scaffolding" (like chain-of-thought prompting, tool use, or retrieval), or combinations of multiple models working together in novel ways. Debates around the scaling laws only tell us about the capabilities of a single foundation model trained in a standard way. For example, by the strong scaling hypothesis we can reach TAI by simply scaling up the same foundation model until meta cognitive abilities emerge. But even if scaling stops, halting capabilities progress on the core foundation model (in either a weak or a strong way), the external techniques that leverage the existing model can still continue advancing.

Think of foundation models like LLMs or LMMs as simply one transistor. Alone they might not be able to do much, but if we combine enough transistors we end up with all the capabilities of a supercomputer. Many researchers think that this is a core element where future capabilities will come from. It is also referred to as "unhobbling" (Aschenbrenner, 2024), "schlep" (Cotra, 2023) and various other terms, but all of them point to the same underlying principle - raw scaling of single model performance is only one part of overall AI capability advancement.

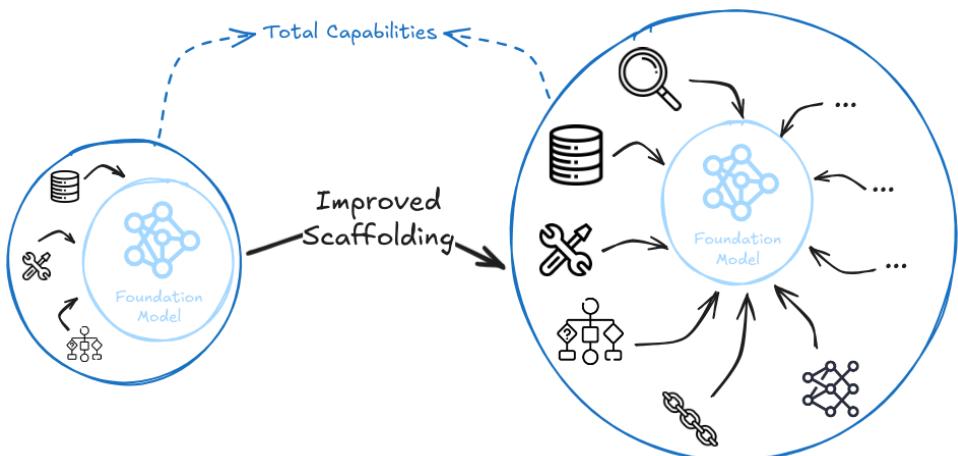


Figure 1.26 : Even if we see no improvements in model scale, other elicitation techniques and scaffolding can keep improving. So overall capabilities keep growing. Realistically, the future is probably going to see both improvement due to scaffolding and scale. So for now, there does not seem to be an upper limit on improving capabilities as long as either one of the two holds.

We go deeper into the arguments and counterarguments for all views on scaling foundation models in the

appendix.

Argument: Against scaling - Memorization vs Synthesis

!!! warning "This is a slightly deeper dive into trying to understand the argument of both sides if you are interested. It is slightly more technical, and can be safely skipped."

When we talk about LLMs as "interpolative databases", we're referring to how they store and manipulate vector programs - these shouldn't be confused with traditional computer programs like python or C++. These templates, or vector programs are transformations in the model's embedding space. Early work on embeddings showed simple transformations (like king - man + woman = queen), but modern LLMs can store millions of much more complex transformations. But due to a function of scale, LLMs can now store arbitrarily complex vector functions — so complex, in fact, that researchers found it more accurate to refer to them as vector programs rather than functions.

So what's happening in LLMs is that they build up a vast database of these vector programs through pre-training. When we say they're doing "template matching" or "memorization", what we really mean is that they're storing millions of these vector transformations that they can retrieve and combine with each prompt.

So the deciding question for/against strong (and even weak scaling) becomes - Is this type of template program combination enough to reach general intelligence. In other words can program synthesis be approximated using recombinations of enough templates (also called abstractions and many other words but the key idea is the same)?

People who argue against this say that no matter how numerous or sophisticated, are fundamentally different from true program synthesis. True program synthesis would mean deriving a new solution from first principles - not just recombining existing transformations. There are some empirical observations to support this view. Like the Caesar cipher example: "LLMs can solve a Caesar cipher with key size 3 or 5, but fail with key size 13, because they've memorized specific solutions rather than understanding the general algorithm" ([Chollet, 2024](#)). Or alternatively, the "reversal curse" which shows that even SOTA language models in 2024 cannot do reverse causal inference - if they are trained on "A is B" they fail to learn "B is A" ([Berglund et al., 2023](#))

But this does still not seem to completely invalidate scaling as of yet. If we scale up the size of the program database and cram more knowledge and patterns into it, we are going to be increasing its performance. ([Chollet, 2024](#)) Both sides of the debate agree on this. So this suggests the real issue isn't whether template recombination has an obvious absolute upper bound, but whether it's the most efficient path to general intelligence. Program synthesis might achieve the same capabilities with far less compute and data by learning to derive solutions rather than memorizing patterns.

1.6 Forecasting

In previous sections, we explored how foundation models leverage computation through scaling laws and the bitter lesson. But how can we actually predict where AI capabilities are headed? This section introduces key forecasting methodologies that help us anticipate AI progress and prepare appropriate safety measures.

Why should we care about forecasting? Forecasting AI progress is critical for AI safety work. The timeline to transformative AI shapes everything from research priorities to governance frameworks – if we expect transformative AI within 5 years versus 50 years, this dramatically changes which safety approaches are viable. For example, if we expect rapid progress, we might need to focus on safety measures that can be implemented quickly rather than long-term theoretical research. Additionally, understanding likely development trajectories helps us anticipate specific capabilities and prepare targeted safety measures before they emerge. This is especially critical given the potential for sudden capability jumps, especially

in dangerous capabilities like malware generation or deception.

Initial Forecast: Before we explore more sophisticated methods, make an initial prediction: When do you think we'll see transformative AI? Keep this forecast in mind as we examine different forecasting approaches.

1.6.1 Methodology

How do we convert beliefs into probabilities and forecasts? We need some ways to actually convert beliefs like "I think AGI is likely this decade" into precise probability estimates. One way we can do this is by decomposition - breaking down complex beliefs into smaller, measurable components and analyzing relevant data. Rather than directly estimating the year in which transformative AI emerges, we can start by separately forecasting things like compute growth, algorithmic progress, and hardware limitations, and then combine these estimates ([Zhang, 2024](#)). This decomposition approach helps us ground predictions in observable trends rather than relying purely on intuitions. So, using this approach there are two main techniques we need to discuss - zeroth-order forecasting for establishing baselines, and first-order forecasting for understanding trajectories of change.

What are reference classes and why do they matter? When analyzing each component of our decomposed forecast, we need relevant historical examples to inform our predictions. This is where reference classes come in - they are categories of similar historical situations we can use to make predictions. For AI development, relevant reference classes might include things like previous technological revolutions (like the industrial or computer revolution), other optimization systems (like biological evolution or economies), or the impact of rapid scientific advances (like CRISPR or mRNA vaccines). The basic point is that they should be meaningfully analogous to what you're trying to predict, but they don't have to be from the same exact category.

What is zeroth-order forecasting? The simplest forecasting approach starts with recognizing that tomorrow often looks pretty close to today. Zeroth-order forecasting uses reference classes - looking at 3-5 similar historical examples and using their average as a baseline prediction. Rather than trying to identify trends or make complex projections, it assumes recent patterns will continue. ([Steinhardt, 2024](#)) These examples can be from different reference classes. For a concrete example of using multiple reference classes in AI safety forecasting: Suppose we want to predict how quickly advanced AI systems might transition from "safe" to "potentially dangerous" capabilities. We could look at:

- How long it took language models to go from generating basic text to being able to plan multi-step deception (an AI-specific reference point)
- How quickly nuclear technology transitioned from peaceful to military applications (a dual-use technology reference point)
- How rapidly biological techniques like CRISPR went from lab discovery to widespread use requiring safety protocols (a biosafety reference point)

Looking at these examples collectively might suggest that dangerous capabilities often emerge within 2-5 years of the underlying technical breakthroughs, which could inform how urgently we need to develop safety measures. Major shifts in development patterns tend to be rare, making recent history a decent baseline predictor of the near future. This doesn't mean changes never happen – but it does mean that deviating from recent patterns requires strong evidence.

What is first-order forecasting? While zeroth-order forecasting uses historical examples from various reference classes as direct predictors, first-order forecasting attempts to identify and project forward patterns in the direct historical data of AI development. In AI, we see some pretty consistent exponential growth patterns. The compute used in frontier models has grown by 4.2x annually since 2010, training datasets have expanded by approximately 2.9x per year, and hardware performance improves by roughly 1.35x every year through architectural advances ([Epoch AI, 2023](#)). First-order forecasting tries to identify

these kinds of patterns and project them forward. This is the approach taken by most systematic AI forecasting work today, including Epoch AI's compute-centric framework and Ajeya Cotra's biological anchors. However, it's worth keeping in mind that even though these trends have been remarkably consistent, they can't continue indefinitely. Physical, thermodynamic, or economic constraints will eventually limit growth. The key question is: when do these limits become relevant? We will explore this in the next section on the compute centric framework.

How do we combine different forecasts? Multiple forecasting approaches often give us different predictions – zeroth-order might suggest one timeline while trend extrapolation indicates another. Just like we can average out over the opinions of many experts, we can integrate these predictions to get a hopefully more accurate picture. One approach is to model each forecast as a probability distribution and combine them using mixture models ([Steinhardt, 2024](#)). For example, if zeroth-order forecasting suggests 3-4 years between major breakthroughs based on recent history, while trend extrapolation points to 1.5-2 years based on compute growth, a combined model might predict 2-3 years but with wider confidence intervals to account for uncertainty in both approaches.

What about situations with limited data or limited reference classes? While decomposition, reference classes and trend analysis form the backbone of AI forecasting, we sometimes face questions where direct data is limited or no clear reference classes exist. For instance, predicting the societal impact of advanced AI systems or forecasting novel capabilities that haven't been demonstrated before. In these cases, we often turn to expert judgment and superforecasters. An advantage of expert forecasting is the ability to integrate qualitative insights that might be missed by pure trend analysis. For example, experts might notice early warning signs of diminishing returns or identify emerging technical approaches that could accelerate progress. This balanced use of both data-driven methods and expert judgment is especially important for AI safety work. While we should ground our predictions in empirical trends whenever possible, we also need frameworks for reasoning about unprecedented developments and potential discontinuities in progress.

How far do empirical findings generalize? There's an ongoing debate about how much we can trust current trends to predict future AI development. Some researchers argue that empirical findings in AI generalize surprisingly far - that patterns we observe today will continue to hold even as systems become more capable ([Steinhardt, 2022](#)). However, our track record with forecasting suggests we should be cautious. When superforecasters predicted MATH dataset accuracy would improve from 44% to 57% by June 2022, actual performance reached 68% - a level they had rated extremely unlikely. Shortly after, GPT-4 achieved 86.4% accuracy. There are a couple of more examples of LLMs surprising most forecasters and experts on certain benchmarks. ([Cotra, 2023](#)).

This pattern of underestimating progress suggests that while empirical trends provide valuable guidance, they may not capture all the dynamics of AI development. Prior to GPT-3, many experts believed tasks like complex reasoning would require specialized architectures. The emergence of these capabilities from scaling alone shows how systems can develop unexpected abilities simply through quantitative improvements. This has critical implications for both forecasting and governance - we need frameworks that can adapt to capabilities emerging faster or differently than current trends suggest.

How does this help us predict transformative AI? These forecasting fundamentals help us critically evaluate claims about AI timelines and takeoff scenarios. When we encounter predictions about discontinuous progress or smooth scaling, we can ask: What trends support this view? What reference classes are relevant? How have similar forecasts performed historically? This systematic approach helps us move beyond intuition to make more rigorous predictions about AI development trajectories.

1.6.2 Biology Inspired Forecasting

What are Biological anchors? Biological anchors are a forecasting technique. To find a reference class, assume that the human brain is indicative of general intelligence. This means we can treat it as a proof of concept. Whatever "amount of compute" it takes to train a human being, might be roughly the same amount it should take to train a TAI. The biological anchors approach estimates the compute required for AI to reach a level of intelligence comparable to humans, outlined through several steps:

- First, assess how much computation the human brain performs, translating this into a quantifiable

measure similar to computer operations in FLOP/s.

- Second, estimate the amount of computation needed to train a neural network to match the brain's inferential capacity, adjusting for future improvements in algorithmic efficiency.
- Third, examine when it would be feasible to afford such vast computational resources, taking into account the decreasing cost of compute, economic growth, and increasing investment in AI.
- Finally, by analyzing these factors, we can predict when it might be economically viable for AI companies to deploy the necessary resources for developing TAI.

Determining the exact computational equivalent for the human brain's training process is complex, leading to the proposal of six hypotheses, collectively referred to as "biological anchors" or "bioanchors." Each anchor has a different weighting contributing to the overall prediction.

Evolution Anchor: Total computational effort across all evolutionary history.

Lifetime Anchor: Brain's computational activity from birth to adulthood (0-32).

Neural Network and Genome Anchors: Various computational benchmarks based on the human brain and genome to gauge the scale of parameters needed for AI to achieve general intelligence.

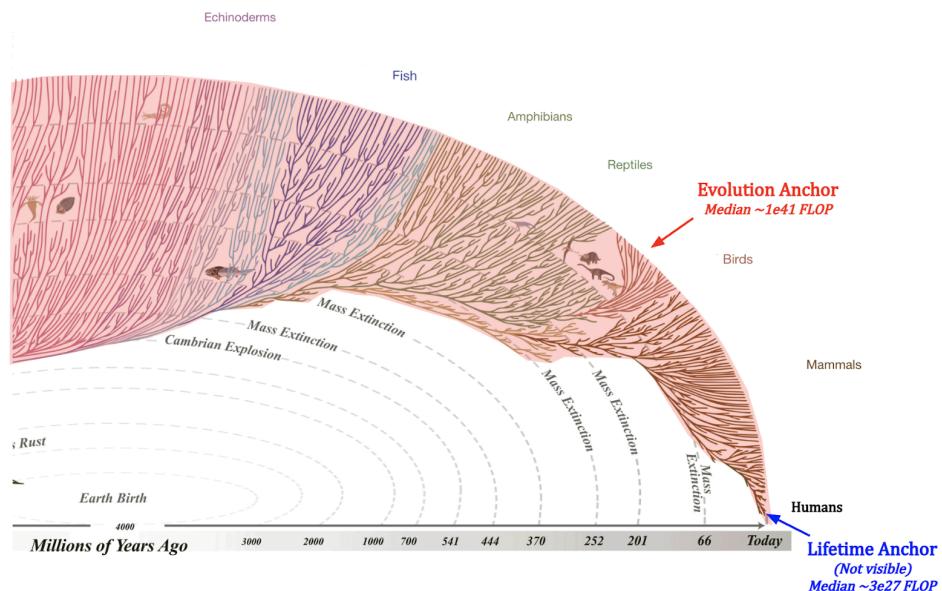


Figure 1.27 : A diagram showing the different starting points that we could use within the Bio anchors report to calculate the amount of effective compute required for human level transformative AI. (Ho, 2022)

Forecasting with Biological Anchors . By integrating these anchors with projections of future compute accessibility, we can outline a potential timeline for TAI. This method aims to provide a "soft upper bound" on TAI's arrival rather than pinpointing an exact year, acknowledging the complexity and unpredictability of AI development. (Karnofsky, 2021) The following image gives an overview of the methodology.

Affordability of compute. The costs related to bio anchors are calculated by considering three different factors: Algorithmic progress, compute price estimates, and willingness to spend on machine learning. The report considers a doubling in algorithmic efficiency every 2-3 years. As for prices, Cotra assumes cost decreases over time, halving every 2.5 years, and further expects this to level off after 6 orders of magnitude. Cotra assumes that the willingness to spend on machine learning training runs should be

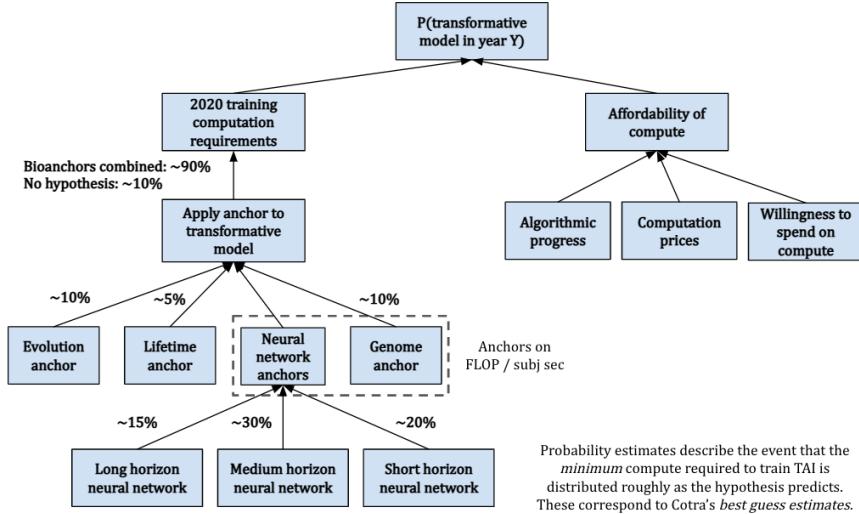


Figure 1.28 : The biological anchor model ([Ho, 2022](#))

capped at 1% of the GDP of the largest country, referencing previous case studies with megaprojects (e.g. the Manhattan Project), and should follow a doubling time of 2 years after 2025. The main uncertainty is whether or not existing trends are going to persist more than several years into the future. For example, Epoch found that OpenAI’s AI and Compute investigation ([OpenAI, 2018](#)) was too aggressive in its findings for compute growth. ([Ho et al., 2022](#)) This suggests taking caution when interpreting the forecasts made by the Bio Anchors report.

