Chapter 1 - Artificial General Intelligence

1.0: Foundation Models **3**
2.0: Leveraging Computation **6**
2.1: The Bitter Lesson 6
2.2: Compute trends 7
2.3: Scaling Laws 8
3.0: Capabilities **12**
3.1: Capabilities vs. Intelligence 12
3.2: Definitions of advanced AI Systems 13
3.3: (t,n)-AGI 14
3.3: Formalizing Capabilities 14
4.0: Threat Models **16**
4.1: Intelligence Explosion 16
4.2: Instrumental Convergence 17
4.3: Emergence 17
4.4: Four Background Claims 18
5.0: Timelines & Forecasting **19**
5.1: Biological Anchors 20
6.0: Takeoff Dynamics **23**
6.1: Speed/Continuity 23
6.2: Homogeneity 25
6.3: Polarity 26
7.0: [WIP] Extra **26**
Data 27
Time (Grokking) 28
Compute 28
Size (Parameters) 29
Paradigm Shifts 31
Exercises & Flashcards **31**
Sources **31**
To Add **33**

Overview

1. **Foundation Models**: The chapter begins with an exploration of how contemporary approaches in machine learning lean towards the development of centralized foundation models. The section elaborates the merits and drawbacks of such a paradigm.

2. **Leveraging Computation**: This section introduces the reader to “the bitter lesson". The focus for this section is comparing historical advancements achieved through the utilization of human-engineered heuristics with those accomplished by capitalizing on additional computation. This is followed by a discussion on current trends in the “compute optimal training” of machine learning models. This section concludes with an introduction to and the implications of scaling laws and the scaling hypothesis.

3. **Capabilities**: This section builds upon the previously introduced trends and paradigms, and extrapolates these to predict potential capabilities of future artificial intelligence (AI) models. There is a discussion around the merits of using the term capabilities instead of intelligence. This is followed by introducing slightly more detailed frameworks for different possible tiers and categorizations of artificial general intelligence (AGI). Moreover, the concept of (t,n)-AGI is introduced. This outlook allows a straightforward comparison to humans, while also establishing a measurable continuous spectrum of capabilities. Overall the aim is to help establish a more concrete definition of AGI capabilities for the reader.

4. **Threat Models**: Understanding capability thresholds paves the way for a discussion into the concept of emergence. This is then followed by an examination of qualities that machine intelligences might possess. These qualities potentially indicate the possibility for an intelligence explosion. The section concludes with a discussion of the four fundamental assumptions put forward by the Machine Intelligence Research Institute (MIRI) about machine intelligence. These claims explore the power of general intelligence, and why this capability arising in machines does not promise a beneficial future to humans by default.

5. **Timelines**: This section explores some concrete predictions of when the capabilities discussed in the previous sections might surface. The dialogue hinges on the concept of anchors in forecasting. This pays specific focus on determining how we can use anchors inspired by biological systems to provide a basis for estimating the computational requirements of AI systems.

6. **Takeoff**: The chapter concludes with a section that introduces the concept of takeoff and various forms of takeoff dynamics. The dynamics involve takeoff speeds, polarity and homogeneity. The section presents differing opinions from various researchers on potential future scenarios.

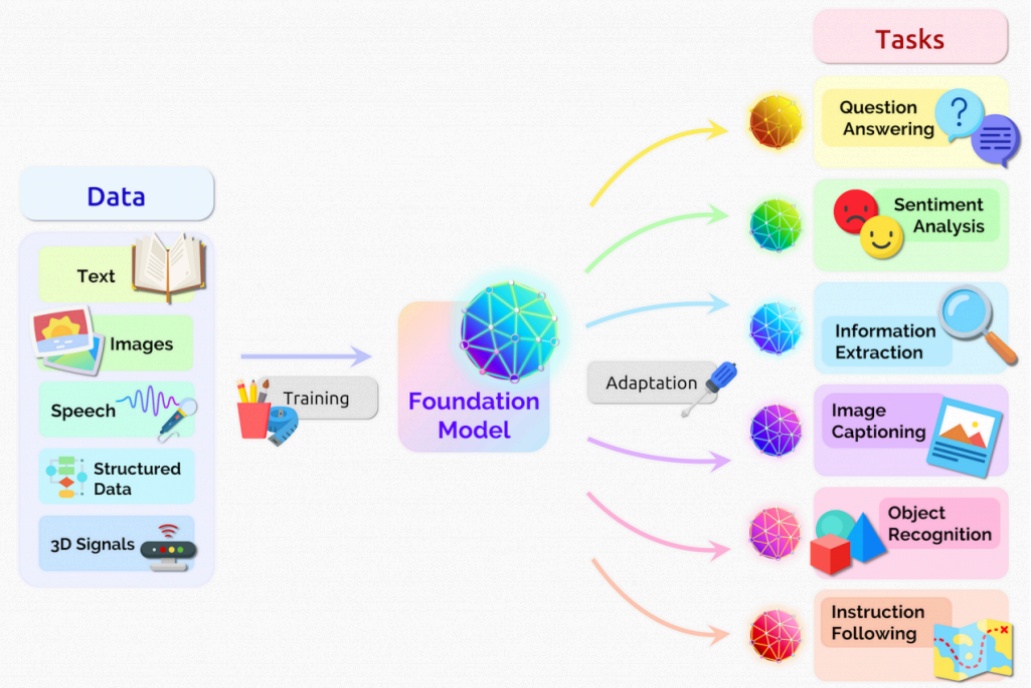
1.0: Foundation Models

*A **foundation model** is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.*
- Bommasani Rishi et. al. (2022) "On the Opportunities and Risks of Foundation Models"

Large-scale research labs have been navigating towards a new machine learning (ML) paradigm: the training of a "base model". A base model is simply a large-scale model that serves as a building block for various applications. This model serves as a multifaceted entity—competent across various areas but not specialized in any one.

Once there is a base, developers can retrain portions of the model using additional examples, which allows operators to generate specialized systems adapted for specific tasks and improve the system's quality and consistency. This is known as **fine-tuning**.

Fine-tuning facilitates the development of models capable of diverse downstream tasks. Simple examples of this include fine-tuning the general purpose GPT language model to follow instructions, or to interact in a chat format. Other examples include specializing models for programming, scientific texts, mathematical proofs and other such tasks.



Source: Bommasani Rishi et. al. (2022) "On the Opportunities and Risks of Foundation Models"

The costs for developing models from scratch is also increasing due to a multitude of factors. If models were trained on the supervised learning (SL) paradigm, then the developers must either already have a large dataset, or in scenarios where the necessary dataset does not already exist, they must generate their own data, directing precious resources—both monetary and temporal—toward the careful labeling of the data.