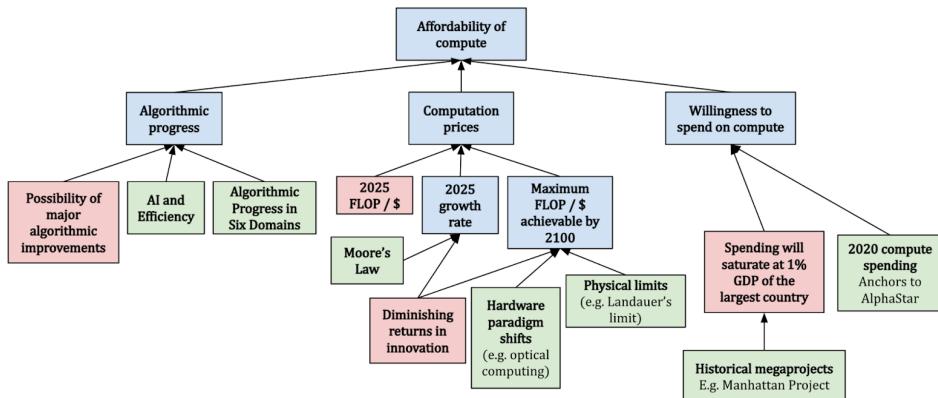


Figure 1.29 : Affordability of compute ([Ho et al., 2022](#))

The following graph gives an overview of the findings. Overall, the graph takes a weighted average of the different ways that the trajectory could flow. This gives us an estimate of a >10% chance of transformative AI by 2036, a 50% chance by 2055, and an 80% chance by 2100. In 2022 a two-year update on the author’s (Ajeya Cotra) timelines was published. The updated timelines for TAI are 15% probability by 2030, 35% probability by 2036, a median of 2040, and a 60% probability by 2050. ([Cotra, 2022](#))

Criticisms. The Biological Anchors framework provides a unique perspective but it’s also crucial to recognize its limitations and the broader debates it sparks within the AI research community. It is not universally accepted as the primary predictive tool among all ML scientists or alignment researchers.

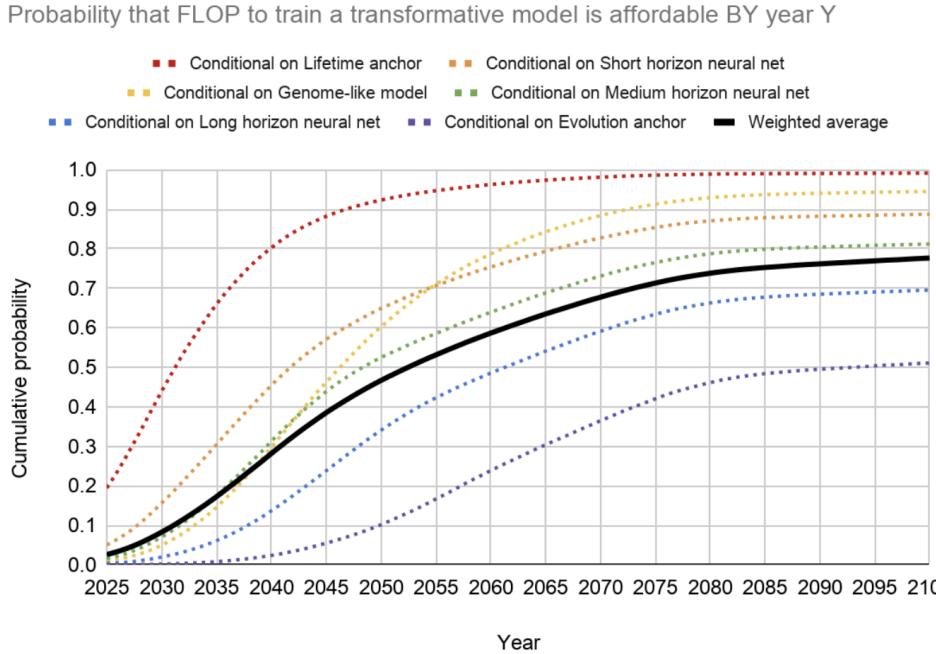


Figure 1.30 : Results from the biological anchor model for different anchors ([Karnofsky, 2021](#))

Platt's Law is a generalized observation named after Charles Platt. It is used to highlight a historical pattern where the estimated arrival of AGI (Artificial General Intelligence) consistently seems to be "just 30 years away".

Vernor Vinge at 1993 NASA speech. ([Yudkowsky, 2021](#))

Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.

Yudkowsky notes that Platt's law seems to fit remarkably well with the prediction made by the Biological Anchors report in 2020. As the statistical aphorism goes: "All models are wrong, but some are useful".

So to get a complete picture of how biological anchors were received, here are some of the criticisms of the Biological Anchors report:

- **Challenges Beyond Compute** : While Biological Anchors highlight compute power as a critical factor for AI development, they may oversimplify the complexity of achieving transformative AI. Factors beyond sheer computing capacity, such as algorithm design, data availability, and the intricacies of learning environments, play pivotal roles. It's overly simplistic to equate the future of AI solely with compute availability, as transformative AI development encompasses more nuanced challenges like algorithm innovation and data accessibility. ([Karnofsky, 2021](#))
- **Potential for Rapid Advancements** : Critics of the Biological Anchors method, such as Eliezer Yudkowsky, emphasize the unpredictability of AI progress and the potential for breakthroughs that could dramatically alter AI capabilities without adhering strictly to computational benchmarks derived from biology. These critiques highlight the importance of considering a range of factors and potential shifts in AI development paradigms that could accelerate progress beyond current forecasts. ([Karnofsky, 2021](#))
- **Purpose and Misinterpretation** : The Biological Anchors approach is intended to provide boundary estimates for AI development timelines rather than precise predictions. Misinterpretations may arise from expecting the method to offer specific year forecasts, whereas its goal is to outline

possible upper and lower bounds, acknowledging the significant uncertainties in AI development. ([Karnofsky, 2021](#))

- **Contemplating Paradigm Shifts :** The AI field's history suggests that major paradigm shifts and technological breakthroughs could substantially impact development timelines. While deep learning currently dominates AI advances, the possibility of new, transformative methodologies emerging remains open, challenging the assumption that current trends will roughly maintain its current growth trajectory into the future.

This is not an exhaustive list of all the criticisms but it serves to highlight the complexity of forecasting AI's future.

1.7 Takeoff

What are takeoff speeds? Before we can talk about different types of AI takeoff, we need to understand what takeoff speed even means. Takeoff speed refers to how quickly AI systems become dramatically more powerful than they are today and cause major societal changes. This is related to, but distinct from, AI timelines (how long until we develop advanced AI). While timelines tell us when transformative AI might arrive, takeoff speeds tell us what happens after it arrives - does AI capability and impact increase gradually over years, or explosively over days or weeks?

How can we think about different takeoff speeds? When analyzing different takeoff scenarios, we can look at several key factors:

- **Speed** : How quickly do AI capabilities improve?
- **Continuity** : Do capabilities improve smoothly or in sudden jumps?
- **Homogeneity** : How similar are different AI systems to each other?
- **Polarity** : How concentrated is power among different AI systems?

Let's discuss each of these factors and what they tell us about potential takeoff scenarios. This will help us better understand the ongoing debate about how transformative AI might develop.

1.7.1 Speed

What is a slow takeoff? In a slow takeoff scenario, AI capabilities improve gradually over months or years. We can see this pattern in recent history - the transition from GPT-3 to GPT-4 brought significant improvements in reasoning, coding, and general knowledge, but these advances happened over several years through incremental progress. Paul Christiano describes slow takeoff as similar to the Industrial Revolution but "10x-100x faster" ([Davidson, 2023](#)). Terms like "slow takeoff" and "soft takeoff" are often used interchangeably.

In mathematical terms, slow takeoff scenarios typically show linear or exponential growth patterns. With linear growth, capabilities increase by the same absolute amount each year - imagine an AI system that gains a fixed number of new skills annually. More commonly, we might see exponential growth, where capabilities increase by a constant percentage, similar to how we discussed scaling laws in earlier sections. Just as model performance improves predictably with compute and data, slow takeoff suggests capabilities would grow at a steady but manageable rate. This might manifest as GDP growing at 10-30% annually before accelerating further.

The key advantage of slow takeoff is that it provides time to adapt and respond. If we discover problems with our current safety approaches, we can adjust them before AI becomes significantly more powerful. This connects directly to what we'll discuss in later chapters about governance and oversight - slow takeoff allows for iterative refinement of safety measures and gives time for coordination between different actors and institutions.

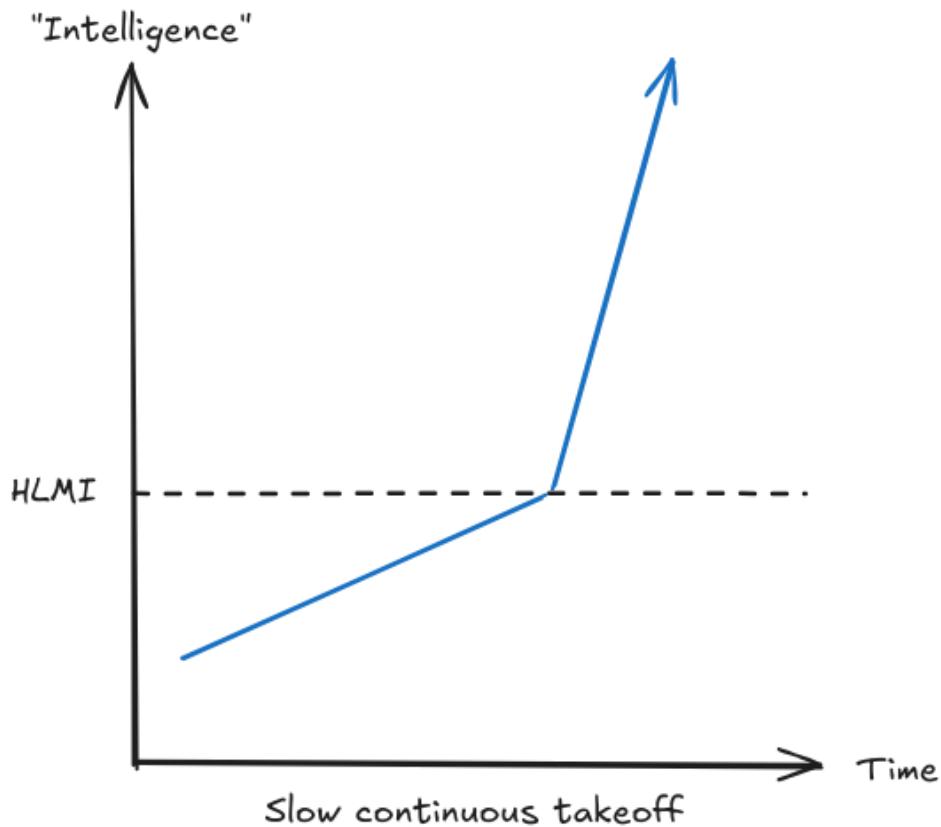


Figure 1.31 : An illustration of slow continuous takeoff. ([Martin & Eth, 2021](#))

What is a fast takeoff? Fast takeoff describes scenarios where AI capabilities increase dramatically over very short periods - perhaps days or even hours. Instead of the gradual improvement we saw from GPT-3 to GPT-4, imagine an AI system making that much progress every day. This could happen through recursive self-improvement, where an AI system becomes better at improving itself, creating an accelerating feedback loop.

Mathematically, fast takeoff involves superexponential or hyperbolic growth, where the growth rate itself increases over time. Rather than capabilities doubling every year as in exponential growth, they might double every month, then every week, then every day. This relates to what we discussed in the scaling section about potential feedback loops in AI development - if AI systems can improve the efficiency of AI research itself, we might see this kind of accelerating progress.

The dramatic speed of fast takeoff creates unique challenges for safety. As we'll explore in the chapter on strategies, many current safety approaches rely on testing systems, finding problems, and making improvements. But in a fast takeoff scenario, we might only get one chance to get things right. If an AI system starts rapidly self-improving, we need safety measures that work robustly from the start, because we won't have time to fix problems once they emerge.

Terms like "fast takeoff", "hard takeoff" and "FOOM" are often used interchangeably.

Why does takeoff speed matter for AI risk? The speed of AI takeoff fundamentally shapes the challenge of making AI safe. This connects directly to what we discussed about scaling laws and trends - if progress follows predictable patterns as our current understanding suggests, we might have more warning and time to prepare. But if new mechanisms like recursive self-improvement create faster feedback loops, we need different strategies.

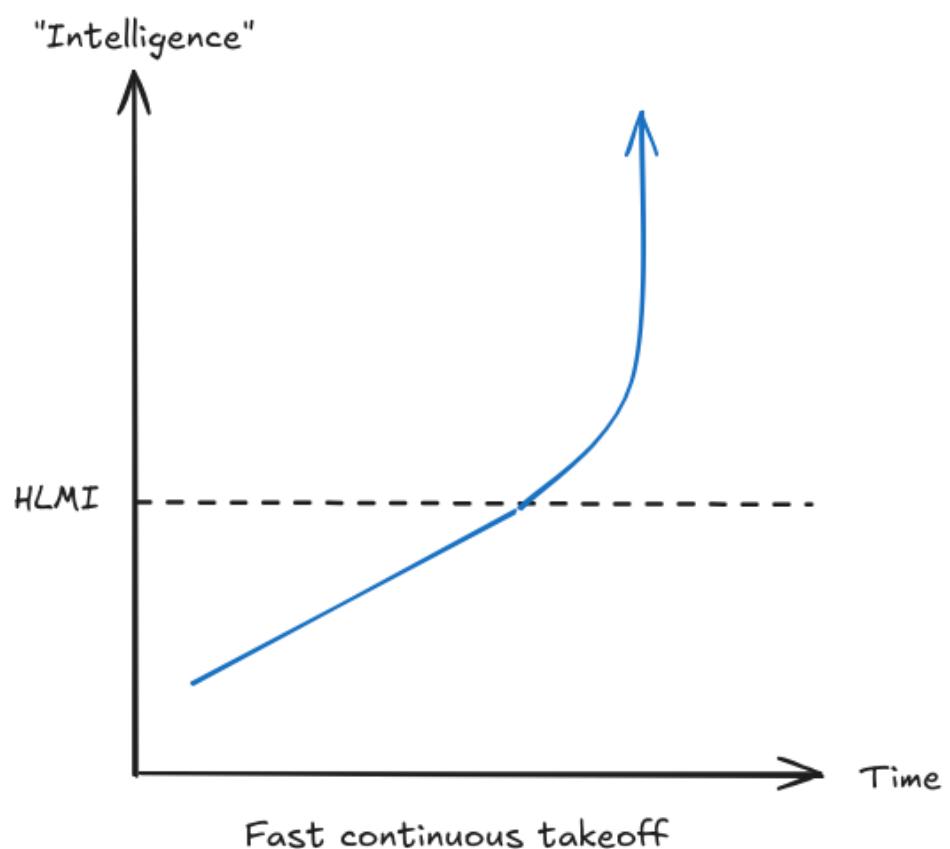


Figure 1.32 : An illustration of fast continuous takeoff, which is usually taken to mean superexponential or hyperbolic growth. The growth rate itself increases. ([Martin & Eth, 2021](#))

A concrete example helps illustrate this: Today, when we discover that language models can be jailbroken in certain ways, we can patch these vulnerabilities in the next iteration. In a slow takeoff, this pattern could continue - we'd have time to discover and fix safety issues as they arise. But in a fast takeoff, we might need to solve all potential jailbreaking vulnerabilities before deployment, because a system could become too powerful to safely modify before we can implement fixes.

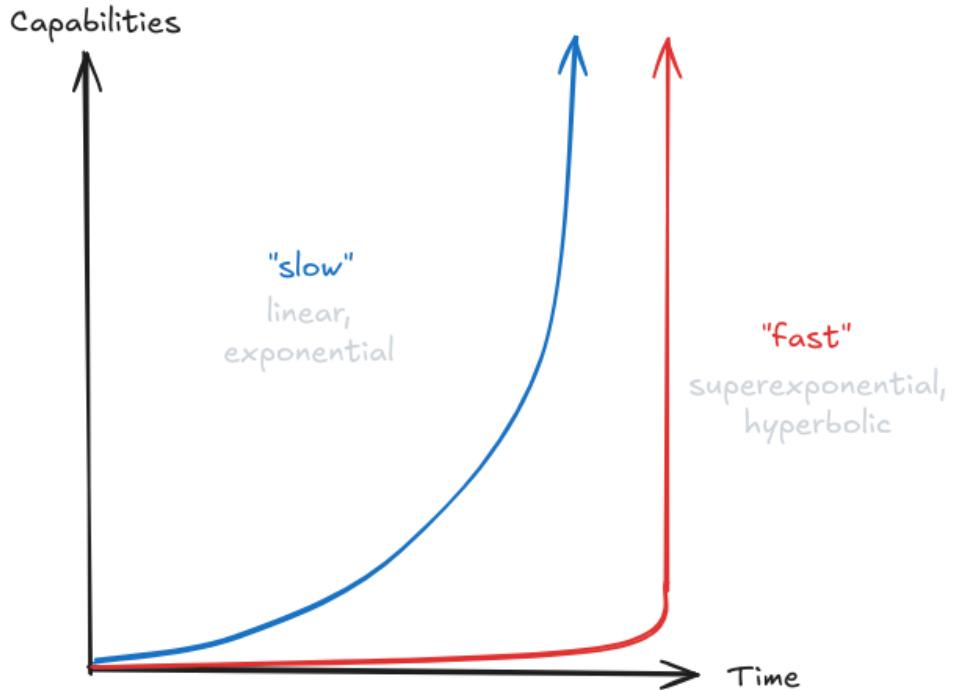


Figure 1.33 : Comparison of "slow" vs "fast" takeoff. Showcasing that while described as linguistically slower than fast, it is by no means "slow". ([Christiano, 2018](#))

1.7.2 Continuity

What is takeoff continuity? Continuity describes whether AI capabilities improve smoothly and predictably or through sudden, unexpected jumps. This is different from speed - even a fast takeoff could be continuous if the rapid progress follows predictable patterns, and a slow takeoff could be discontinuous if it involves surprising breakthroughs. Understanding continuity helps us predict whether we can extrapolate from current trends, like the scaling laws we discussed earlier, or if we should expect sudden departures from these patterns. So if you think of speed as a measure of how quickly the AI becomes superintelligent, continuity can be thought of as a measure of "surprise".

What is a continuous takeoff? In a continuous takeoff, AI capabilities follow smooth, predictable trends. The improvements we've seen in language models provide a good example - each new model tends to be somewhat better than the last at tasks like coding or math, following patterns we can roughly predict from scaling laws and algorithmic improvements. As we saw in the forecasting section, many aspects of AI progress have shown this kind of predictable behavior.

Continuous progress doesn't mean linear or simple progress. It might still involve exponential or even superexponential growth, but the key is that this growth follows patterns we can anticipate. Think of how GPT-4 is better than GPT-3, which was better than GPT-2 - each improvement was significant but not completely surprising given the increase in scale and improved training techniques.

A continuous takeoff suggests that current trends in scaling laws and algorithmic progress might extend even to transformative AI systems. This would give us more warning about upcoming capabilities and more ability to prepare appropriate safety measures. As we'll discuss in the governance chapter, even

though progress is fast, this kind of predictability makes it comparatively easier to develop and implement regulation before AI systems become extremely powerful or uncontrollable. Keeping in mind of course that comparatively easier does not mean "easy".

What is a discontinuous takeoff? A discontinuous takeoff involves sudden jumps in capability that break from previous patterns. Instead of steady improvements in performance as we add compute or data, we might see the emergence of entirely new capabilities that weren't predicted by existing trends. One hypothetical example would be if an AI system suddenly developed robust general reasoning capabilities after appearing to only handle narrow tasks - this would represent a discontinuity in the pattern of AI development.

Discontinuities could arise through various mechanisms. We might discover fundamentally new training approaches that are dramatically more efficient than current methods. Or, as we discussed in the scaling section, we might hit tipping points where quantitative improvements in scale lead to qualitative changes in capability. An AI system might even discover such improvements about itself, leading to unexpected jumps in capability.

The historical record provides some precedent for both continuous and discontinuous scientific progress. The development of nuclear weapons represented a discontinuous jump in explosive power, while improvements in computer processing power have followed more continuous trends. However, as we saw in the forecasting section, technological discontinuities have historically been rare, which some researchers cite as evidence favoring continuous takeoff scenarios.

The terms 'fast takeoff' and 'discontinuous takeoff' are often used interchangeably. However, the images below displaying different takeoff trajectories might help in clarifying the subtle differences between the concepts.

Why does continuity matter for AI safety? The continuity of AI progress has crucial implications for how we approach safety. In a continuous takeoff, we can more reliably test safety measures on less capable systems and be more confident they'll work on more advanced ones. We can also better predict when we'll need different safety measures and plan accordingly.

1.7.3 Similarity

What is takeoff homogeneity? Homogeneity describes how similar or different AI systems are to each other during the takeoff period. Will we see many diverse AI systems with different architectures and capabilities, or will most advanced AI systems be variations of the same basic design? This isn't just about technical diversity - it's about whether advanced AI systems will share similar behaviors, limitations, and safety properties. ([Hubinger, 2020](#))

What is a homogeneous takeoff? In a homogeneous takeoff, most advanced AI systems would be fundamentally similar. We can see hints of this pattern today - many current language models are based on the transformer architecture and trained on similar data, leading to similar capabilities and limitations. In a homogeneous takeoff, this pattern would continue. Perhaps most AI systems would be fine-tuned versions of a few base models, or different implementations of the same core breakthrough in AI design.

A key factor that could drive homogeneity is the sheer scale required to train advanced AI systems. If training transformative AI requires massive compute resources, as scaling laws suggest, then only a few organizations might be capable of training base models from scratch. Other organizations would build on these base models rather than developing entirely new architectures, leading to more homogeneous systems.

Homogeneous takeoff could be safer in some ways but riskier in others. If we solve alignment for one AI system, that solution might work for other similar systems. However, if there's a fundamental flaw in the common architecture or training approach, it could affect all systems simultaneously. It's like having a monoculture in agriculture - while easier to manage, it's also more vulnerable to shared weaknesses.

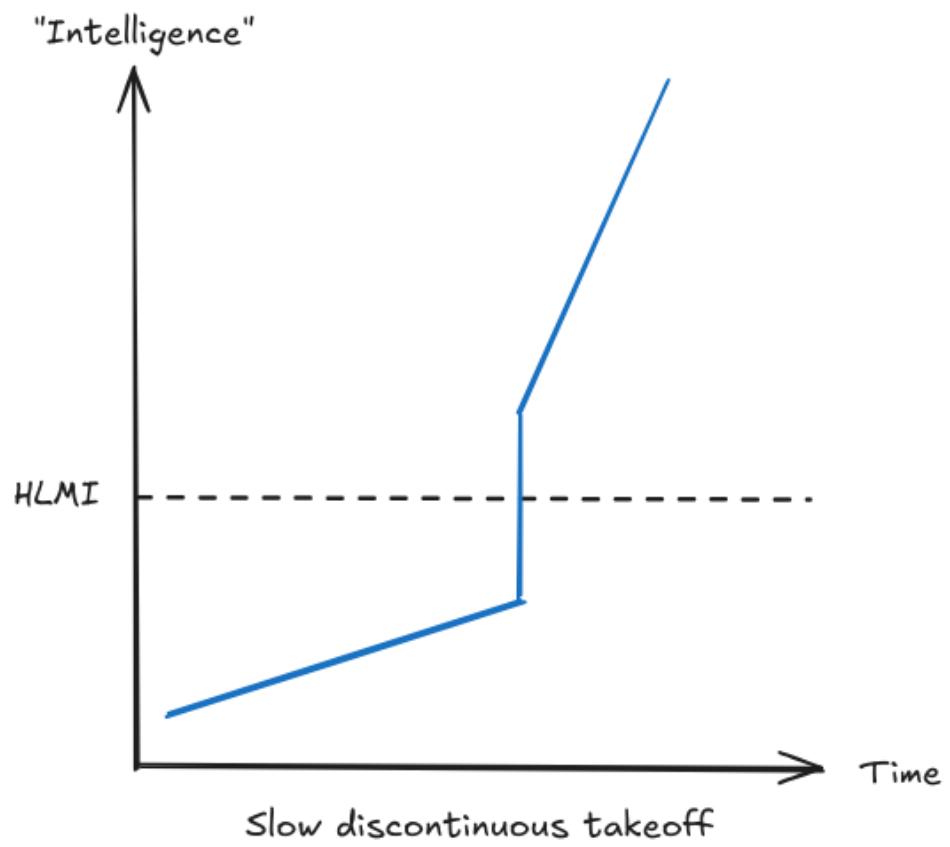


Figure 1.34 : One example illustration of slow discontinuous takeoff, where even though progress keeps increasing we might see sudden "jumps" in progress. ([Martin & Eth, 2021](#))

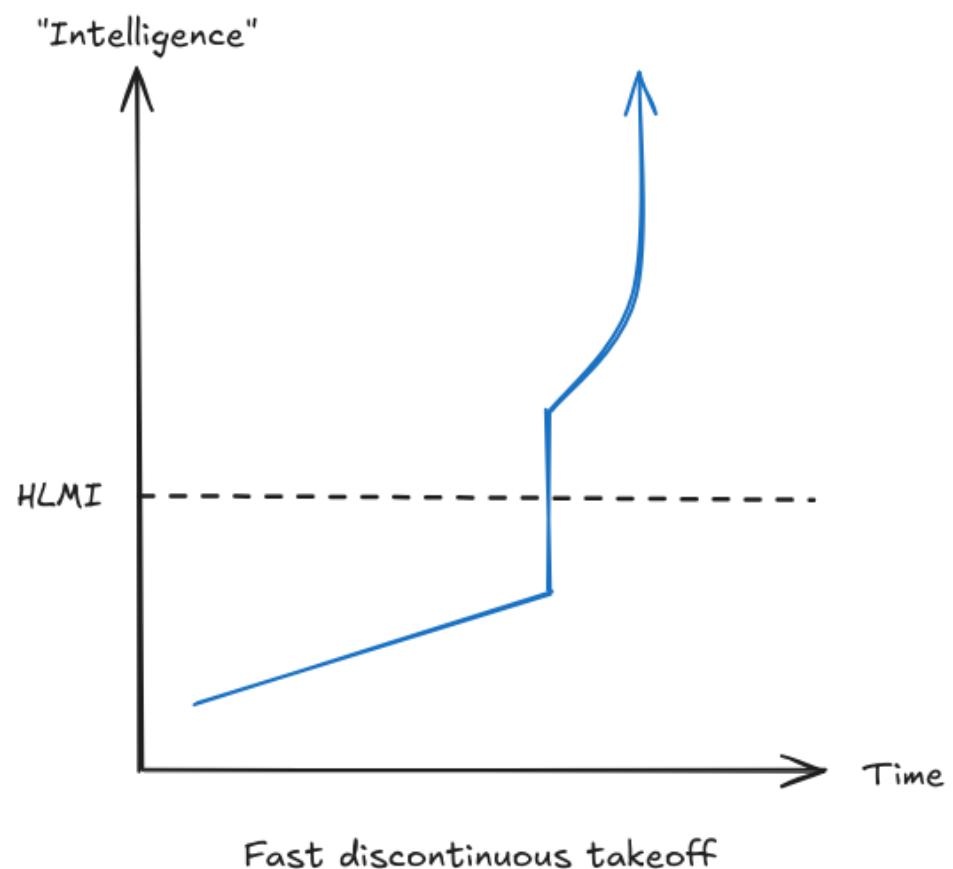


Figure 1.35 : One example illustration of fast discontinuous takeoff. Even though progress keeps accelerating, in addition to that we might also see sudden "jumps" in progress. ([Martin & Eth, 2021](#))

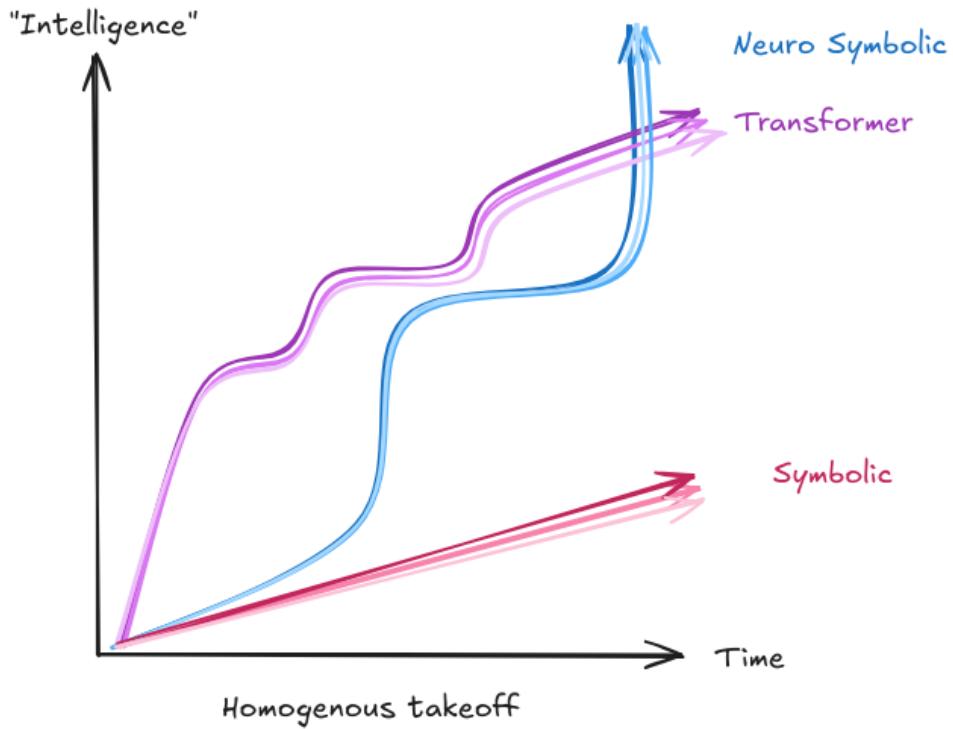


Figure 1.36 : An illustration of homogeneous takeoff. We can see multiple different overarching model architectures. The figure shows three in different colors. Within each architecture the takeoff is roughly the same due to similarity in design, regulations, and safety mitigations. **NOTE :** The curves here with architectures are purely illustrative, and are not meant to indicate predicted growth trajectories and comparisons between different architectures.

What is a heterogeneous takeoff? In a heterogeneous takeoff, we'd see significant diversity among advanced AI systems. Different organizations might develop fundamentally different approaches to AI, leading to systems with distinct strengths, weaknesses, and behaviors. Some might be specialized for specific domains while others are more general, some might be more transparent while others are more opaque, some might be more aligned with human values while others might not be. Competitive dynamics among AI projects could exacerbate diversity, as teams race to achieve breakthroughs without necessarily aligning on methodologies or sharing crucial information. As an example, we might have a future where AI becomes a strategic national asset, and AI development is closely guarded. In this environment, the pursuit of AI capabilities becomes siloed, each company or country would then employ different development methodologies, potentially leading to a wide range of behaviors, functionalities, and safety levels.

Heterogeneous takeoff creates different challenges for safety. We'd need to develop safety measures that work across diverse systems, and we couldn't necessarily apply lessons learned from one system to others. However, diversity might provide some protection against systemic risks - if one approach proves dangerous, alternatives would still exist.

How does takeoff homogeneity affect the broader picture? The degree of homogeneity during takeoff has significant implications for how transformative AI might develop. In a homogeneous scenario, progress might be more predictable but also more prone to winner-take-all dynamics. A heterogeneous scenario might be more robust against single points of failure but harder to monitor and control.

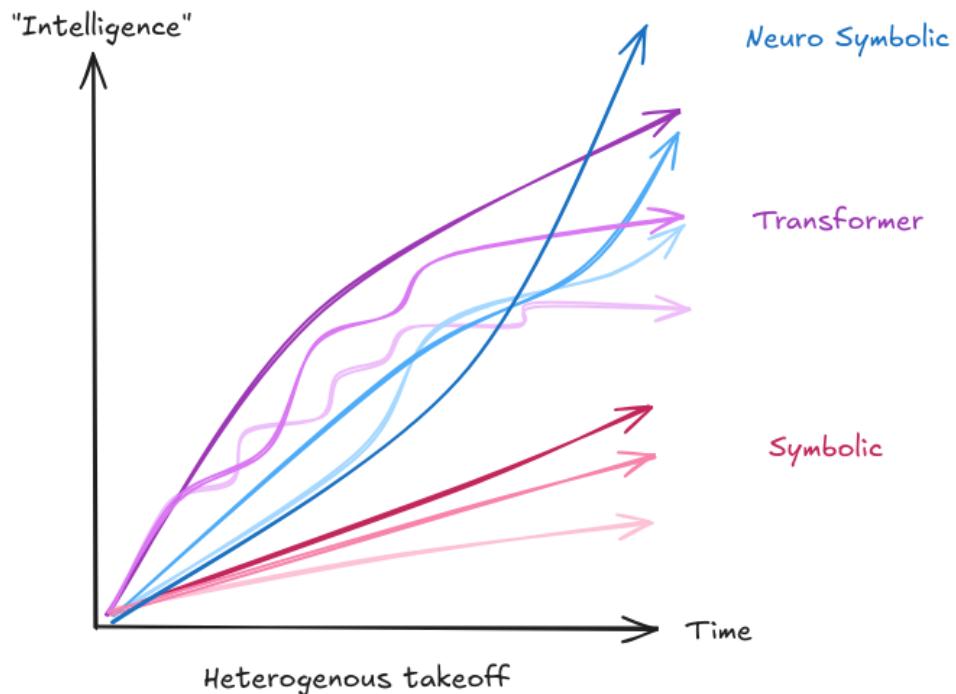


Figure 1.37 : One example of heterogeneous takeoff. We can see multiple different overarching model architectures. The figure shows three in different colors. Within each architecture the takeoff is different due to differences in design, regulations, and safety mitigations. **NOTE :** The curves here with architectures are purely illustrative, and are not meant to indicate predicted growth trajectories and comparisons between different architectures.

1.7.4 Polarity

What is takeoff polarity? Polarity describes whether power and capability becomes concentrated in a single AI system or organization, or remains distributed among multiple actors. In other words, will one AI system or group pull dramatically ahead of all others, or will multiple AI systems advance in parallel with comparable capabilities?