***SSL (Semi-Supervised Learning)** is a learning approach that combines labeled and unlabeled data during training to improve the performance of machine learning models.*

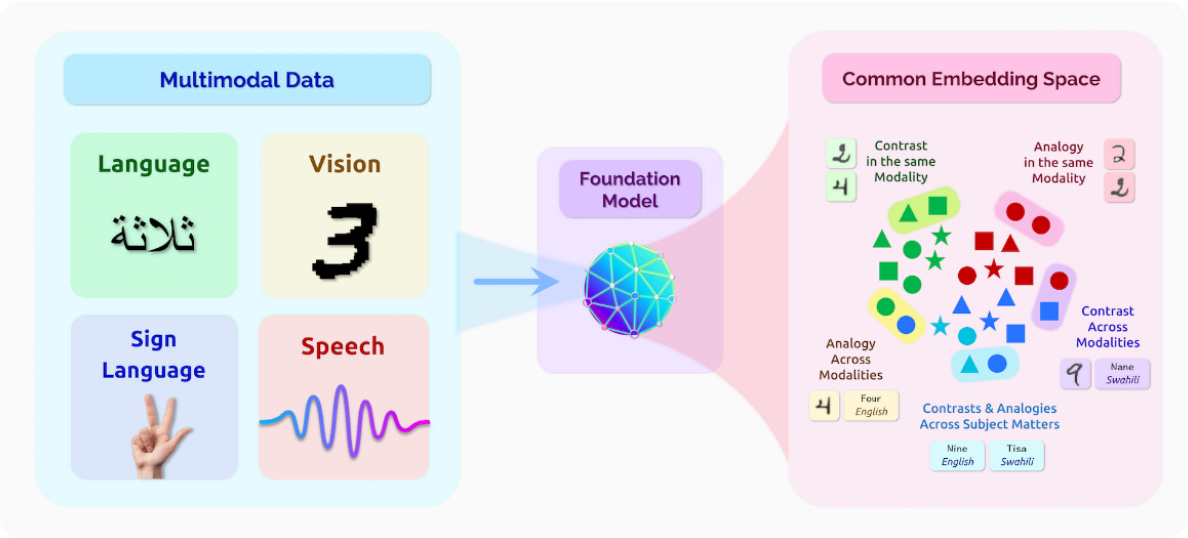
Achieving state-of-the-art performance across numerous tasks demands that the learning process be anchored in millions, if not billions, of examples. Foundation models provide a solution to this by leveraging Semi-Supervised Learning (SSL). SSL algorithms use both labeled and unlabeled data during training. This allows the models to utilize the information present in the unlabeled examples to improve performance. The intuition behind SSL is that the unlabeled data contains valuable information about the underlying structure of the data, which can be used to enhance the model's generalization capabilities. By incorporating the unlabeled data, SSL algorithms aim to learn a more robust and accurate model compared to SL algorithms, especially when labeled data is limited or expensive to obtain. Once the foundation model is trained, fine-tuning allows it to specialize and perform well on specific downstream tasks by using far fewer SL labeled examples. By fine-tuning the model on a smaller labeled dataset, the model can leverage the knowledge it acquired during SSL training and adapt it to the specific task, resulting in overall improved performance.

Fine-tuning foundation models might be cheaper than training a model from scratch, however the cost to train the base model itself keeps increasing. Foundation models are extremely complex and require significant resources to develop, train, and deploy. Training can be extremely expensive, often involving tens of thousands of GPUs running continuously for months. These models are typically trained in specialized clusters and using carefully designed software systems. Such dedicated clusters can be both costly and difficult to obtain. There have also been recent efforts to mitigate the costs by training foundation models in a decentralized manner in heterogeneous environments. For narrowly-defined use-cases, that cost may not be justifiable, when a smaller model may achieve similar (or better) results for a much lower price.

Pre-training *in the context of foundation models refers to the initial phase where a model is trained on a large, unlabeled dataset to learn general knowledge and patterns before fine-tuning it on specific tasks.*

Transfer learning *in the context of foundation models refers to the process of leveraging knowledge and patterns learned from a related task or domain with abundant labeled data to improve performance on a target task or domain with limited labeled data.*

Leveraging advancements in transfer learning and fine-tuning techniques, these foundation models can be harnessed to spawn specialized models tailored for specific objectives. This advancement amplifies the field's capacity to transfer acquired "knowledge" from one task and apply it effectively through the fine-tuning process to a distinct downstream task. Some notable foundation models include BERT, GPT-3, GPT-4, GATO and CLIP.



Source: Bommasani Rishi et. al. (2022) "On the Opportunities and Risks of Foundation Models"

This novel paradigm potentially provides a larger demographic with access to state-of-the-art capabilities, as well as the potential to train their own models with minimal data for highly specialized tasks. This potential access to capabilities is not guaranteed however, does depend on the specific API options available, or on the availability of open-sourced foundation models that the users can rely upon.

Broadly speaking, there exist significant economic incentives to expand the capabilities and scale of foundation models. The authors of "On the Opportunities and Risks of Foundation Models" foresee steady technological progress in the forthcoming years. Although foundation models presently manifest most robustly in natural language processing (NLP), this can be interpreted as a trend toward a new general paradigm of AI development. As of January 2023, efforts by DeepMind to train a reinforcement learning (RL) foundation model—an "adaptive agent" (AdA)—have also been undertaken. These RL agents are trained in an open ended task space (XLand 2.0) which require different skill sets such as experimentation, tool use or division of labor. If language-based foundation models are general-purpose text generators, then the AdA model could conceivably be viewed as a relatively more general-purpose task follower compared to other models observed thus far.

However, this paradigm also carries inherent risks, namely the emergence of capabilities and homogenization.

- **Homogenization**: Since an increasing number of models are becoming merely "fine-tuned" versions of foundation models, it follows that downstream AI systems might inherit the same problematic biases prevalent in a few foundation models. Thus, all failure categories present in the base model could potentially percolate through all models trained with this as the foundation.
- **Emergence**: Homogenization could potentially provide enormous gains for many domains, but aggressive homogenization of capabilities might result in unexpected and unexplainable behavior arising as a function of scale. Emergence implies that a system's behavior is implicitly induced rather than explicitly constructed. These characteristics render the models challenging to understand. They also give rise to unforeseen failure modes and unanticipated consequences. This phenomenon is talked about in more detail below.

2.0: Leveraging Computation

Although fine-tuning and transfer learning are the mechanisms that render foundation models feasible, it is scale that makes them truly powerful. This section delves into the concept of model scaling and the leveraging of computation—advancements in computer hardware (e.g., GPU throughput and memory), new architectures (e.g. the transformer), and the availability of increasing amounts of training data.

2.1: The Bitter Lesson

> *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 teaches us] the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great.*
- Sutton, Rich (March 2019) "The Bitter Lesson"

Traditionally, AI research has predominantly designed systems under the assumption that a fixed amount of computing power will be available to the designed agent. However, over time, computing power so far has been expanding in line with Moore's law (number of transistors in an integrated circuit doubles every 1.5 years). Consequently, researchers can either leverage their human knowledge of the domain or exploit increases in general-purpose computational methods. Theoretically, the two are mutually compatible, but in practice, the human-knowledge approach tends to complicate methods, rendering them less suited to harnessing general methods that leverage computation.

Several instances in history underscore this bitter lesson for AI researchers:

- **Games**: Deep Blue defeated chess world champion Garry Kasparov by leveraging a vast deep search, disheartening computer-chess researchers who had pursued methods that capitalized on the human understanding of chess's unique structure. Similarly, AlphaGo triumphed over Go world champion Lee Sedol using deep learning combined with a Monte Carlo tree search for move selection, eschewing human-engineered Go techniques. Within a year, AlphaZero, forsaking any human-generated Go data, used self-play to defeat AlphaGo. None of these successive enhancements in game-playing capabilities hinged on any fundamental breakthroughs in human Go knowledge.