What is a unipolar takeoff? In a unipolar takeoff, one AI system or organization gains a decisive lead over all others. This could happen through a single breakthrough, exceptional scaling advantages, or recursive self-improvement. For example, if one AI system becomes capable enough to substantially accelerate its own development, it might rapidly outpace all other systems. The mathematics of training compute provide one path to a unipolar outcome. If a doubling of compute leads to reliable improvements in capability, then an organization that gets far enough ahead in acquiring compute could maintain or extend their lead. Their improved systems could then help them develop even better training methods, hardware, and attract investment creating a positive feedback loop that others can't match. But compute isn't the only path to unipolarity. A single organization might discover a fundamentally better training approach, or develop an AI system that's better at improving itself than at helping humans build alternatives. Once any actor gets far enough ahead, it might become practically impossible for others to catch up.

What is a multipolar takeoff? In a multipolar takeoff, multiple AI systems or organizations develop advanced capabilities in parallel. This could look like several large labs developing different but comparably powerful AI systems, or like many actors having access to similar AI capabilities through open source models or AI services. Today's AI landscape shows elements of multipolarity - multiple organizations can train large language models, and techniques developed by one lab are often quickly adopted by others. A multipolar takeoff might continue this pattern, with multiple groups maintaining similar capabilities even as those capabilities become transformative. A unipolar scenario raises concerns about the concentration of power, while a multipolar world presents challenges in coordination among diverse entities or AI systems.

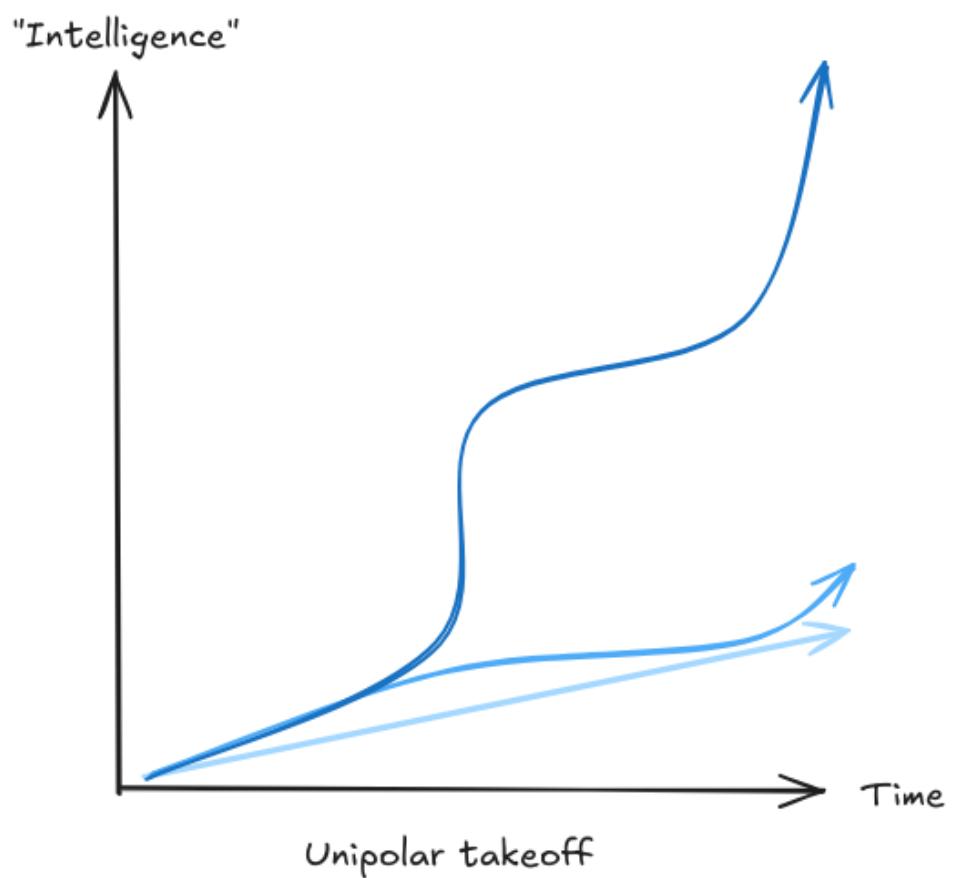


Figure 1.38 : An illustration of unipolar takeoff. One model (dark blue here) significantly outperforms all others.

Both unipolar and multipolar worlds have the potential for misuse of advanced AI capabilities by human actors.

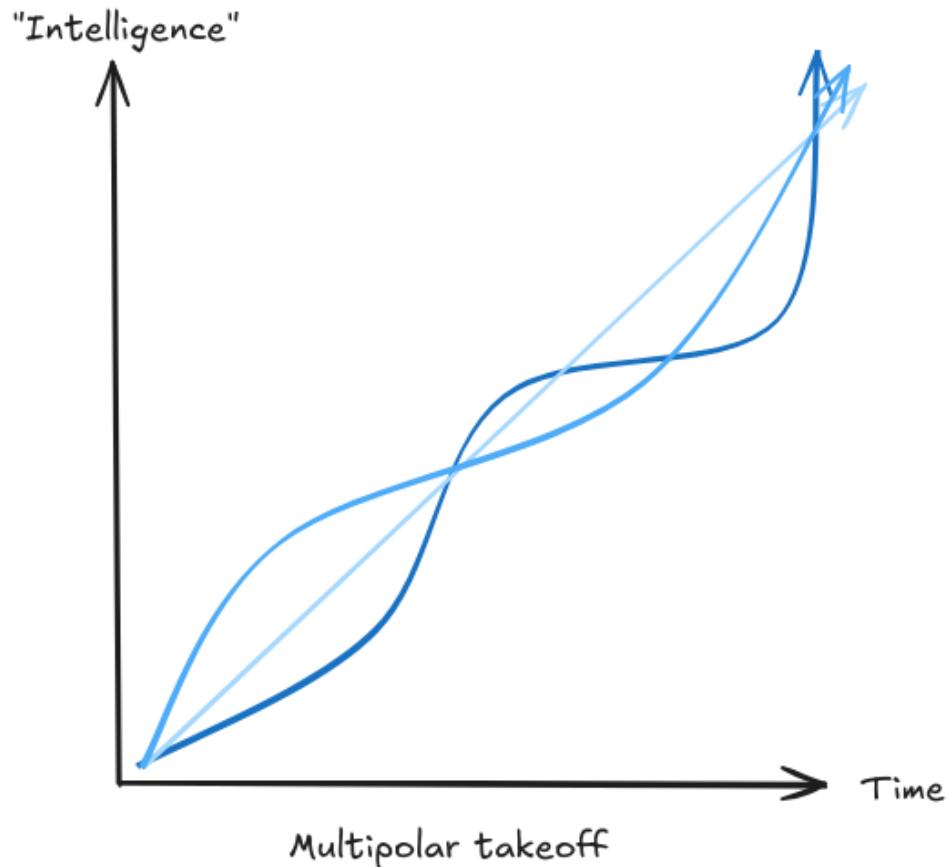


Figure 1.39 : An illustration of multipolar takeoff. No model significantly outperforms all others, and they all takeoff at a roughly competitive rate relative to each other.

Why does polarity matter? The polarity of takeoff has major implications for both safety risks and potential solutions. In a unipolar scenario, the actions and alignment of a single system or organization become crucial - they might gain the ability to shape the long-term future unilaterally. This concentrates risk in a single point of failure, but might also make coordination easier since fewer actors need to agree. A multipolar scenario creates different challenges. Multiple advanced systems might act in conflicting ways or compete for resources. This could create pressure to deploy systems quickly or cut corners on safety. There's also an important interaction between polarity and the other aspects of takeoff we've discussed. A fast takeoff might be more likely to become unipolar, as the first system to make rapid progress could quickly outpace all others. A slow takeoff might tend toward multipolarity, giving more actors time to catch up to any initial leads.

Factors Influencing Polarity . Several key elements influence whether takeoff polarity leans towards a unipolar or multipolar outcome:

- Speed of AI Development: A rapid takeoff might favor a unipolar outcome by giving a significant advantage to the fastest developer. In contrast, a slower takeoff could lead to a multipolar world where many entities reach advanced capabilities more or less simultaneously.
- Collaboration vs. Competition: The degree of collaboration and openness in the AI research community can significantly affect takeoff polarity. High levels of collaboration and information sharing could support a multipolar outcome, while secretive or highly competitive environments might push towards unipolarity.

- **Regulatory and Economic Dynamics:** Regulatory frameworks and economic incentives also play a crucial role. Policies that encourage diversity in AI development and mitigate against the accumulation of too much power in any single entity's hands could foster a multipolar takeoff.

1.7.5 Takeoff Arguments

The Overhang Argument . There might be situations where there are substantial advancements or availability in one aspect of the AI system, such as hardware or data, but the corresponding software or algorithms to fully utilize these resources haven't been developed yet. The term 'overhang' is used because these situations imply a kind of 'stored' or 'latent' potential. Once the software or algorithms catch up to the hardware or data, there could be a sudden unleashing of this potential, leading to a rapid leap in AI capabilities. Overhangs provide one possible argument for why we might favor discontinuous or fast takeoffs. There are two types of overhangs commonly discussed:

- **Hardware Overhang :** This refers to a situation where there is enough computing hardware to run many powerful AI systems, but the software to run such systems hasn't been developed yet. If such hardware could be repurposed for AI, this would mean that as soon as one powerful AI system exists, probably a large number of them would exist, which might amplify the impact of the arrival of human-level AI.
- **Data Overhang :** This would be a situation where there is an abundance of data available that could be used for training AI systems, but the AI algorithms capable of utilizing all that data effectively haven't been developed or deployed yet.

Overhangs are also used as a counter argument to why AI pauses do not affect takeoff. One counter argument to the overhang argument is that it relies on the assumption that during the time that we are pausing AI development, the rate of production of chips will remain constant. It could be argued that the companies manufacturing these chips will not make as many chips if data centers aren't buying them. However, this argument only works if the pause is for any appreciable length of time, otherwise the data centers might just stockpile the chips. It is also possible to make progress on improved chip design, without having to manufacture as many during the pause period. However, during the same pause period we could also make progress on AI safety techniques. ([Elmore, 2024](#))

The Economic Growth Argument . Historical patterns of economic growth, driven by human population increases, suggest a potential for slow and continuous AI takeoff. This argument says that as AIs augment the effective economic population, we might witness a gradual increase in economic growth, mirroring past expansions but at a potentially accelerated rate due to AI-enabled automation. Limitations in AI's ability to automate certain tasks, alongside societal and regulatory constraints (e.g. that medical or legal services can only be rendered by humans), could lead to a slower expansion of AI capabilities. Alternatively, growth might far exceed historical rates. Using a similar argument for a fast takeoff hinges on AI's potential to quickly automate human labor on a massive scale, leading to unprecedented economic acceleration.

Compute Centric Takeoff Argument . This argument, similar to the Bio Anchors report, assumes that compute will be sufficient for transformative AI. Based on this assumption, Tom Davidson's 2023 report on compute-centric AI takeoff discusses feedback loops that may contribute to takeoff dynamics.

- **Investment feedback loop :** There might be increasing investment in AI, as AIs play a larger and larger role in the economy. This increases the amount of compute available to train models, as well as potentially leading to the discovery of novel algorithms. All of this increases capabilities, which drives economic progress, and further incentivizes investment.
- **Automation feedback loop :** As AIs get more capable, they will be able to automate larger parts of the work of coming up with better AI algorithms, or helping in the design of better GPUs. Both of these will increase the capability of the AIs, which in turn allow them to automate more labor.

Depending on the strength and interplay of these feedback loops, they can create a self-fulfilling prophecy leading to either an accelerating fast takeoff if regulations don't curtail various aspects of such loops, or a

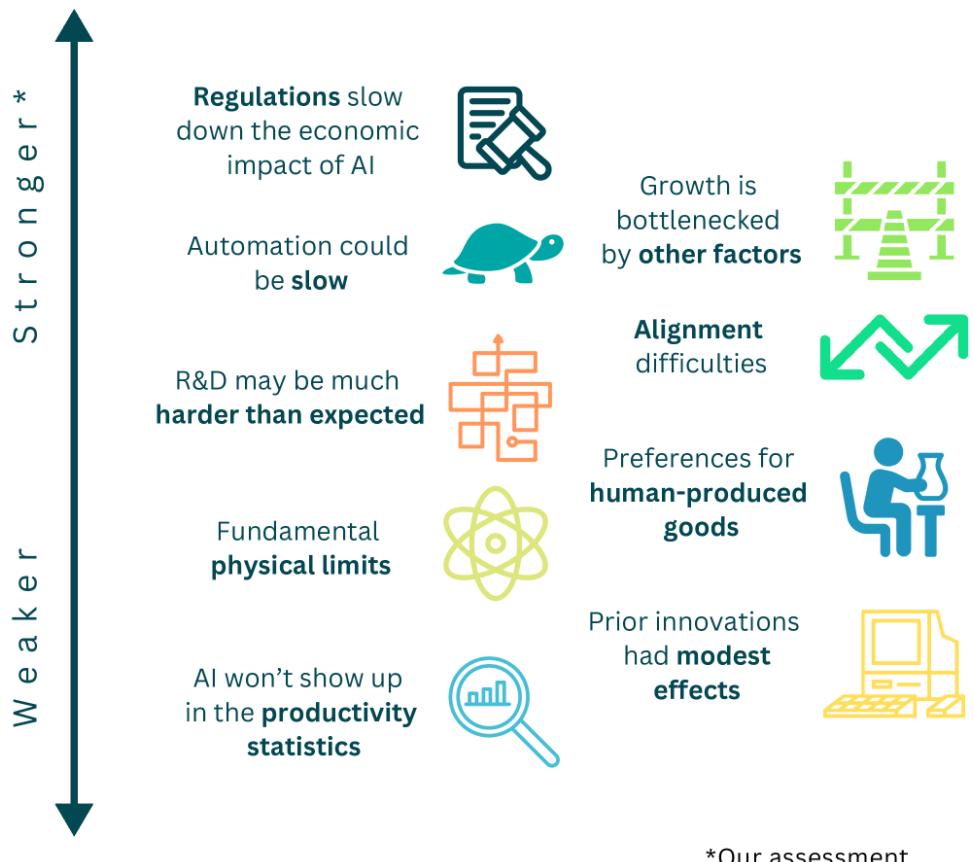


Figure 1.40 : A visualization of the ranking of arguments for explosive economic growth, both in favor and against. By Epoch AI. ([Erdil & Besiroglu, 2024](#))

slow takeoff if the loops are weaker or counterbalanced by other factors. The entire model is shown in the diagram below:

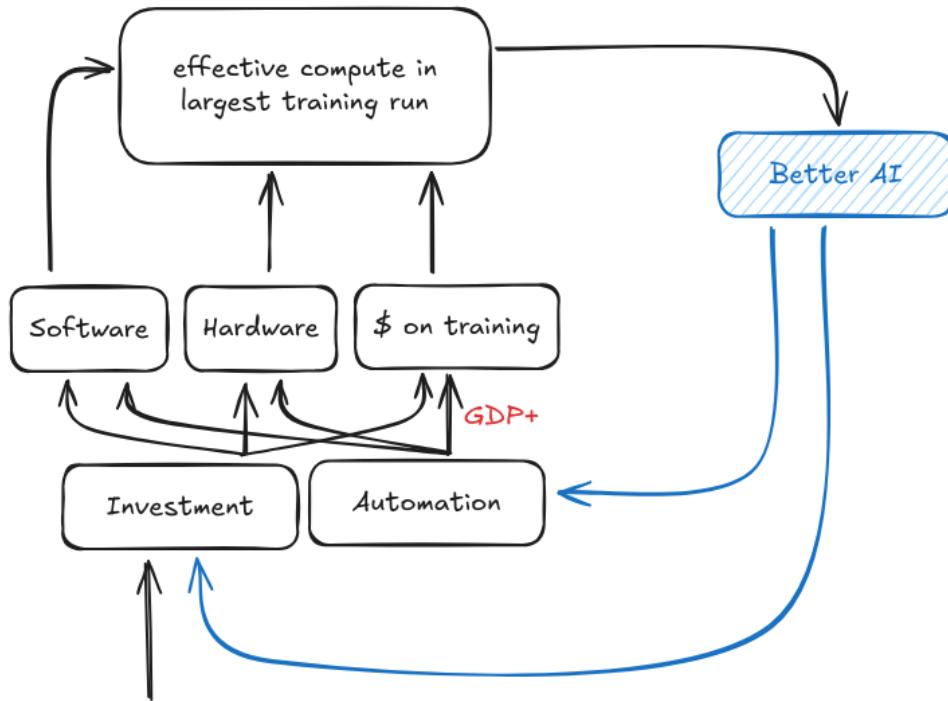


Figure 1.41 : A summary of What a Compute-Centric Framework Says About Takeoff Speeds ([Davidson, 2024](#))

Automating Research Argument. Researchers could potentially design the next generation of ML models more quickly by delegating some work to existing models, creating a feedback loop of ever-accelerating progress. The following argument is put forth by Ajeya Cotra:

Currently, human researchers collectively are responsible for almost all of the progress in AI research, but are starting to delegate a small fraction of the work to large language models. This makes it somewhat easier to design and train the next generation of models.

The next generation is able to handle harder tasks and more different types of tasks, so human researchers delegate more of their work to them. This makes it significantly easier to train the generation after that. Using models gives a much bigger boost than it did the last time around.

Each round of this process makes the whole field move faster and faster. In each round, human researchers delegate everything they can productively delegate to the current generation of models — and the more powerful those models are, the more they contribute to research and thus the faster AI capabilities can improve. ([Cotra, 2023](#))

So before we see a recursive explosion of intelligence, we see a steadily increasing amount of the full RnD process being delegated to AIs. At some point, instead of a significant majority of the research and design being done by AI assistants at superhuman speeds, it will become that - all of the research and design for AIs is done by AI assistants at superhuman speeds.

At this point there is a possibility that this might eventually lead to a full automated recursive intelligence explosion.

The Intelligence Explosion Argument . This concept of the 'intelligence explosion' is also central to the conversation around discontinuous takeoff. It originates from I.J. Good's thesis, which posits that sufficiently advanced machine intelligence could build a smarter version of itself. This smarter version

could in turn build an even smarter version of itself, and so on, creating a cycle that could lead to intelligence vastly exceeding human capability ([Yudkowsky, 2013](#)).

In their 2012 report on the evidence for Intelligence Explosions, Muehlhauser and Salamon delve into the numerous advantages that machine intelligence holds over human intelligence, which facilitate rapid intelligence augmentation. ([Muehlhauser, 2012](#)) These include:

- **Computational Resources** : Human computational ability remains somewhat stationary, whereas machine computation possesses scalability.
- **Speed** : Humans communicate at a rate of two words per second, while GPT-4 can process 32k words in an instant. Once LLMs can write "better" than humans, their speed will most probably surpass us entirely.
- **Duplicability** : Machines exhibit effortless duplicability. Unlike humans, they do not need birth, education, or training. While humans predominantly improve individually, machines have the potential to grow collectively. Humans take 20 years to become competent from birth, whereas once we have one capable AI, we can duplicate it immediately. Once AIs reach the level of the best programmer, we can just duplicate this AI. The same goes for other jobs.
- **Editability** : Machines potentially allow more regulated variations. They exemplify the equivalent of direct brain enhancements via neurosurgery in opposition to laborious education or training requirements. Humans can also improve and learn new skills, but they don't have root access to their hardware: we are just starting to be able to understand the genome's "spaghetti code," while AI could use code versioning tools to improve itself, being able to attempt risky experiments with backup options in case of failure. This allows for much more controlled variation.
- **Goal coordination** : Copied AIs possess the capability to share goals effortlessly, a feat challenging for humans.

1.8 Appendix: Expert Opinions

Everything in the appendices is optional, and is meant as additional knowledge and context. You do not need to read this to understand the core points made in either this chapter or future chapters.

1.8.1 Surveys

According to a recent survey conducted by AI Impact ([AI Impacts, 2022](#)): "*Expected time to human-level performance dropped 1–5 decades since the 2022 survey . As always, our questions about 'high-level machine intelligence' (HLMI) and 'full automation of labor' (FAOL) got very different answers, and individuals disagreed a lot (shown as thin lines below), but the aggregate forecasts for both sets of questions dropped sharply. For context, between 2016 and 2022 surveys, the forecast for HLMI had only shifted about a year. "*

It is also possible to compare the predictions of the same study in 2022 to the current results. It is interesting to note that the community has generally underestimated the speed of progress over the year 2023 and has adjusted its predictions downward. Some predictions are quite surprising. For example, tasks like "Write High School Essay" and "Transcribe Speech" are arguably already automated with ChatGPT and Whisper, respectively. However, it appears that researchers are not aware of these results. Additionally, it is surprising that the forecast for when we are able to build an "AI researcher" has longer timelines than when we are able to build "High-level machine intelligence (all human tasks)".

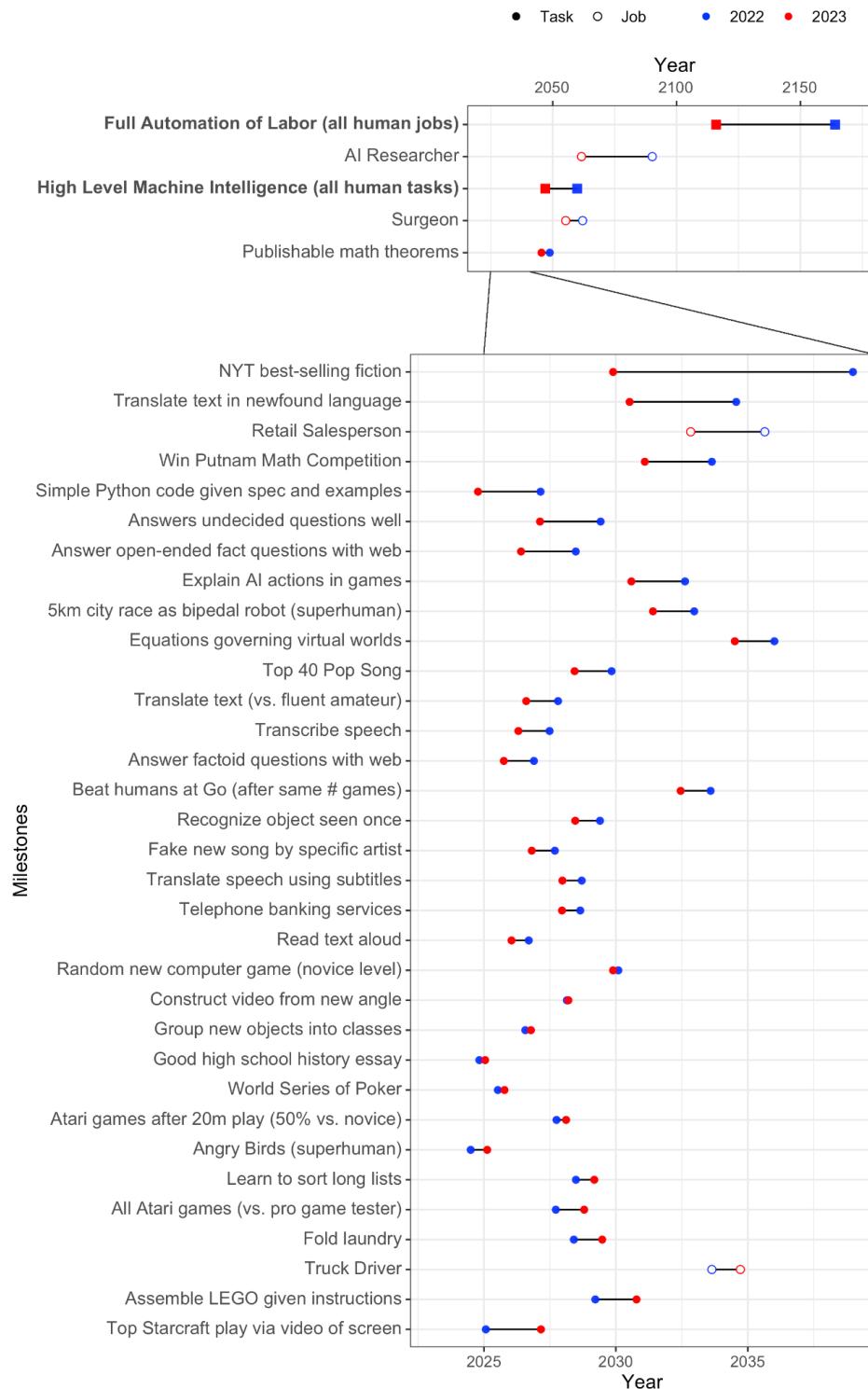


Figure 1.42 : 2024 Survey of AI Experts (AI Impacts, 2022)

The median of the 2024 expert survey predicts human-level machine intelligence (HLMI) in 2049.

1.8.2 Quotes

Here are many quotes from people regarding transformative AI.

AI Experts

Note that Hinton, Bengio, and Sutskever are some of the most cited researchers in the field of AI. And that Hinton, Bengio, and LeCun are the recipients of the Turing Award in Deep Learning. Some users on reddit have put together a comprehensive list of publicly stated AI timelines forecasts from famous researchers and industry leaders.

Geoffrey Hinton (Godfather of modern AI, Turing Award Recipient)

"The research question is: how do you prevent them from ever wanting to take control? And nobody knows the answer [...] The alarm bell I'm ringing has to do with the existential threat of them taking control [...] If you take the existential risk seriously, as I now do, it might be quite sensible to just stop developing these things any further [...] it's as if aliens had landed and people haven't realized because they speak very good English"

Yoshua Bengio (One of most cited scientists ever, Godfather of modern AI, Turing Award Recipient)

"It's very hard, in terms of your ego and feeling good about what you do, to accept the idea that the thing you've been working on for decades might actually be very dangerous to humanity.... I think that I didn't want to think too much about it, and that's probably the case for others [...] Rogue AI may be dangerous for the whole of humanity. Banning powerful AI systems (say beyond the abilities of GPT-4) that are given autonomy and agency would be a good start."

Yann LeCun (Godfather of modern AI, Turing Award Recipient, Chief AI Scientist at Meta)

"There is no question that machines will become smarter than humans—in all domains in which humans are smart—in the future. It's a question of when and how, not a question of if."

Stuart Russell (Co-Author of leading AI textbook, Co-Founder of Center for Human-Compatible AI)

"If we pursue [our current approach], then we will eventually lose control over the machines."

Demis Hassabis (Co-Founder and CEO of DeepMind)

"We must take the risks of AI as seriously as other major global challenges, like climate change. It took the international community too long to coordinate an effective global response to this, and we're living with the consequences of that now. We can't afford the same delay with AI [...] then maybe there's some kind of equivalent one day of the IAEA, which actually audits these things."

Dario Amodei (Co-Founder and CEO of Anthropic, Former Head of AI Safety at OpenAI)

"When I think of why am I scared [...] I think the thing that's really hard to argue with is like, there will be powerful models; they will be agentic; we're getting towards them. If such a model wanted to wreak havoc and destroy humanity or whatever, I think we have basically no ability to stop it."

Mustafa Suleyman (CEO of Microsoft AI, Co-Founder of DeepMind)

"[About a Pause] I don't rule it out. And I think that at some point over the next five years or so, we're going to have to consider that question very seriously."

Ilya Sutskever (One of the most cited scientists ever, Co-Founder and Former Chief Scientist at OpenAI)

"The future is going to be good for the AIs regardless; it would be nice if it would be good for humans as well [...] It's not that it's going to actively hate humans and want to harm them, but it's just going to be too powerful, and I think a good analogy would be the way humans treat animals [...] And I think by default that's the kind of relationship that's going to be between us and AGIs which are truly autonomous and operating on their own behalf."

Shane Legg (Co-Founder and Chief AGI Scientist at DeepMind)

"Do possible risks from AI outweigh other possible existential risks...? It's my number 1 risk for this century [...] A lack of concrete AGI projects is not what worries me, it's the lack of concrete plans on how to keep these safe that worries me."

Jan Leike (Former co-lead of the Superalignment project at OpenAI)

"[After resigning at OpenAI, talking about sources of risks] These problems are quite hard to get right, and I am concerned we aren't on a trajectory to get there [...] OpenAI is shouldering an enormous responsibility on behalf of all of humanity. But over the past years, safety culture and processes have taken a backseat to shiny products. We are long overdue in getting incredibly serious about the implications of AGI."

Sam Altman (Co-Founder and CEO of OpenAI)

"[Suggesting about how to ask for a global regulatory body:] "any compute cluster above a certain extremely high-power threshold – and given the cost here, we're talking maybe five in the world, something like that – any cluster like that has to submit to the equivalent of international weapons inspectors" [...] I did a big trip around the world this year, and talked to heads of state in many of the countries that would need to participate in this, and there was almost universal support for it."

Greg Brockman (Co-Founder and Former CTO of OpenAI)

"The exact way the post-AGI world will look is hard to predict — that world will likely be more different from today's world than today's is from the 1500s [...] We do not yet know how hard it will be to make sure AGIs act according to the values of their operators. Some people believe it will be easy; some people believe it'll be unimaginably difficult; but no one knows for sure"

John Schulman (Co-Founder of OpenAI)

"[Talking about times near the creation of the first AGI] you have the race dynamics where everyone's trying to stay ahead, and that might require compromising on safety. So I think you would probably need some coordination among the larger entities that are doing this kind of training [...] Pause either further training, or pause deployment, or avoiding certain types of training that we think might be riskier."

Jaan Tallinn (Co-Founder of Skype, Future of Life Institute)

"I've not met anyone in AI labs who says the risk [from training a next-gen model] is less than 1% of blowing up the planet. It's important that people know lives are being risked [...] One thing that a pause achieves is that we will not push the Frontier, in terms of risky pre-training experiments."

Politicians

Rishi Sunak (Former UK Prime Minister)

"Get this wrong, and AI could make it easier to build chemical or biological weapons. Terrorist groups could use AI to spread fear and destruction on an even greater scale. Criminals could exploit AI for cyber-attacks, disinformation, fraud, or even child sexual abuse. And in the most unlikely but extreme cases, there is even the risk that humanity could lose control of AI completely through the kind of AI sometimes referred to as 'super intelligence'."

Kamala Harris (Former US Vice President)

"[...] just as AI has the potential to do profound good, it also has the potential to cause profound harm. From AI-enabled cyberattacks at a scale beyond anything we have seen before to AI-formulated bio-weapons that could endanger the lives of millions, these threats are often referred to as the "existential threats of AI" because, of course, they could endanger the very existence of humanity. These threats, without question, are profound, and they demand global action."

Zhang Jun (China's UN Ambassador)

"The potential impact of AI might exceed human cognitive boundaries. To ensure that this technology always benefits humanity, we must regulate the development of AI and prevent this technology from turning into a runaway wild horse [...] We need to strengthen the detection and evaluation of the entire lifecycle of AI, ensuring that mankind has the ability to press the pause button at critical moments."

Donald Trump (Former US President)

"[...] the other thing that I think is maybe the most dangerous thing out there of anything, because there's no real solution — the AI, as they call it."

Joe Biden (US President)

"Artificial Intelligence is one of the most powerful tools of our time, but to seize its opportunities, we must first mitigate its risks. [...] Social media has shown us the harm that powerful technology can do without the right safeguards in place [...] we must be clear-eyed and vigilant about the threats emerging — of emerging technologies that can pose — don't have to, but can pose — to our democracy and our values."

Vladimir Putin (President of Russia)

"Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world [...] If we become leaders in this area, we will share this know-how with [the] entire world, the same way we share our nuclear technologies today."

Li Qiang (China's Head of Government)

"AI must be guided in a direction that is conducive to the progress of humanity. So there should be a red line in AI development, a red line that must not be crossed [...] It should not just benefit only a small group of people, but benefit the overwhelming majority of mankind [...] It is essential that we work together and coordinate with each other."

Ursula von der Leyen (Head of EU Executive Branch)

"[We] should not underestimate the real threats coming from AI [...] It is moving faster than even its developers anticipated [...] We have a narrowing window of opportunity to guide this technology responsibly."

António Guterres (UN Secretary-General)

"AI poses a long-term global risk. Even its own designers have no idea where their breakthrough may lead. I urge [the UN Security Council] to approach this technology with a sense of urgency [...] Its creators themselves have warned that much bigger, potentially catastrophic and existential risks lie ahead."

Academics

Eliezer Yudkowsky (Co-Founder of Machine Intelligence Research Institute)

"I do not expect something actually smart to attack us with marching robot armies with glowing red eyes where there could be a fun movie about us fighting them. I expect an actually smarter and uncaring entity will figure out strategies and technologies that can kill us quickly and reliably and then kill us."

Stephen Hawking (Theoretical Physicist)

"The development of full artificial intelligence could spell the end of the human race [...] It would take off on its own, and re-design itself at an ever increasing rate."

Alan Turing (Father of Computer Science and AI)

"It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers... They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control."

I. J. Good (Cryptologist at Bletchley Park)

"An ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Tech Entrepreneurs

Elon Musk (Founder/Co-Founder of OpenAI, Neuralink, SpaceX, xAI, PayPal, CEO of Tesla, CTO of X/Twitter)

"AI is a rare case where I think we need to be proactive in regulation than be reactive [...] I think that [digital super intelligence] is the single biggest existential crisis that we face and the most pressing one. It needs to be a public body that has insight and then oversight to confirm that everyone is developing AI safely [...] And mark my words, AI is far more dangerous than nukes. Far. So why do we have no regulatory oversight? This is insane."

Bill Gates (Co-Founder of Microsoft)

"Superintelligent AIs are in our future. [...] There's the possibility that AIs will run out of control. [Possibly,] a machine could decide that humans are a threat, conclude that its interests are different from ours, or simply stop caring about us."

Join Declarations

The Bletchley Declaration (Multiple Nations & EU, 2023)

"Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent. These issues are in part because those capabilities are not fully understood [...] There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models."

AI Risk Statement (Multiple AI Experts, 2023)

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

1.8.3 Prediction Markets

Prediction markets are like betting systems where people can buy and sell shares based on their predictions of future events. For example, if there's a prediction market for a presidential election, you can buy shares for the candidate you think will win. If many people believe Candidate A will win, the price of shares for Candidate A goes up, indicating a higher probability of winning.

These markets are helpful because they gather the knowledge and opinions of many people, often leading to accurate predictions. For example, a company might use a prediction market to forecast whether a new product will succeed. Employees can buy shares if they believe the product will do well. If the majority think it will succeed, the share price goes up, giving the company a good indication of the product's potential success.

By allowing participants to profit from accurate predictions, these markets encourage the sharing of valuable information and provide real-time updates on the likelihood of various outcomes. The argument is that either prediction markets are more accurate than experts, or experts should be able to make a lot of money from these markets and, in doing so, correct the markets. So the incentive for profit leads to the most accurate predictions. Examples of prediction markets include [manifold](#), or [metaculus](#).