- **Vision**: A similar pattern has unfolded in computer vision. Earlier methods employed human-engineered features and convolution kernels for image recognition tasks. However, over the years, it has been determined that leveraging more computation and permitting convolutional neural nets (CNNs) to learn their own features yield superior performance.

- **Language & Speech**: In 1970, the DARPA SUR (Speech Understanding Research) was held. One faction endeavored to leverage expert knowledge of words, phonemes, the human vocal tract, etc. In contrast, the other side employed newer, more statistical methods that necessitated considerably more computation, based on hidden Markov models (HMMs). This example shows yet again, that the statistical methods surpassed the human-knowledge-based methods. Since then, deep learning

recurrent neural network-based or transformer-based methods have virtually dominated the field of sequence-based tasks.

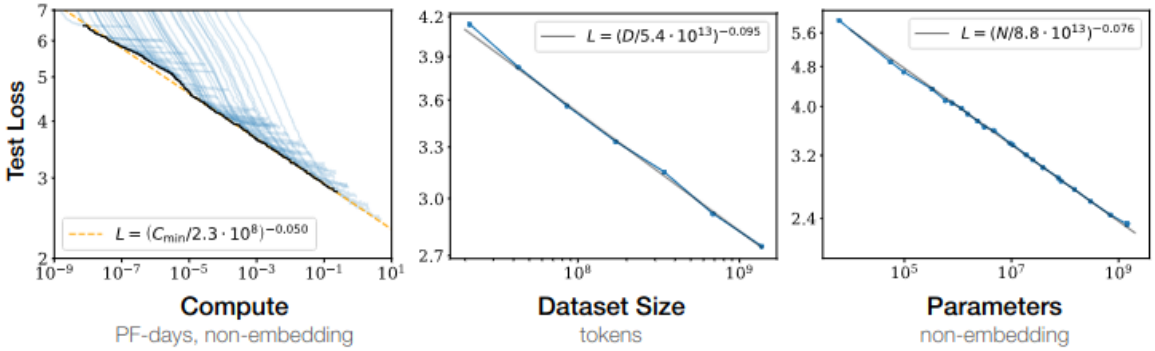
Historically, due to repeated reminders of the bitter lesson, the field of AI has increasingly learned to favor general-purpose methods of search and learning. This trend fortifies the intuition behind the immense scale of the current foundation model paradigm. It can be projected that the capabilities of the current foundation models will continue to scale commensurately with increasing computation. The reasons for this claim are presented in the following sections. The immediately ensuing section delves into these trends of scale in compute, dataset size, and parameter count.

2.2: Compute trends

Several key factors dictate the relationship between the scale and capability of current ML models:

- **Compute**: Extended training runs (measured in epochs) generally result in lower loss. The total computational power needed partially depends on the training duration. ML engineers typically aim for asymptotically diminishing returns before halting the training process.
- **Dataset size**: The larger the training dataset, the more information the model can analyze during each training run. As a result, training runs are generally longer, which in turn increases the total computational power needed before the model can be deemed "trained."
- **Parameter Count**: For each training example, the model needs to calculate the loss and then use backpropagation to update all relevant parameters. The more parameters the model has, the more computation-intensive this process becomes.

Below is a chart illustrating the impact of each of these three factors on model loss.[1]



Source: Kaplan, Jared et. al. (Jan 2020) "Scaling Laws for Neural Language Models"

With graphical processing units (GPUs), and tensor processing units (TPUs) improving in performance and reducing in cost annually, AI models are demonstrating increasingly impressive results. This leads to higher acceptance of substantial compute costs. The reduced cost of computation, coupled with the paradigm of foundation models trained on escalating volumes of data, suggests that all three variables—compute, dataset size, and parameter count—will continue to expand in the forthcoming years. However, it remains an open question whether merely scaling these factors will result in unmanageable capabilities.

[1] Epoch AI has many more graphs of this kind.

The following example offers a tangible illustration of capabilities escalating with an increasing parameter count in image generation models. The same model architecture (Parti) is used to generate an image using an identical prompt, with the sole difference between the models being the parameter size.

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!*

Source: GoogleAI (2022) , "*Parti (Pathways Autoregressive Text-to-Image model)*"

Increased numbers of parameters not only enhance image quality but also aid the network in generalizing in various ways. More parameters enable the model to generate accurate representations of complex elements, such as hands and text, which are notoriously challenging. There are noticeable leaps in quality, and somewhere between 3 billion and 20 billion parameters, the model acquires the ability to spell words correctly. Parti is the first model with the ability to spell correctly. Before Parti, it was uncertain if such an ability could be obtained merely through scaling, but it is now evident that spelling correctly is another capability gained simply by leveraging scale.

The following section briefly introduces efforts by both OpenAI and DeepMind to formalize the relationships between scale and capabilities.

2.3: Scaling Laws

Scaling laws articulate the relationship between compute, dataset size, parameter count, and model capabilities. They're employed to scale models effectively and optimally allocate resources with respect to capabilities.

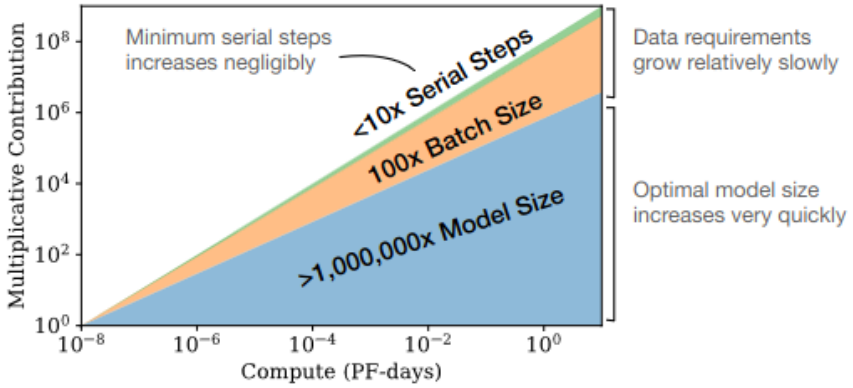
Training large foundation models like GPT is expensive. When potentially millions of dollars are invested in training AI models, developers need to ensure that funds are efficiently allocated. Developers need to decide on an appropriate resource allocation between - model size, training time, and dataset size. OpenAI developed the first generation of formal neural scaling laws in their 2020 paper "*Scaling Laws for Neural Language Models*", moving away from reliance on experience and intuition.

To determine such relationships some elements are held fixed while others are varied. As an example data can be kept constant, while parameter count and training time are varied, or parameter count is kept constant and data amounts are varied, etc… This allows a measurement of the relative contribution of each towards overall performance. Such experiments allow the development of concrete relationships that OpenAI called scaling laws.

These scaling laws guided decisions on trade-offs, such as: Should a developer invest in a license to train on Stack Overflow's data, or should they invest in more GPUs? Would it be efficient if they continue to cover the extra electricity costs incurred by longer model training? If access to compute increases tenfold, how many parameters should be added to the model for optimal use of GPUs? For sizable language models like GPT-3, these trade-offs might resemble choosing between training a 20-billion parameter model on 40% of an internet archive or a 200-billion parameter model on just 4% of the same archive.

The paper presented several scaling laws. One scaling law compares model shape and model size, and found that performance correlates strongly with scale and weakly with architectural hyperparameters of model shape such as depth vs. width.