When using prediction markets to estimate the reproducibility of scientific research it was found that they outperformed expert surveys ([Dreber et al., 2015](#)). So if a lot of experts participate, prediction markets might be one of our best probabilistic forecasting tools, better even than surveys or experts.

The live charts below show the results of the prediction markets from Metaculus for - "When will the first weakly general AI system be devised, tested, and publicly announced?" At the time of writing, weakly general systems are expected in 2027, and general systems in 2032.

1.9 Appendix: Discussion on LLMs

Everything in the appendices is optional, and is meant as additional knowledge and context. You do not need to read this to understand the core points made in either this chapter or future chapters.

Current LLMs, although trained on abundant data, are still far from perfect.

Will these problems persist in future iterations, or will they disappear? This section examines the main criticisms of those models and tries to determine if they are valid even for future LLMs.

This kind of qualitative assessment is important to know whether LLMs represent the most likely route to AGI or not.

1.9.1 Empirically insufficiency?

Can LLMs be creative ? The creativity of LLMs is often debated, but there are clear indications that AI, in principle, is capable of creative processes in various ways:

- **Autonomous Scientific Research :** Recent advancements have shown that LLMs can indeed make novel discoveries. For example, a study by DeepMind demonstrated that an LLM "discovered new solutions for the cap set problem, a long-standing open problem in mathematics" ([DeepMind, 2023](#)) which was a favorite open problem of Terence Tao. This indicates that AI can not only understand existing knowledge but also contribute new insights in complex fields like mathematics.
- **Autonomous Discovery :** AI has the capability to rediscover human strategies and openings independently. AlphaGo, for example, rediscovered human Go strategies and openings through self-play ([McGrath et al., 2021](#)), without any human data input. This demonstrates an AI's ability to independently learn and innovate within established domains.
- **Creative Optimization :** AI can optimize in surprisingly creative ways. The phenomena of specification gaming, where AI finds unintended solutions to problems, illustrate this. Although this unpredictability poses its challenges, it also shows that AI systems can come up with novel, creative solutions that might not be immediately obvious or intuitive to human problem solvers. DeepMind's blog post on Specification Gaming illustrates this point vividly. ([Krakovna et al., 2020](#))

Aren't LLMs just too slow at learning things? Arguments against transformer based language models often state that they are too sample inefficient, and that LLMs are extremely slow to learn new concepts when compared to humans. To increase performance in new tasks or situations, it's often argued that LLMs require training on vast amounts of data — millions of times more than a human would need. However, there's a growing trend towards data efficiency, and an increasing belief that this can be significantly improved in future models.

EfficientZero is a reinforcement learning agent that surpasses median human performance on a set of 26 Atari games after just two hours of real-time experience per game. ([Ye et al., 2021; Wang et al., 2024](#)) This is a considerable improvement over previous algorithms, showcasing the potential leaps in data efficiency. The promise here is not just more efficient learning but also the potential for rapid adaptation and proficiency in new tasks, akin to a child's learning speed. EfficientZero is not an LLM, but it shows that deep learning can sometimes be made efficient.

Scaling laws indicate that larger AIs tend to be more data efficient, requiring less data to reach the same level of performance as their smaller counterparts. Papers such as "Language Models are Few-Shot Learners" ([Brown et al., 2020](#)) and the evidence that larger models seem to take less data to reach the same level of performance ([Kaplan et al., 2020](#)), suggest that as models scale, they become more proficient with fewer examples. This trend points towards a future where AI might be able to rapidly adapt and learn from limited data, challenging the notion that AIs are inherently slow learners compared to humans.

Are LLMs robust to distributional shifts? While it is true that AI has not yet achieved maximal robustness, for example being able to perform perfectly after a change in distribution, there has been considerable progress:

- **Robustness correlates with capabilities :** Robustness is closely linked to the capabilities of AI models when AIs are trained on difficult tasks. For instance, there is a significant improvement in robustness and transfer learning from GPT-2 to GPT-4. In computer vision, recent models like Segment Anything ([Kirillov et al., 2023](#)) are far more robust and capable of transfer learning than their less capable predecessors. This progression isn't due to any mysterious factors but rather a result of scaling and improving upon existing architectures.
- **Robustness is a continuum, and perfect robustness may be not necessary:** Robustness in AI should not be viewed as a binary concept, but rather as existing on a continuum. This continuum is evident in the way AI models, like those in image classification, often surpass human performance

in both capability and robustness (Korzekwa, 2022). However, it's important to recognize that no system is completely immune to challenges such as adversarial attacks. This is exemplified by advanced AIs like Katago in Go, which, despite being vulnerable to such attacks (Wang et al., 2022), still achieves a superhuman level of play. However, the quest for perfect robustness may not be essential to create capable transformative AI, as even systems with certain vulnerabilities can achieve superhuman levels of competence. However, while robustness may not be necessary to create capable AI, the creation of safe, aligned AI will have to solve the problem of misgeneralizing goals.

1.9.2 Shallow Understanding?

Stochastic Parrots: Do AIs only memorize information without truly compressing it?

François Chollet (Prominent AI Researcher) (Chollet, 2023)

Unfortunately, too few people understand the distinction between memorization and understanding. It's not some lofty question like "does the system have an internal world model?", it's a very pragmatic behavior distinction: "is the system capable of broad generalization, or is it limited to local generalization?"

There are two archetypal ways to represent information in an LLM: either memorize point by point, like a look-up table, or compress the information by only memorizing higher-level features, which we can then call "the world model". This is explained in the very important paper "Superposition, Memorization, and Double Descent" (Anthropic, 2023): it turns out that to store points, initially the model learns the position of all the points (pure memorization), then, if we increase the number of points, the model starts to compress this knowledge, and the model is now capable of generalization (and implements a simple model of the data).

AI is capable of compressing information, often in a relevant manner. For example, when examining the representations of words representing colors in LLMs like "red" and "blue", the structure formed by all the embeddings of those colors creates the correct color circle (This uses a nonlinear projection such as a T-distributed stochastic neighbor embedding (T-SNE) to project from high-dimensional space to the 2D plane). Other examples of world models are presented in a paper called "Eight Things to Know about Large Language Models" (Bowman, 2023).

Of course, there are other domains where AI resembles more of a look-up table, but it is a spectrum, and each case should be examined individually. For example, for "factual association," the paper "Locating and Editing Factual Associations in GPT" shows that the underlying data structure for GPT-2 is more of a look-up table (Meng et al., 2023), but the paper "Emergent Linear Representations in World Models of Self-Supervised Sequence Models" demonstrates that a small GPT is capable of learning a compressed world model of OthelloGpt. (Nanda et al., 2023) There are more examples in the section dedicated to world models in the paper "Eight Things to Know about Large Language Models" (Bowman, 2023).

It's clear that LLMs are compressing their representations at least a bit. Many examples of impressive capabilities are presented in the work "The Stochastic Parrot Hypothesis is debatable for the last generation of LLMs", which shows that it cannot be purely a memorization. (Feuillade-Montixi & Peigné, 2023)

Will LLMs Inevitably Hallucinate? LLMs are prone to "hallucinate," a term used to describe the generation of content that is nonsensical or factually incorrect in response to certain prompts. This issue, highlighted in studies such as "On Faithfulness and Factuality in Abstractive Summarization" by Maynez et al. (Maynez et al., 2020) and "TruthfulQA: Measuring How Models Mimic Human Falsehoods" by Lin et al. (Lin et al., 2022), poses a significant challenge. However, it's important to see that these challenges are anticipated due to the training setup and can be mitigated:

- **Inherent Bias in Source Texts :** One of the fundamental reasons LLMs may produce untrue content is training data, which may not always be entirely factual or unbiased. In essence, LLMs

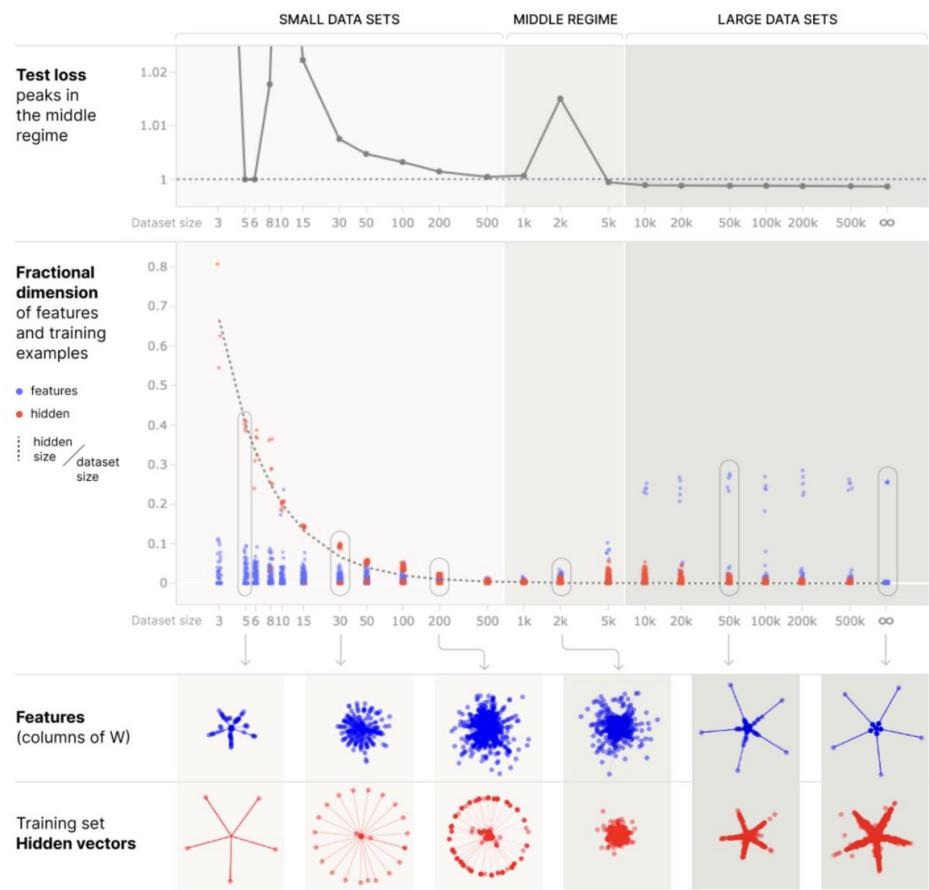


Figure 1.43 : From Superposition, Memorization, and Double Descent ([Anthropic, 2023](#))

are reflecting the diverse and sometimes contradictory nature of their training data. In this context, LLMs are constantly 'hallucinating', but occasionally, these hallucinations align with our perception of reality.

- **Strategies to Enhance Factual Accuracy** : The tendency of LLMs to generate hallucinations can be significantly diminished using various techniques. See the box below for a breakdown of those.
- **Larger models can be more truthful than smaller ones.** This is the case with TruthfulQA. OpenAI reports that GPT-4 is 40% more accurate and factually consistent than its predecessor.

Many techniques can be used to increase the truthfulness of LLMs

- **Fine-tuning LLMs for Factuality** : In this paper ([Tian et al., 2023](#)), the authors recommend fine-tuning methods using Direct Preference Optimization (DPO) to decrease the rate of hallucinations. By applying such techniques, a 7B Llama 2 model saw a 58% reduction in factual error rate compared to its original model.
- **Retrieval Augmented Generation (RAG)** . This method works by incorporating a process of looking up real-world information (retrieval, like a Google search) and then using that information to guide the AI's responses (generation, based on the document retrieved). By doing so, the AI is better anchored in factual reality, reducing the chances of producing unrealistic or incorrect content. Essentially, it's like giving the AI a reference library to check facts against while it learns and responds, ensuring its output is more grounded in reality. This approach is particularly useful in the context of in-context learning, where the AI learns from the information and context provided in each interaction.
- **Prompting techniques** in AI have evolved to include sophisticated methods like

<tab>

- **Consistency checks** ([Fluri et al., 2023](#)), that involve comparing the output from multiple instances of the model on the same prompt, identifying and resolving any disagreements in the responses. This method enhances the accuracy and credibility of the information provided. For example, if different iterations of the model produce conflicting answers, this discrepancy can be used to refine and improve the model's understanding.
- **Reflexion.** The Reflexion technique ("Reflexion: Language Agents with Verbal Reinforcement Learning"): It's possible to simply ask the LLM to take a step back, to question whether what it has done is correct or not, and to consider ways to improve the previous answer, and this enhances a lot the capabilities of GPT-4. This technique is emergent and does not work well with previous models. ([Shinn et al., 2023](#)).
- **Verification chains**, like **selection inference** ([Creswell et al., 2022](#)). Chain-of-Thought has access to the whole context, so each reasoning step is not necessarily causally connected to the last. But selection inference enforces a structure where each reasoning step necessarily follows from the last, and therefore the whole reasoning chain is causal. This process involves the AI model examining its own reasoning or the steps it took to arrive at a conclusion. By doing so, it can verify the logic and consistency of its responses, ensuring they are well-founded and trustworthy.
- **Allowing the AI to express degrees of confidence** in its answers, acknowledging uncertainty when appropriate. For example, instead of a definitive "Yes" or "No," the model might respond with "I am not sure," reflecting a more nuanced understanding akin to human reasoning. This approach is evident in advanced models like Gopher ([Rae et al., 2022](#)), which contrasts with earlier models such as WebGPT which may not exhibit the same level of nuanced responses.

- **Process-based training** ensures that the systems are accustomed to detailing their thoughts in much greater detail and not being able to skip too many reasoning steps. For example, see OpenAI's Improving Mathematical Reasoning with process supervision ([Lightman et al., 2023](#)).
- **Training for metacognition** : Models can be trained to give the probability of what they assert, a form of metacognition. For example, the paper "Language Models (Mostly) Know What They Know" ([Kadavath et al., 2022](#)) demonstrates that AIs can be Bayesian calibrated about their knowledge. This implies that they can have a rudimentary form of self-awareness, recognizing the likelihood of their own accuracy. Informally, this means it is possible to query a chatbot with "Are you sure about what you are telling me?" and receive a relatively reliable response. This can serve as training against hallucinations.

It's worth noting that these techniques enable substantial problem mitigation for the current LLMs, but they don't solve all the problems that we encounter with AI that are potentially deceptive, as we will see in the chapter on goal misgeneralization.

</tab>

1.9.3 Structural inadequacy?

Are LLMs missing System 2? System 1 and System 2 are terms popularized by economist Daniel Kahneman in his book "Thinking, Fast and Slow," describing the two different ways our brains form thoughts and make decisions. System 1 is fast, automatic, and intuitive; it's the part of our thinking that handles everyday decisions and judgments without much effort or conscious deliberation. For example, when you recognize a face or understand simple sentences, you're typically using System 1. On the other hand, System 2 is slower, more deliberative, and more logical. It takes over when you're solving a complex problem, making a conscious choice, or focusing on a difficult task. It requires more energy and is more controlled, handling tasks such as planning for the future, checking the validity of a complex argument, or any activity that requires deep focus. Together, these systems interact and influence how we think, make judgments, and decide, highlighting the complexity of human thought and behavior.

A key concern is whether LLMs are able to emulate System 2 processes, which involve slower, more deliberate, and logical thinking. Some theoretical arguments about the depth limit in transformers show that they are provably incapable of internally dividing large integers ([Delétang et al., 2023](#)). However, this is not what we observe in practice: GPT-4 is capable of detailing some calculations step-by-step and obtaining the expected result through a chain of thought or via the usage of tools like a code interpreter.

Emerging Metacognition . Emerging functions in LLMs, like the Reflexion technique ([Shinn et al., 2023](#)), allow these models to retrospectively analyze and improve their answers. It is possible to ask the LLM to take a step back, question the correctness of its previous actions, and consider ways to improve the previous answer. This greatly enhances the capabilities of GPT-4, enhancing its capabilities and aligning them more closely with human System 2 operations. Note that this technique is emergent and does not work well with previous models.

These results suggest a blurring of the lines between these two systems. System 2 processes may be essentially an assembly of multiple System 1 processes, appearing slower due to involving more steps and interactions with slower forms of memory. This perspective is paralleled in how language models operate, with each step in a System 1 process akin to a constant time execution step in models like GPT. Although these models struggle with intentionally orchestrating these steps to solve complex problems, breaking down tasks into smaller steps (Least-to-most prompting) or prompting them for incremental reasoning (Chain-of-Thought (CoT) prompting) significantly improves their performance.

Are LLMs missing an internal world model? The notion of a "world model" in AI need not be confined to explicit encoding within an architecture. Contrary to approaches like H-JEPA ([LeCun, 2022](#)), which advocate for an explicit world model to enhance AI training, there's growing evidence that a world model can be effectively implicit. This concept is particularly evident in reinforcement learning (RL), where the distinction between model-based and model-free RL can be somewhat misleading. Even in model-free RL, algorithms often implicitly encode a form of a world model that is crucial for optimal performance.

- **Time and geographical coordinates:** Research on Llama-2 models reveals how these models can represent spatial and temporal information ([Gurney & Tegmark, 2024](#)). LLMs like Llama-2 models encode approximate real-world coordinates and historical timelines of cities. Key findings include the gradual emergence of geographical representations across model layers, the linearity of these representations, and the models' robustness to different prompts. Significantly, the study shows that the models are not just passively processing this information but actively learning the global geometry of space and time.
- **Board representation:** In the paper "Emergent Linear Representations in World Models of Self-Supervised Sequence Models" ([Nanda et al., 2023](#)), the author presents significant findings on the nature of representations in AI models. The paper delves into how the Othello-GPT model, trained to predict legal moves in the game of Othello, develops an emergent world representation of the game board! Contrary to previous beliefs that this representation was non-linear, he demonstrates that it is, in fact, linear. He discovers that the model represents board states not in terms of black or white pieces, but as "my color" or "their color," aligning with the model's perspective of playing both sides. This work sheds light on the potential of AI models to develop complex, yet linear, world representations through simple objectives like next-token prediction.
- **Other examples** are presented in the paper: "Eight Things to know about LLMs". ([Bowman, 2023](#))

Can LLMs learn continuously, and have long term memory? Continual learning and the effective management of long-term memory represent significant challenges in the field of AI in general.