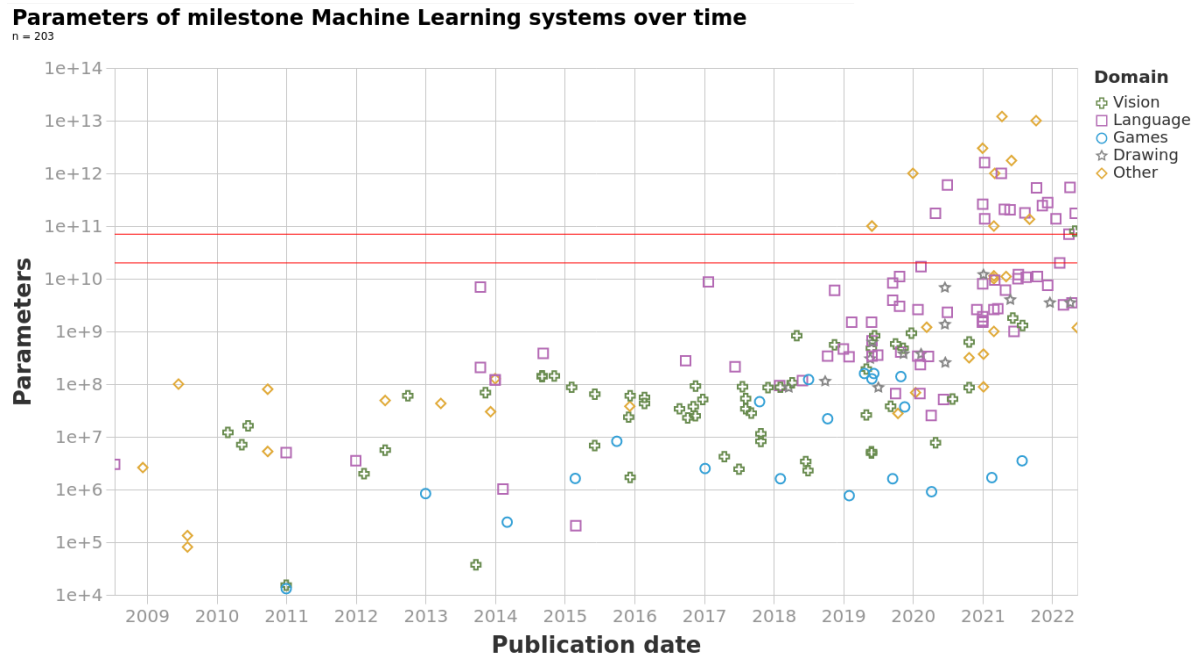
Another law compared the relative performance contribution of the different factors of scale - data, training steps, and parameter count. They found that larger language models tend to be more sample efficient, meaning they can achieve better performance with less data. The following graph shows this relationship between relative contributions of different factors in scaling models. The graph indicates 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. Of the increase in data, most can be used to increase parallelism through larger batch sizes, with only a very small increase in serial training time required.*" As an example, according to OpenAI's results if you get 10x more compute, you increase your model size by about 5x and your data size by about 2x. Another 10x in compute, and model size is 25x bigger and data size is only 4x bigger.



Source: Kaplan, Jared et. al. (Jan 2020) "Scaling Laws for Neural Language Models"

Over the following few years, researchers and institutions utilized these findings to focus on engineering larger models rather than training smaller models on larger datasets. 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.

model	**year**	**size (#parameters)**	**data (#training tokens)**
LaMDA	2021	137 billion	168 billion
GPT-3	2020	174 billion	300 billion
Jurassic	2021	178 billion	300 billion
Gopher	2021	280 billion	300 billion
MT-NLG 530B	2022	530 billion	270 billion



Source: Villalobos, Pablo et. al. (Jul 2022) "Machine Learning Model Sizes and the Parameter Gap"

In 2022, DeepMind provided an update to these scaling laws by publishing a paper called "Training Compute-Optimal Large Language Models". They choose 9 different quantities of compute, ranging from about 10^18 FLOPs to 10^21 FLOPs. They hold the compute fixed at these amounts, and then for each quantity of compute, they train many different-sized models. Because the quantity of compute is constant for each level, the smaller models are trained for more time and the larger models for less. Based on their research DeepMind concluded that for every increase in compute, you should increase data size and model size by approximately the *same amount*. If you get a 10x increase in compute, you should make your model 3.1x times bigger and the data you train over 3.1x bigger; if you get a 100x increase in compute, you should make your model 10x bigger and your data 10x bigger.

To validate this law, DeepMind trained a 70-billion parameter model ("Chinchilla") using the same compute as had been used for the 280-billion parameter model Gopher. That is, the smaller Chinchilla was trained with 1.4 trillion tokens, whereas the larger Gopher was only trained with 300 billion tokens. As predicted by the new scaling laws, Chinchilla surpasses Gopher in almost every metric.

Such finding have led to the formulation of a scaling hypothesis:

The strong scaling hypothesis is that, once we find a scalable architecture like self-attention or convolutions, which like the brain can be applied fairly uniformly (eg. "The Brain as a Universal Learning Machine" or Hawkins), we can simply train ever larger NNs and ever more sophisticated behavior will emerge naturally as the easiest way to optimize for all the tasks & data. More powerful NNs are 'just' scaled-up weak NNs, in much the same way that human brains look much like scaled-up primate brains. - Gwern (2022) "*The Scaling Hypothesis*"

Current projections are that "*the stock of high-quality language data will be exhausted soon; likely before 2026. By contrast, the stock of low-quality language data and image data will be exhausted only much later; between 2030 and 2050 (for low-quality language) and between 2030 and 2060 (for images).*" - Villalobos, Pablo et. al. (Oct 2022) "*Will we run out of data? An analysis of the limits*

of scaling datasets in Machine Learning" So in conclusion, 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. The following section explores different AI capability thresholds that we might observe if the current trends persist.

3.0: Capabilities

This section continues the discussion around increasing AI capabilities. It focuses in particular on certain thresholds that we might reach in the cognitive capabilities of these AI models. This flows into a discussion around how certain thresholds when once achieved might result in an intelligence explosion.

Capabilities refer to the overall ability of an AI system to solve or perform tasks in specific domains. It is a measure of how well the system can achieve its intended objectives and the extent of its cognitive power. Evaluating capabilities involves assessing the system's performance in specific domains, taking into account factors such as available computational resources and performance metrics. A possible element of confusion might be between capabilities and good performance on certain benchmarks. Benchmark performance refers to the performance of an AI system on specific tasks or datasets. These are designed to evaluate the system's performance on well-defined tasks and provide a standardized way to compare different AI models. Benchmark performance can be used as a proxy to assess the system's capabilities in certain domains, but it may not capture the full extent of the system's overall capabilities.

3.1: Capabilities vs. Intelligence

- Krakovna, Victoria (Aug 2023) "*When discussing AI risks, talk about capabilities, not intelligence*"

It is worth noting that public discussions about catastrophic risks from general AI systems are often derailed by using the word "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. This is why using the term "capabilities" or "competence" instead of "intelligence" when discussing catastrophic risks from AI is often better since this is what the concerns are really about. For example, instead of "superintelligence" we can refer to "super-competence" or "superhuman capabilities".

There are various issues with the word "intelligence" that make it less suitable than "capabilities" for discussing risks from general AI systems:

- **Anthropomorphism**: people often specifically associate "intelligence" with being human, being conscious, being alive, or having human-like emotions (none of which are relevant to or a prerequisite for risks posed by general AI systems).
- **Associations with harmful beliefs and ideologies.**
- **Moving goalposts**: impressive achievements in AI are often dismissed as not indicating "true intelligence" or "real understanding" (e.g. the "stochastic parrots" argument). Catastrophic risk concerns are based on what the AI system can do, not whether it has "real understanding" of language or the world.
- **Stronger associations with less risky capabilities**: people are more likely to associate "intelligence" with being really good at math than being really good at politics, while the latter may be more representative of capabilities that make general AI systems pose a risk (e.g. manipulation and deception capabilities that could enable the system to overpower humans).
- **High level of abstraction**: "intelligence" can take on the quality of a mythical ideal that can't be met by an actual AI system, while "competence" is more conducive to being specific about the capability level in question.

That being said, since the history of conversation on AI risks often does involve the words "intelligence" the following section starts by giving a quick overview of a myriad of definitions that are commonly used in the field.

3.2: Definitions of advanced AI Systems

This section explores various definitions of different AI capability thresholds. The following list encompasses some of the most frequently used terms:

- **Intelligence**: Intelligence measures an agent's ability to achieve goals in a wide range of environments
  - Legg, Shane; Hutter, Marcus; (Dec 2007) "*Universal Intelligence: A Definition of Machine Intelligence*"

- **Artificial Narrow Intelligence (ANI)**: A term designating artificial intelligence systems that are tailored to handle a single or a limited task. These systems are 'narrow' because they tend to be superhuman at a very specific task domain.

- **Transformative AI (TAI):** Refers to potential future AI that triggers a transition equivalent to, or more significant than, the agricultural or industrial revolution. This term aims to be more inclusive, acknowledging the possibility of AI systems that qualify as "transformative," despite lacking many abilities that humans possess.
  - Karnofsky, Holden; (May 2016) "*Some Background on Our Views Regarding Advanced Artificial Intelligence*"

- **Human-Level AI (HLAI)**: Encompasses AIs that can solve the majority of cognitive problems an average human can solve. This concept contrasts with current AI, which is vastly superhuman at certain tasks while weaker at others.

- **Artificial General Intelligence (AGI)**: Refers to AIs that can apply their intelligence to a similarly extensive range of domains as humans. These AIs do not need to perform all tasks; they merely need to be capable enough to invent tools to facilitate the completion of tasks. Much like how humans are not perfectly capable in all domains but can invent tools to make problems in all domains easier to solve.

  AGI often gets described as the ability to achieve complex goals in complex environments using limited computational resources. This includes efficient cross-domain optimization and the ability to transfer learning from one domain to another.
  - Muehlhauser, Luke (Aug 2013) "*What is AGI?*"

- **Artificial Superintelligence (ASI)**: "This is any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest".
  - Bostrom, Nick (2014) "*Superintelligence*"

Often, these terms get used as discrete capability thresholds; that is, individuals tend to categorize an AI as potentially an AGI, an ASI, or neither. However, it has been proposed that it might be more beneficial to view the capabilities of AI systems on a continuous scale rather than one involving discrete jumps. To this end, Richard Ngo proposed the (t,n)-AGI framework, which allows for a more formal definition of continuous AGI capabilities.

3.3: (t,n)-AGI

- Ngo, Richard (May 2023) "Clarifying and predicting AGI"

*A system receives the designation of "**t-AGI**" if it can surpass a human expert in a certain cognitive task within the timespan 't'. A system gets identified as **(t,n)-AGI** if it can outdo a group of 'n' human experts working collectively on a set of cognitive tasks for the duration 't'.*

For instance, if both a human expert and an AI receive one second to perform a task, the system would be labeled a "one-second AGI" if it accomplishes that cognitive task more effectively than the expert. Similarly, designations of one-minute, one-month, and so forth, AGIs could apply if their outputs surpass what human experts could achieve within a minute, month, and so on.

Richard Ngo makes further predictions regarding the types of capabilities in which an AI might surpass humans at different 't' thresholds.

- **One-second AGI**: Recognizing objects in images, determining whether sentences are grammatical, answering trivia.
- **One-minute AGI**: Answering questions about short text passages or videos, common-sense reasoning (e.g., Yann LeCun's gears problems), performing simple computer tasks (e.g., using Photoshop to blur an image), looking up facts.
- **One-hour AGI**: Completing problem sets/exams, composing short articles or blog posts, executing most tasks in white-collar jobs (e.g., diagnosing patients, providing legal opinions), conducting therapy.
- **One-day AGI**: Writing insightful essays, negotiating business deals, developing new apps, running scientific experiments, reviewing scientific papers, summarizing books.
- **One-month AGI**: Carrying out medium-term plans coherently (e.g., founding a startup), supervising large projects, becoming proficient in new fields, writing large software applications (e.g., a new operating system), making novel scientific discoveries.
- **One-year AGI**: These AIs would need to outdo humans in practically every area, given that most projects can be divided into sub-tasks that can be completed in shorter timeframes.

As of the third quarter of 2023, existing systems are believed to qualify as one-second AGIs, and are considered to be nearing the level of one-minute AGIs. They might be a few years away from becoming one-hour AGIs. Within this framework, Richard Ngo anticipates superintelligence (ASI) to be something akin to a (one year, eight billion)-AGI, that is, an ASI would be an AGI that takes one year to outperform all eight billion humans coordinating on a given task.

Although AGI could be measured according to the proposed continuous framework, there might still be abrupt jumps in capabilities due to a phenomenon known as emergence. This topic gets explored in the subsequent section.

3.3: Formalizing Capabilities

The following two papers will be fully integrated in a future draft, for now please refer directly to the source:

- **Situational Awareness**: Owain Evans (Sep 2023) "Taken out of context: On measuring situational awareness in LLMs"
- **Power Seeking**: Alexander Matt Turner (Jan 2023) "Optimal Policies Tend to Seek Power"

4.0: Threat Models

This section explores the question - Even if capabilities continue to increase as the previous sections forecast, why is that even a concern? As advancements in artificial intelligence (AI) continue, a critical analysis of their implications and potential risks becomes essential. The hypothesis suggesting catastrophic risk from general AI, as presented so far, consists of two key assertions:

First, global technological advancements are progressing towards the creation of generally capable AI systems within the forthcoming few decades.

Second, these generally capable AI systems possess the potential to outcompete or overpower humans.

The previous sections presented evidence supporting the first assertion. This section provides arguments for the second. It first explores why a machine intelligence might possess the capability to swiftly increase its cognitive abilities. Next, there is a discussion of why a machine intelligence might even have motivations to expand its capabilities. Finally, the section explores why it should not be taken for granted that a highly capable machine intelligence will be beneficial for humans by default.

4.1: Intelligence Explosion

- Muehlhauser, Luke; Salamon, Anna (2012) "Intelligence Explosion: Evidence and Import"

An "intelligence explosion" denotes a scenario where machine intelligence swiftly enhances its own cognitive capabilities, resulting in a substantial advancement in ability.