Catastrophic Forgetting. A crucial obstacle in this area is catastrophic forgetting, a phenomenon where a neural network, upon learning new information, tends to entirely forget previously learned information. This issue is an important focus of ongoing research, aiming to develop AI systems that can retain and build upon their knowledge over time. For example, suppose we train an AI on an Atari game. At the end of the second training, the AI has most likely forgotten how to play the first game. This is an example of catastrophic forgetting.

But now suppose we train a large AI on many ATARI games, simultaneously, and even add some Internet text and some robotic tasks. This can just work. For example, the AI GATO illustrates this training process and exemplifies what we call the **blessing of scale**, which is that what is impossible in small regimes can become possible in large regimes.

Other techniques are being developed to solve long-term memory, for example, **Scaffolding-based approaches** have also been employed for achieving long-term memory and continual learning in AI. Scaffolding in AI refers to the use of hard-coded wrappers explicitly programmed structures by humans that involve a for loop to query continuously the model:

- **LangChain** addresses these challenges by creating extensive memory banks. LangChain is a Python library that allows LLM to retrieve and utilize information from large datasets, essentially providing a way for AI to access a vast repository of knowledge and use this information to construct more informed responses. However, this approach may not be the most elegant due to its reliance on external data sources and complex retrieval mechanisms. A potentially more seamless and integrated solution could involve utilizing the neural network's weights as dynamic memory, constantly evolving and updating based on the tasks performed by the network.
- **Voyager:** A remarkable example of a scaffolding-based long-term memory is the AI Voyager, an AI system developed under the "AutoGPT" paradigm. This system is notable for its ability to engage in continuous learning within a 3D game environment like Minecraft. In a single game session, AI Voyager demonstrates the capacity to learn basic controls, achieve initial goals such as resource acquisition, and eventually advance to more complex behaviors, including combat with enemies and crafting tools for gathering sophisticated resources. This demonstrates a significant stride in LLM's ability to learn continually and manage long-term memory within dynamic environments.

It should be noted that scaffold-based long-term memory is not considered an elegant solution, and purists would prefer to use the system's own weights as long-term memory.

Planning Planning is an area that AIs currently struggle with, but there is significant progress. Some paradigms, such as those based on scaffolding, enable task decomposition and breaking down objectives into smaller, more achievable sub-objectives.

Furthermore, the paper "Voyager: An Open-Ended Embodied Agent with Large Language Models" demonstrates that it is possible to use GPT-4 for planning in Natural language in Minecraft. ([Wang et al., 2023](#))

1.9.4 Differences with the brain

It appears that there are several points of convergence between the LLMs and the linguistic cortex:

- **Behavioral similarities.** From ([Canell, 2022](#)), it's highlighted that LLMs show a close comparison to human linguistic abilities and the linguistic cortex. These models have excelled in mastering syntax and a significant portion of semantics in human language. Of course, today, they still lag in aspects such as long-term memory, coherence, and general reasoning - faculties that in humans depend on various brain regions like the hippocampus and prefrontal cortex, but we explained in the last sections that those problems may be solvable.
- **Convergence in internal Representations :** LLMs have a representation that converges with scale toward the brain representation. This is supported by the study, "Brains and algorithms partially converge in natural language processing." ([Caucheteux & King, 2022](#)) Additional insights can be found in the works "The Brain as a Universal Learning Machine" ([Canell, 2015](#)) and "Brain Efficiency: Much More than You Wanted to Know." ([Canell, 2022](#)) At comparable learning stages, LLMs and the linguistic cortex develop similar or equivalent feature representations. In some evaluations, advanced LLMs have been able to predict 100% of the explainable neural variance, as detailed by Schrimpf, Martin, et al. in "The neural architecture of language: Integrative modeling converges on predictive processing." ([Schrimpf et al., 2021](#))
- **Scale is also important in primates.** The principal architectural difference between human and other primate brains seems to be the number of neurons rather than anything else, as demonstrated in various studies. ([Houzel, 2012](#); [Pearson et al., 2023](#); [Charvet, 2021](#)).

1.9.5 Further reasons to continue scaling LLMs

Following are some reasons to believe that labs will continue to scale LLMs.

Scaling Laws on LLM implies further qualitative improvements. The scaling laws might not initially appear impressive. However, linking these quantitative measures can translate to a qualitative improvement in algorithm quality. An algorithm that achieves near-perfect loss, though, is one that necessarily comprehends all subtleties, and displays enormous adaptability. The fact that the scaling laws are not binding is very significant and means that we can make the model a qualitatively better reasoner.

From simple correlations to understanding. During a training run, GPTs go from basic correlations to deeper and deeper understanding. Initially, the model merely establishes connections between successive words. Gradually, it develops an understanding of grammar and semantics, creating links between sentences and subsequently between paragraphs. Eventually, GPT masters the nuances of writing style¹.

Exercise: Scaling Laws on LLM implies further qualitative improvements.

Let's calculate the difference in loss, measured in bits, between two model outputs: "Janelle ate some ice cream because he likes sweet things like ice cream." and "Janelle ate some ice cream because she likes sweet things like ice cream." The sentence contains approximately twenty tokens. If the model vacillates between "He" or "She," choosing randomly (50/50 odds), it incurs a loss of 2 bits on the pronoun token when incorrect. The loss for other tokens remains the same in both models. However, since the model is only incorrect half the time, a factor of 1/2 should be

¹See also "The Scaling Hypothesis," to delve into this progression in a fascinating story.

applied. This results in a difference of $(1/2) * (2/20) = 1/20$, or 0.05 bits. Thus, a model within 0.05 bits of the minimal theoretical loss should be capable of understanding even more nuanced concepts than the one discussed above.

Text completion is probably an AI-complete test ([Wikipedia, 2022](#)).

Current LLMs have only as many parameters as small mammals have synapses, no wonder they are still imperfect. Models like GPT-4, though very big compared to other models, should be noted for their relatively modest scale compared to the size of a human brain. To illustrate, the largest GPT-3 model has a similar number of parameters to the synapses of a hedgehog. We don't really know how many parameters GPT-4 has, but if it is the same size as PALM, which has 512 B parameters, then GPT-4 has only as many parameters as a chinchilla has synapses. In contrast, the human neocortex contains about 140 trillion synapses, which is over 200 times more synapses than a chinchilla. For a more in-depth discussion on this comparison, see the related discussion [here](#). For a discussion of the number of parameters necessary to emulate a synapse, see the discussion on biological anchors.

GPT-4 is still orders of magnitude cheaper than other big science projects. : Despite the high costs associated with training large models, the significant leaps in AI capabilities provided by scaling justify these costs. For example, GPT-4 is expensive compared to other ML models. It is said to cost 50M in training. But the Manhattan Project cost 25B, which is 500 times more without accounting for inflation, and achieving Human-level intelligence, may be more economically important than achieving the nuclear bomb.

Collectively, these points support the idea that AGI can be achieved by only scaling current algorithms.

1.10 Appendix: Forecasting - Trends

Everything in the appendices is optional, and is meant as additional knowledge and context. You do not need to read this to understand the core points made in either this chapter or future chapters.

Generally, the three main components recognized as the main variables of advancement in deep learning are: computational power available, algorithmic improvements, and the availability of data. These three variables are also sometimes called the inputs to the AI production function, or the AI triad. ([Buchanan, 2022](#))

We can anticipate that models will continue to scale in the near future. Increased scale combined with the increasingly general-purpose nature of foundation models could potentially lead to a sustained growth in general-purpose AI capabilities.

1.10.1 Compute

The first thing to look at is the trends in the overall amount of training compute required when we train our model. Training compute grew by 1.58 times/year up until the Deep Learning revolution around 2010, after which growth rates increased to 4.2 times/year. We also find a new trend of "large-scale" models that emerged in 2016, trained with 2-3 OOMs (Order Of Magnitude) more compute than other systems in the same period.

Hardware advancements are paralleling these trends in training compute and data. GPUs are seeing a yearly 1.35 times increase in floating-point operations per second (FLOP/s). However, memory constraints

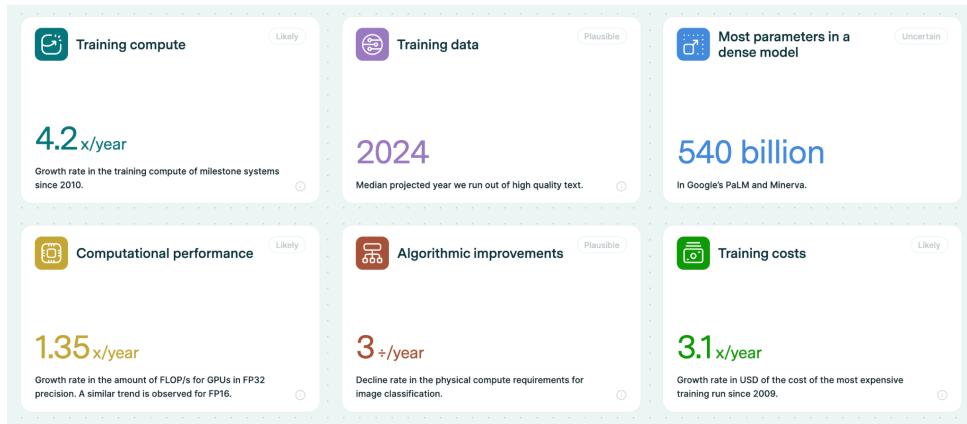


Figure 1.44 : Key trends and figures in Machine Learning (*Epoch, 2023*)

are emerging as potential bottlenecks, with DRAM capacity and bandwidth improving at a slower rate. Investment trends reflect these technological advancements

In 2010, before the deep learning revolution, DeepMind co-founder Shane Legg predicted human-level AI by 2028 using compute-based estimates (Legg, 2010). OpenAI co-founder Ilya Sutskever, whose AlexNet paper sparked the deep learning revolution, was also an early proponent of the idea that scaling up deep learning would be transformative.

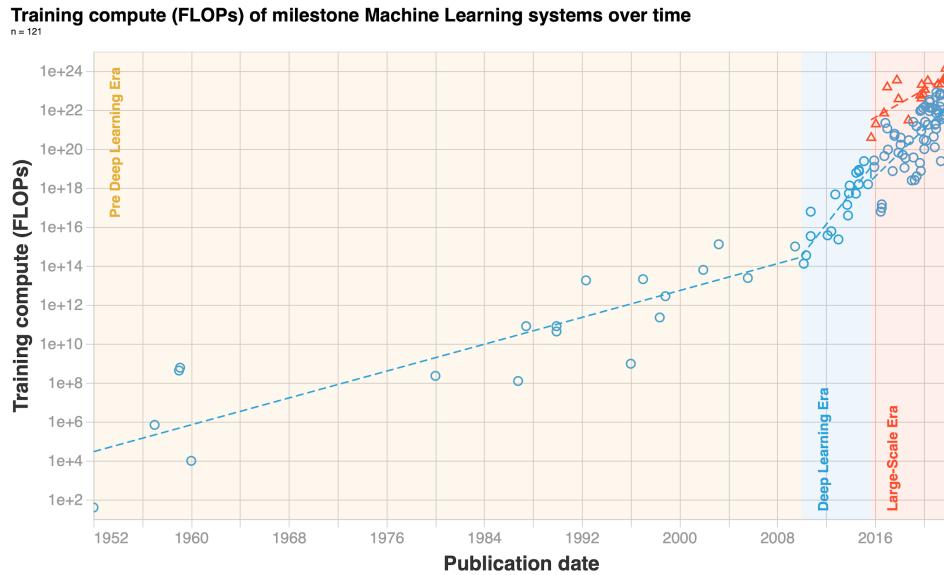


Figure 1.45 : Key trends and figures in Machine Learning (*Epoch AI, 2023*)

1.10.2 Parameters

In this section, let's look at the trends in model parameters. The following graph shows how even though parameter counts have always been increasing, in the new 2018+ era, we have really entered a different phase of growth. Overall, between the 1950s and 2018, models have grown at a rate of 0.1 orders of magnitude per year (OOM/year). This means that in the 68 years between 1950 and 2018 models grew by a total of 7 orders of magnitude. However, post-2018, in just the last 5 years models have increased by yet another 4 orders of magnitude (not accounting for however many parameters GPT-4 has because we don't know).

The following table and graph illustrate the trend change in machine learning models' parameter growth. Note the increase to half a trillion parameters with constant training data.

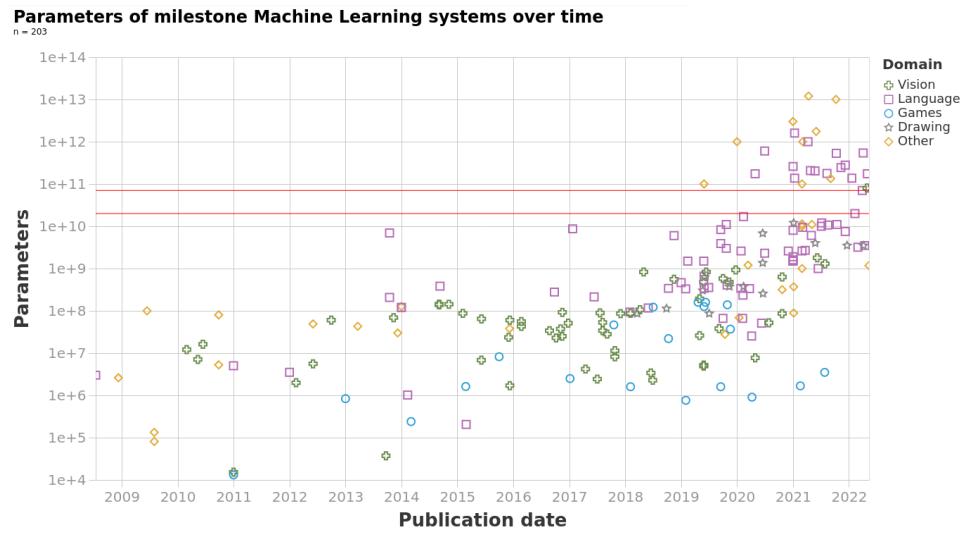


Figure 1.46 : Machine Learning Model Sizes and the Parameter Gap ([Villalobos et al., 2022](#))

1.10.3 Data

We are using ever-increasing amounts of data to train our models. The paradigm of training foundation models to fine-tune later is accelerating this trend. If we want a generalist base model then we need to provide it with ‘general data’, which means: all the data we can get our hands on. You have probably heard that models like ChatGPT and PaLM are trained on data from the internet. The internet is the biggest repository of data that humans have. Additionally, as we observed from the Chinchilla papers scaling laws, it is possible that data to train our models is the actual bottleneck, and not compute or parameter count. So the natural question is how much data is left on the internet for us to keep training our models? and how much more data do we humans generate every year?

How much data do we generate? The total amount of data generated every single day in 2019 was on the order of 463EB ([World Economic Forum, 2019](#)). But in this post, we will assume that models are not training on ‘all the data generated’ (yet), rather they will continue to only train on open-source internet text and image data. The available stock of text and image data grew by 0.14 OOM/year between 1990 and 2018 but has since slowed to 0.03 OOM/year.

How much data is left? The median projection for when the training dataset of notable ML models exhausts the stock of professionally edited texts on the internet is 2024. The median projection for the year in which ML models use up all the text on the internet is 2040. Overall, projections by Epochai predict that we will have exhausted high-quality language data before 2026, low-quality language data somewhere between 2030 and 2050, and vision data between 2030 and 2060. This might be an indicator of slowing down ML progress after the next couple of decades. These conclusions from Epochai, like all the other conclusions in this entire leveraging computation section, rely on the unrealistic assumptions that current trends in ML data usage and production will continue and that there will be no major innovations in data efficiency, i.e. we are assuming that the amount of capabilities gained per training datapoint will not change from current standards.

Even if we run out of Data, many solutions are proposed, from using synthetic data, for example, filtering and preprocessing the data with GPT-3.5 to create a new cleaner dataset, an approach used in the paper “Textbooks are all you need” with models like Phi 1.5B that demonstrate excellent performance for their size through the use of high-quality filtered data, to the use of more efficient trainings, or being more

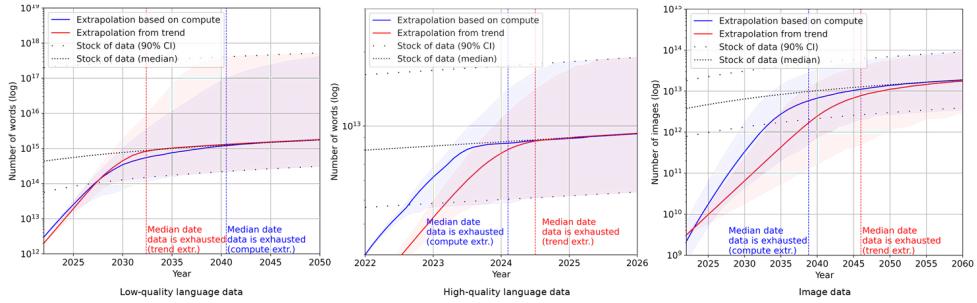


Figure 1.47 : ML data consumption and data production trends for low-quality text, high-quality text, and images. (Epoch AI, 2023)

efficient by training on more epochs.

1.10.4 Algorithms

Algorithmic advancements also play a role. For example, between 2012 and 2021, the computational power required to match the performance of AlexNet has been reduced by a factor of 40, which corresponds to a threefold yearly reduction in the compute required for achieving the same performance on image classification tasks like ImageNet. Improving the architecture also counts as algorithmic advancement. A particularly influential architecture is that of Transformers, central to many recent innovations, especially in chatbots and autoregressive learning. Their ability to be trained in parallel over every token of the context window fully exploits the power of modern GPUs, and this is thought to be one of the main reasons why they work so well compared to their predecessor, even if this point is controversial.

Does algorithmic architecture really matter?

This is a complicated question, but some evidence suggests that once an architecture is expressive and scalable enough, the architecture matters less than we might have thought:

In a paper titled ‘ConvNets Match Vision Transformers at Scale,’ Google researchers found that Visual Transformers (ViT) can achieve the same results as CNNs (Convolutional neural network) simply by using more compute. (Smith et al., 2023) They took a special CNN architecture and trained it on a massive dataset of four billion images. The resulting model matched the accuracy of existing ViT systems that used similar training compute.

Variational AutoEncoders (long an also-ran to GANs or autoregressive models in terms of image generation) catch up if you make them very deep (Child, 2021; Vahdat & Kautz, 2021).

Progress in late 2023, such as the mamba architecture (Gu & Dao, 2023), appears to be an improvement on the transformer. It can be seen as an algorithmic advancement that reduces the amount of training computation needed to achieve the same performance.

The connections and normalizations in the transformer, which were thought to be important, can be taken out if the weights are set up correctly. This can also make the transformer design simpler (Note however that this architecture is slower to converge than the others). (He et al., 2023)

On the other side of the argument, certain attention architectures are significantly more scalable when dealing with long context windows, and no feasible amount of training could compensate for this in more basic transformer models. Architectures specifically designed to handle long sequences, like Sparse Transformers (Child et al., 2019) or Longformer (Beltagy et al., 2020), can outperform standard transformers by a considerable margin for this usage. In computer vision, architectures like CNNs are inherently structured to recognize spatial hierarchies in images, making them more efficient for these tasks than architectures not specialized in handling spatial data when the amount of data is limited, and the "prior" encoded in the architecture makes the

model learn faster.

1.10.5 Costs

Understanding the dollar cost of ML training runs is crucial for several reasons. Firstly, it reflects the real economic expense of developing machine learning systems, which is essential for forecasting the future of AI development and identifying which actors can afford to pursue large-scale AI projects. Secondly, by combining cost estimates with performance metrics, we can track the efficiency and capabilities of ML systems over time, offering insights into how these systems are improving and where inefficiencies may lie. Lastly, these insights help determine the sustainability of current spending trends and guide future investments in AI research and development, ensuring resources are allocated effectively to foster innovation while managing economic impact.

Moore's Law, which predicts the doubling of transistor density and thus computational power approximately every two years, has historically led to decreased costs of compute power. However, the report finds that spending on ML training has grown much faster than the cost reductions suggested by Moore's Law. This means that while hardware has become cheaper, the overall expense of training ML systems has escalated due to increasing demand for computational resources. This divergence emphasizes the rapid pace of advancements in ML and the significant investments required to keep up with the escalating computational demands.

To measure the cost of ML training runs, the report employs two primary methods. The first method uses historical trends in the price-performance of GPUs to estimate costs. This approach leverages general trends in hardware advancements and cost reductions over time. The second method bases estimates on the specific hardware used to train the ML systems, such as NVIDIA GPUs, providing a more detailed and accurate picture of the costs associated with particular technologies. Both methods involve calculating the hardware cost—the portion of the up-front hardware cost used for training—and the energy cost, which accounts for the electricity required to power the hardware during training. These calculations provide a comprehensive view of the economic burden of training ML models.

Measuring the cost of development extends beyond the final training run of an ML system to encompass a range of factors. This includes research and development costs, which cover the expenditures on preliminary experiments and model refinements that lead up to the final product. It also involves personnel costs, including salaries and benefits for researchers, engineers, and support staff. Infrastructure costs, such as investments in data centers, cooling systems, and networking equipment, are also significant. Additionally, software and tools, including licenses and cloud services, contribute to the overall cost. Energy costs throughout the development lifecycle, not just during the final training run, and opportunity costs—potential revenue lost from not pursuing other projects—are also crucial components. Understanding these broader costs provides a more comprehensive view of the economic impact of developing advanced ML systems, informing strategic decisions about resource allocation.

The findings suggest that the cost of ML training runs will continue to grow, but the rate of growth might slow down in the future. The report estimates that the cost of ML training has grown by approximately 2.8 times per year for all systems. For large-scale systems, the growth rate is slower, at around 1.6 times per year. This substantial year-on-year increase in training costs highlights the need for significant efficiency improvements in both hardware and training methodologies to manage future expenses effectively.

The report forecasts that if current trends continue, the cost for the most expensive training runs could exceed significant economic thresholds, such as 1% of the US GDP, within the next few decades. This implies that without efficiency improvements, the economic burden of developing state-of-the-art ML systems will increase substantially. Consequently, understanding and managing these costs is essential for ensuring the sustainable growth of AI capabilities and maintaining a balanced approach to AI investment and development.

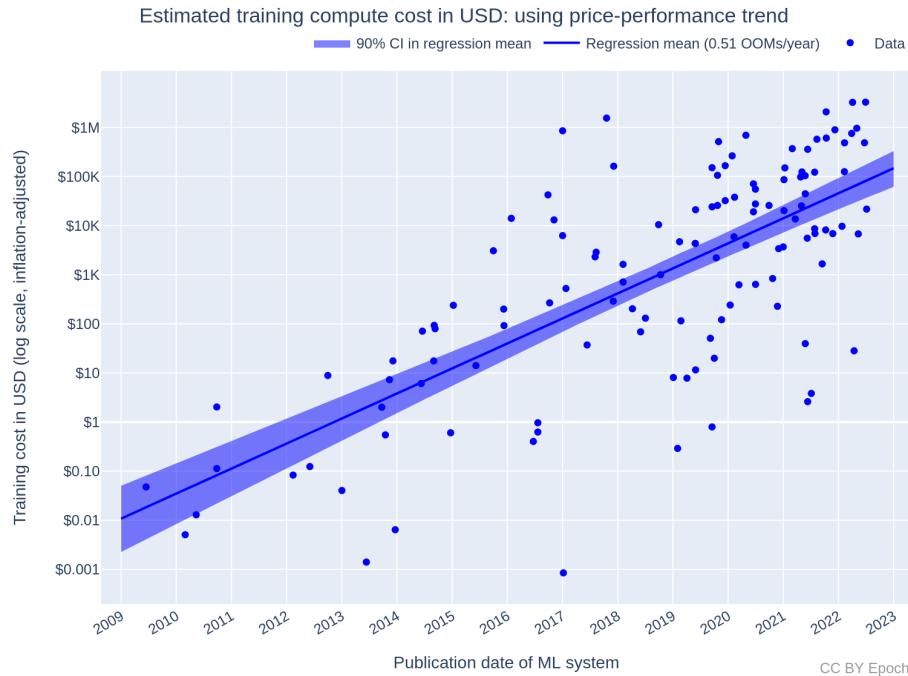


Figure 1.48 : Estimated cost of compute in US dollars for the final training run of ML systems. ([Epoch AI, 2023](#))

1.11 Appendix: Forecasting - Bio Anchors

Everything in the appendices is optional, and is meant as additional knowledge and context. You do not need to read this to understand the core points made in either this chapter or future chapters.

1.11.1 Evolution Anchor

This anchor quantifies the computational effort invested by evolution in shaping the human brain. It considers the vast amount of processing and learning that has taken place from the emergence of the first neurons to the development of the modern human brain. This method suggests that evolution has served as a form of "pre-training" for the human brain, enhancing its ability to adapt and survive. To estimate the computational power of this evolutionary "pre-training", the report considers the total amount of compute used by all animal brains over the course of evolution. This includes not just the brains of humans, but also those of our ancestors and other animals with nervous systems. The idea is that all of this brain activity represents a form of learning or adaptation that has contributed to the development of the modern human brain. While the exact calculations involved in this estimate are complex and subject to considerable uncertainty, the basic idea is to multiply the number of animals that have ever lived by the amount of compute each of their brains performed over their lifetimes. This gives an estimate of the total compute performed by all animal brains over the course of evolution.

Cotra accounts for these considerations and assumes that the "average ancestor" performed as many FLOP/s as a nematode, and that there were on average $1e21$ ancestors at any time. This yields a **median of $1e41$ FLOP**, which seems extraordinarily high compared to modern machine learning. As an example, Google's PaLM model was trained with $2.5e24$ FLOP (17 orders of magnitude smaller). She

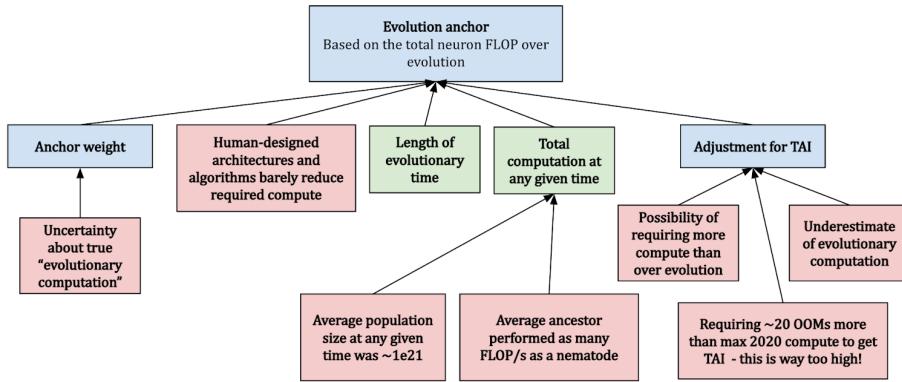


Figure 1.49 : The Evolution anchor (Ho, 2022)

gives this anchor a **weight of 10%**. (Ho, 2022)

</tab>

1.11.2 Lifetime Anchor

This refers to the total computational activity the human brain performs over a human lifetime. This anchor is essentially a measure of the "training" a human brain undergoes from birth to adulthood and incorporates factors such as the number of neurons in the human brain, the amount of computation each neuron performs per year, and the number of years it takes for a human to reach adulthood. The human brain has an estimated 86 billion neurons. Each of these neurons performs a certain number of computations per second, which can be calculated as a certain number of operations per second in FLOP/s. When calculating the total amount of compute over a lifetime, these factors are multiplied together, along with the number of years a human typically lives.

For example, if we assume that a neuron is able to perform about 1000 operations per second, and there are about 31.5 million seconds in a year, then a single neuron would perform about 31.5 billion operations in a year. Now, if we multiply this by the estimated number of neurons in the human brain (86 billion), we get an estimate of the total brain-compute performed in one year. We can then multiply this by the number of years in a typical human lifespan to estimate the total brain-compute over a lifetime. Plugging in the numbers about brain FLOP/s seems to suggest that 1e27 FLOP would be required to reach TAI. This seems low since examples from other technological domains suggest that the efficiency of things we build (on relevant metrics) is generally not great when compared to nature.

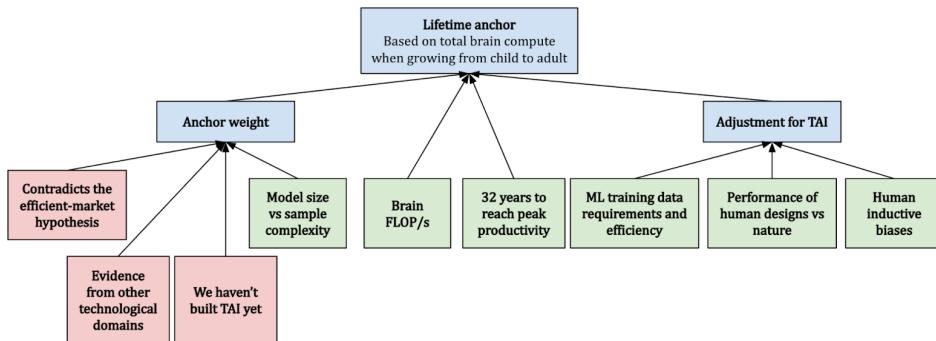


Figure 1.50 : The Lifelong anchor (Ho, 2022)

Cotra's report finds a median of 1e28 FLOP and only gives the lifetime anchor a 5% weight, indicating

that it's not the most influential factor in the overall model. The report acknowledges the inherent uncertainties and complexities involved in such a calculation and uses this anchor as one of several to provide a range of estimates for the compute required to achieve AI with human-level performance. (Ho, 2022)

</tab>

1.11.3 Neural Network Anchors

Each of the neural network anchors serves to provide a different perspective on the amount of compute that might be required to train a TAI. There are three Neural Network Anchors presented in the report: long (1e37 FLOP), medium (1e34 FLOP), and short horizon (1e32 FLOP). These anchors hypothesize that the ratio of parameters to compute used by a TAI should be similar to the ratio observed in today's neural networks. Additionally, a TAI should perform approximately as many FLOPs per subjective second as a human brain. A "subjective second" is the time it takes a model to process as much data as a human can in one second (Ho, 2022). As an example a typical human reads about 3-4 words per second for non-technical material, so "one subjective second" for a language model would correspond to however much time that the model takes to process about 3-4 words of data. (Cotra, 2020) Cotra determines the training data requirements based on a mix of machine learning theory and empirical considerations. She puts 15% weight on short horizons, 30% on medium horizons, and 20% on long horizons, for a total of 65% on the three anchors. (Ho, 2022)

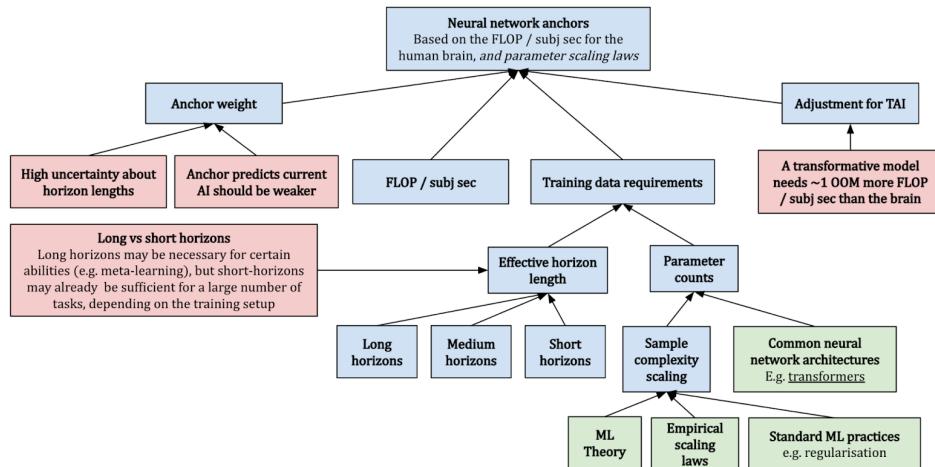


Figure 1.51 : Neural network Anchors (Ho, 2022)

1.11.4 Genome Anchor

The genome anchor looks at the FLOP/subj sec of the human brain and expects TAI to require as many parameters as there are bytes in the human genome. This hypothesis implicitly assumes a training process that's structurally analogous to evolution, and that TAI will have some critical cognitive ability that evolution optimized for. This differs from the evolution anchor in that it assumes we can search over possible architectures/algorithms a lot more efficiently than evolution, using gradients. Due to this structural similarity, and because feedback signals about the fitness of a particular genome configuration are generally sparse, this suggests that the anchor only really makes sense with long horizon lengths. (Ho, 2022)

At least at the time of writing (May 2022), machine learning architectures don't look very much like the human genome, and we are yet to develop TAI – thus Cotra updates against this hypothesis towards

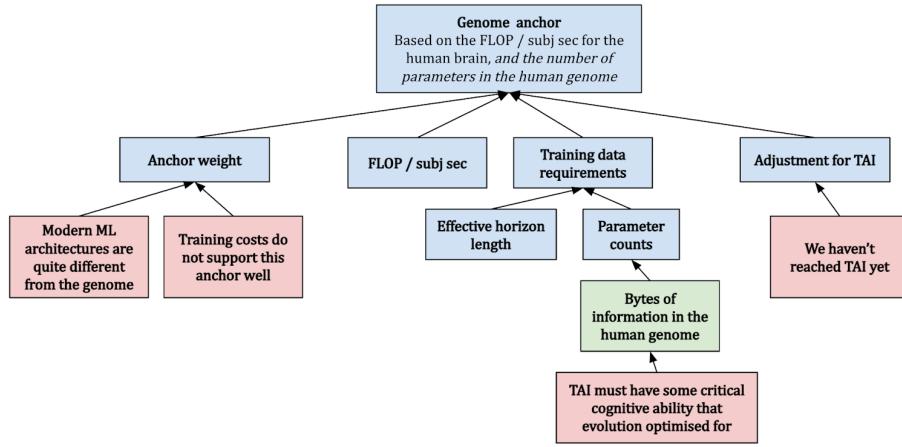


Figure 1.52 : The Genome Anchor ([Ho, 2022](#))

requiring more FLOP. Overall, she finds a median of $1e33$ FLOP and places 10% weight on this anchor. ([Ho, 2022](#))