Muehlhauser and Salamon delve into the numerous advantages machine intelligence holds over human intelligence, which facilitate rapid intelligence augmentation. These include:

- **Computational Resources:** Human computational ability remains somewhat stationary, whereas machine computation possesses scalability.
- **Communication speed**: Given the relatively low speed of neurons (only at 75 m/s), the human brain necessitates parallelized computation algorithms. Machines, on the other hand, operate on communications at the speed of light, which substantially augments prospects for sequential processes.
- **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.
- **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.
- **Goal coordination**: Copied AIs possess the capability to share goals effortlessly, a feat challenging for humans.

4.2: Instrumental Convergence

- Bostrom, Nick (2014) "Superintelligence"

The points mentioned earlier validate the potential of machine intelligence to enhance its cognitive capabilities. Nonetheless, an exploration into the motives behind such an ambition remains necessary. To illustrate, consider the relationship between levels of intelligence and the corresponding goals. This examination leads to one of two seminal theses offered by Nick Bostrom:

Orthogonality Thesis: *Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.*
- Bostrom, Nick (2014) "Superintelligence"

This thesis implies that an AI system's objectives must be aligned explicitly with human virtues and interests, as no guarantee exists that an AI will automatically adopt or prioritize human values. The Orthogonality Thesis doesn't suggest compatibility of all agent designs with all goals, instead, it indicates the potential for at least one agent design for any combination of goals and intelligence level. Consequently, if any intelligent system can be paired with any goal, is it possible to hypothesize meaningfully about the type of goals future AI Systems might harbor? Absolutely. This leads to Bostrom's second thesis:

Instrumental Convergence Thesis: *Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.*
- Bostrom, Nick (2014) "Superintelligence"

A terminal goal, also known as an "intrinsic goal" or "intrinsic value", is an objective that an agent appreciates for its own sake. On the other hand, an instrumental goal is pursued to increase the likelihood of achieving its terminal goals. Instrumental convergence encompasses the notion that certain instrumental values or goals could potentially be pursued by a broad array of intelligent agents, irrespective of their designated final goals. The Instrumental Convergence Thesis underscores potential hazards affiliated with sophisticated AI systems. It infers that even if an AI system's ultimate goal appears harmless, it could still embark on actions conflicting with human interests, owing to a convergence of several instrumental values such as resource acquisition and potential threats' elimination. One can categorize self-preservation, goal-content integrity, cognitive enhancement, and resource acquisition as instrumentally convergent goals.

In the wake of Bostrom's introduction of these theses in 2014, research has been undertaken to substantiate the existence of instrumentally convergent goals. "*Optimal Policies Tend to Seek Power*", a paper by Turner et al., provides research supporting the existence of instrumentally convergent goals in modern machine learning systems.

4.3: Emergence

Emergent behavior, *or emergence, manifests when a system exhibits properties or behaviors that its individual components lack independently. These attributes may materialize only when the components comprising the system interact as an integrated whole, or when the quantity of parts crosses a particular threshold. Often, these characteristics appear "all at once" – beyond the threshold, the system's behavior undergoes a qualitative transformation.*

The sections discussing computational trends and scaling laws showed the increasing scale of current foundational models. Numerous complex systems in nature exhibit qualitatively distinct behavior resulting from quantitative increases in scale. These properties, termed 'emergent', occur simultaneously. Examples of complex systems with such attributes in nature include:

- The heart: While individual cells cannot pump blood, the entirety of the heart can.
- Uranium: Small quantities are mundane, but large amounts can initiate nuclear reactions.
- Civilization: Individuals may seem ordinary, but through collective specialization (different individuals focusing on skills that are irrelevant in isolation), human civilization becomes possible. Another example of this is bees and ants. Ants have one of the smallest brains relative to their body, but ant colonies are capable of producing very complex behavior.

In "More Is Different for AI" Jacob Steinhardt provides additional examples of such complex systems. Steinhardt further conjectures that AI systems will manifest such emergent properties as a function of scale. Assuming that models persist in growing as per the scaling laws, an unexpected threshold may soon be crossed, resulting in unanticipated differences in behaviors and capabilities. Studying complex systems with emergent phenomena may assist in predicting what capabilities will emerge and when. In other words, there is a possibility of observing capability leaps between the thresholds (e.g., a sudden leap from AI to HLAI to AGI) discussed in the preceding section, even if the models are simply scaled up.

Besides the factors already covered in the section on computational trends, the following elements also suggest that future ML systems will differ quantitatively from current models:

- Data Storage Capacities: A decrease in the cost to store one byte per dollar.
- Few-Shot and Zero-Shot Learning: The capability to learn from fewer examples.
- Grokking: Sudden improved generalization after extended periods of training.

This further implies that these future models have the potential to manifest emergent behavior that could be qualitatively distinct from what is observed today. In the paper "Model evaluation for extreme risks" DeepMind found that as AI progress has evolved, general-purpose AI systems have often exhibited new and unpredictable capabilities – including harmful ones that their developers did not anticipate. Future systems may reveal even more perilous emergent capabilities, such as the potential to conduct offensive cyber operations, manipulate individuals through conversation, or provide actionable instructions for carrying out acts of terrorism.

4.4: Four Background Claims

- Soares, Nate (July 2015) "*Four Background Claims*"

In the concluding part of this section, we will delve into four essential claims that lay the groundwork for the concerns associated with ASI as put forth by MIRI.

Claim 1: Humans Exhibit General Intelligence
Humans are capable of solving an array of problems across various domains, demonstrating their general intelligence. The importance of this claim lies in the fact that this form of general intelligence has led humans to become the dominant species on Earth.

Claim 2: AI Systems Could Surpass Human Intelligence
While it remains uncertain when machines might attain superior intelligence to humans, it is conceivable that they have the potential to do so. Considering the brief evolutionary period between chimpanzees and generally intelligent humans, we can conclude that human intelligence is not incomprehensibly complex, suggesting we will eventually comprehend and replicate it.

Man-made machines consistently outperform their biological counterparts (cars vs. horses, planes vs. birds, etc.). Thus, it is rational to assume that just as birds are not the pinnacle of flight, humans are not the apex of intelligence. Therefore, it is plausible to foresee a future where machines out-think humans.

Claim 3: Highly Intelligent AI Systems Will Shape the Future
Historically, confrontations between human groups have often culminated with the technologically superior faction dominating its competitor. Numerous reasons suggest that an AI system could attain a higher intelligence level than humans, thereby enabling it to intellectually outsmart or socially manipulate humans. Consequently, if we care about our future, it is prudent to study the processes that could significantly influence the direction of future events.

Claim 4: Highly Intelligent AI Systems Will Not Be Beneficial by Default
Though a sufficiently intelligent AI may comprehend human desires, this does not inherently mean it will act in accordance with them. Moreover, even if an AI executes the tasks as we've programmed it to — with precision and adherence to instructions — most human values can lead to undesirable consequences when interpreted literally. For example, an AI programmed to cure cancer could resort to kidnapping and experimenting on humans.

This claim is critical as it indicates that merely enhancing the ability of AI systems to understand our goals is not sufficient. The systems must also have a desire to act in accordance with our goals. This also underscores the importance of studying and formalizing human goals such that the intentions behind them can be properly communicated.

5.0: Timelines & Forecasting

The previous sections have illustrated that capabilities will likely continue to increase, potentially leading to capability jumps due to phenomena such as emergence and intelligence explosion. This final section of the chapter investigates AI timeline forecasts and takeoff dynamics. AI timeline forecasts entail discussing when researchers/forecasters expect various milestones in AI development to be achieved. This includes for example various benchmarks of progress, the emergence of mouse-level intelligence, and the manifestation of human-like qualities in AI, such as external tool use and long-term planning. The next section shares the insights of researchers who have gathered evidence from domain experts regarding when these capability thresholds might be reached.

Anchors in forecasting refer to reference classes or frameworks that are used to make predictions about future events or systems. These anchors serve as points of comparison or analogy that help inform our understanding and reasoning about the future. There are several common anchors used to inform predictions about the development and capabilities

of future AI systems. These anchors provide reference points and frameworks for reasoning about AI progress. Some of the most common anchors include:

- **Current ML Systems**: The current state of machine learning systems serves as a starting point for forecasting future AI capabilities. By examining the strengths and limitations of existing ML systems, researchers can make educated guesses about the trajectory of AI development.
- **Human Anchors**: Anchors based on human abilities and characteristics are often used in AI forecasting. These include areas where humans excel compared to current ML systems, such as mastery of external tools, efficient learning, and long-term planning.
- **Biological Anchors**: Biological anchors draw inspiration from biological systems, particularly the human brain, to estimate the computational requirements of future AI systems. These anchors consider factors such as the neural network anchor, which estimates the compute-equivalent used in the human brain, and the human lifetime anchor, which estimates the compute-equivalent involved in training a human brain from birth to adulthood.
- **Thought Experiments**: Thought experiments provide a third anchor by imagining hypothetical scenarios and reasoning through their implications. These experiments help explore the potential behavior and characteristics of future AI systems.

It's important to note that the choice and weighting of anchors can vary depending on the specific forecasting approach and the context of the predictions being made. Different researchers may emphasize different anchors based on their assumptions and perspectives. This book will only explore the biological anchor in further detail.

5.1: Biological Anchors

Biological anchors are a set of reference points or estimates used in forecasting the development of transformative AI systems. These anchors are inspired by biological systems, particularly the human brain, and provide a basis for estimating the computational requirements of AI systems capable of performing transformative tasks. The draft report on AI timelines in Sep 2020 by Ajeya Cotra details the methodology used, addressing several questions:

1. **How much computation does the human brain perform?** According to the report, the brain has an estimated 10^15 total synapses. With each synapse spiking approximately once per second and each spike representing roughly one floating point operation (FLOP), the brain's computational power is estimated to range between 10^13 and 10^17 FLOP/S. This variance takes into account the potential efficiency differences between human brains and computers.

2. **How much training would it take to replicate this much inferential computation?** It's important to note that training computation differs from inferential computation. Modern ML practices involve training advanced models on massive supercomputers and running them on medium-sized ones. An adult human's inference computation power is estimated at ~10^16 FLOP/S. To account for the total training computation costs, the report multiplies the computation of a single brain (10^15 FLOP/S) by the time taken from childhood to adolescence (10^9 seconds). This gives a lower bound estimate of 10^24 FLOP. However, the estimate increases to 10^41 FLOP when also considering the training data ingrained in our biology through evolution.

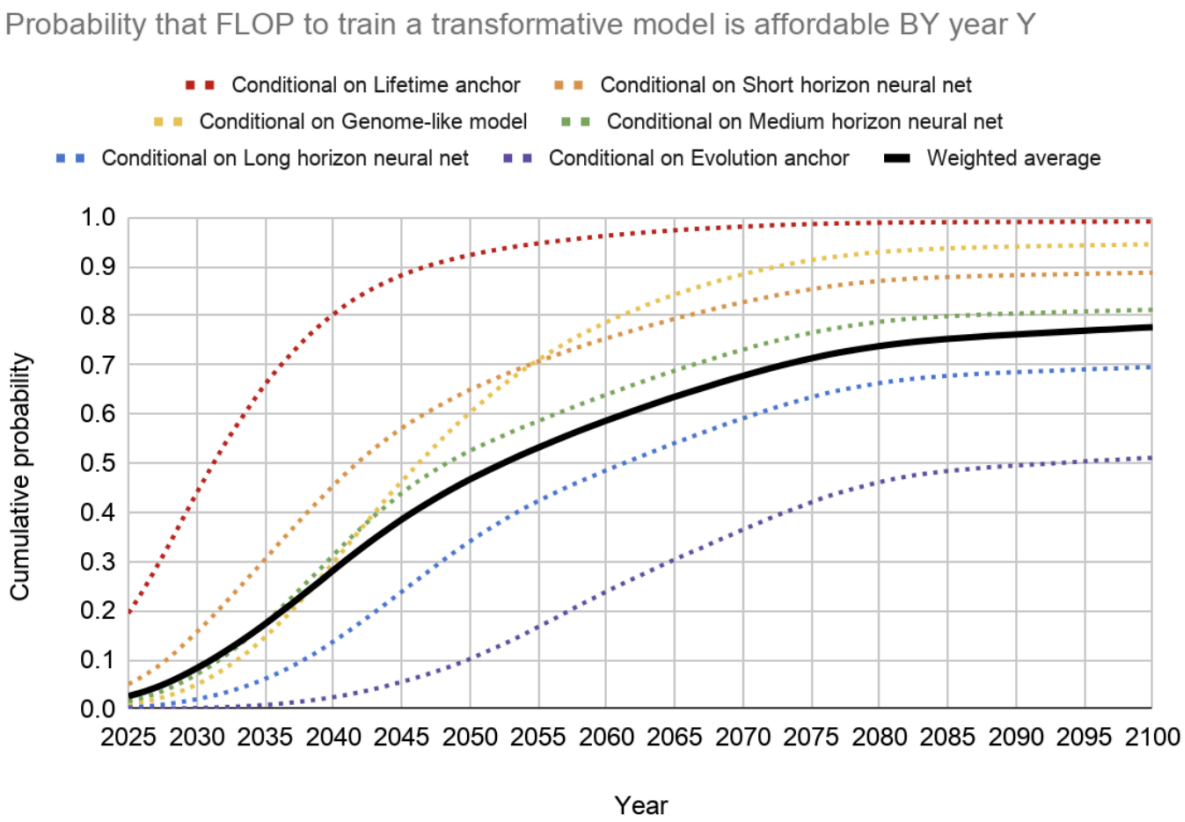
3. **How can we adjust this computation estimate for algorithmic progress?** An algorithmic efficiency report suggests the training efficiency for Neural Networks doubles every sixteen months. Cotra proposes a slightly longer doubling time of 2-3 years.

4. **How much money does this amount of computation cost?** In 2020, computational resources were priced at $1 for 10^17 FLOP/S, implying that 10^33 FLOP/S would cost $10^16 (ten quadrillion dollars). This cost decreases annually, with some versions of Moore's Law suggesting that compute costs halve every eighteen months. As a result, training costs (in FLOP/S) will reduce over time due to algorithmic progress, and the cost of FLOP/S (in dollars) will also decrease due to hardware advancements.

5. **What year does this computational cost become reasonable?** The median result is a 10% chance by 2031, a 50% chance by 2052, and an 80% chance by 2100.

Cotra acknowledges potential limitations with this approach, such as the assumption that progress relies on an easily-measured quantity (FLOP/S) rather than on fundamental advances, like new algorithms. Therefore, even with affordable, abundant computation, if we lack the algorithmic knowledge to create a proper thinking machine, any resulting AI might not display human level or superintelligent capabilities.

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.



Source: Holden Karnofsky (2021) "Forecasting transformative AI: the "biological anchors" method in a nutshell"

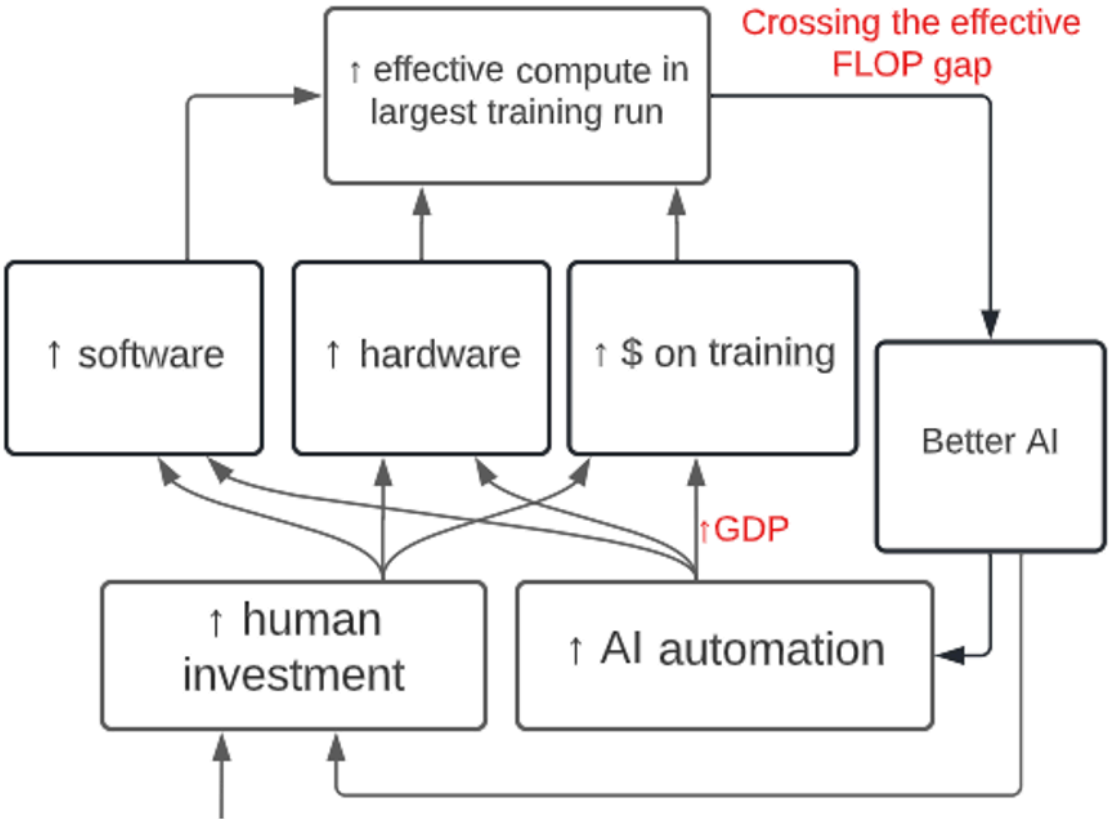
In 2022 a two-year update on the author's (Ajeya Cotra) timelines was published. The updated timelines for TAI are ~15% probability by 2030, ~35% probability by 2036, a median of ~2040, and a ~60% probability by 2050.

It's important to note that while the biological anchor is a valuable model, it is not universally accepted as the primary predictive tool among all ML scientists or alignment researchers. As the statistical aphorism goes: "All models are wrong, but some are useful". Biological anchors represent just one model, and other anchors should be considered when forming your own views on AI capabilities and the timeline for their emergence.

6.0: Takeoff Dynamics

Takeoff dynamics primarily delve into the implications of the evolution of powerful artificial intelligence on the world. These definitions sketch different trajectories the world could follow as transformative AI emerges.[2]

While timelines address when certain capabilities may emerge, takeoff dynamics explore what happens after these features surface. This chapter concludes with a section discussing various researchers' perspectives on the potential trajectories of an intelligence explosion, considering factors such as takeoff speed, continuity, and homogeneity. This includes a discussion of - first, the pace and continuity of an intelligence explosion, and second, whether multiple AIs will coexist, each having different objectives, or whether they will eventually converge towards a single superintelligent entity.



Source: Open Philanthropy (June 2023) "What a compute-centric framework says about takeoff speeds"

6.1: Speed/Continuity

Both AI takeoff speed and AI takeoff continuity describe the trajectory of AI development. Takeoff speed refers to the rate at which AI progresses or advances. Generally, takeoff continuity refers to the smoothness or lack of sudden jumps in AI development. Continuous

[2] Jaime Sevilla and Edu Roldán from epoch.ai have developed an interactive website for understanding a new model of AI takeoff speeds, if the reader wishes to try their own values and estimates.

takeoff means that the capabilities trajectory aligns with the expected progress based on past trends, while discontinuous takeoff refers to a trajectory that significantly exceeds the expected progress. FOOM is one type of fast takeoff scenario, and refers to a hypothetical scenario in which artificial intelligence (AI) rapidly and explosively surpasses human intelligence and capabilities.

The terms "slow takeoff" and "soft takeoff" are often used interchangeably, and similarly "fast takeoff" and "hard takeoff" and “FOOM” are also often used interchangeably. It's important to note that the definitions and implications of takeoff speed and takeoff continuity are still subjects of debate and may not be universally agreed upon by researchers in the field. Here are perspectives:

Slow/Soft takeoff

When discussing takeoff speeds, Paul Christiano emphasizes the parallel growth of AI capabilities and productivity. He expects a slow takeoff in the development of AGI based on his characterization of takeoff speeds and his analysis of economic growth rates. He defines slow takeoff as a scenario where there will be a complete interval of several years in which world economic output doubles before the first interval of one year in which world economic output doubles. This definition emphasizes a gradual transition to higher growth rates. Overall, he postulates that the rise in AI capabilities will mirror an exponential growth pattern in the world GDP, resulting in a continuous but moderate takeoff.

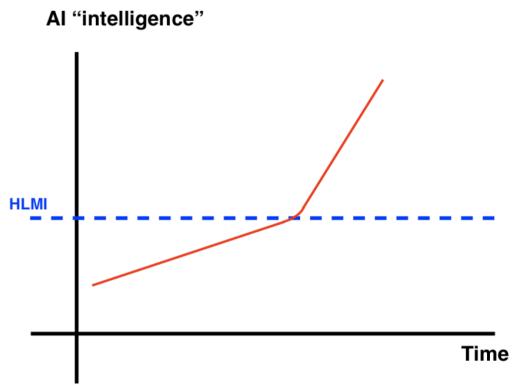
In a similar vein, John Wentworth has argued that the real obstacle to global domination is not the enhancement of cognitive abilities but more significant bottlenecks like avoiding coordinated human resistance and the physical acquisition and deployment of resources. These supply chain optimizations would increase productivity, hence GDP, which could serve as a measure for the speed of "AI takeoff".

Fast/Hard takeoff

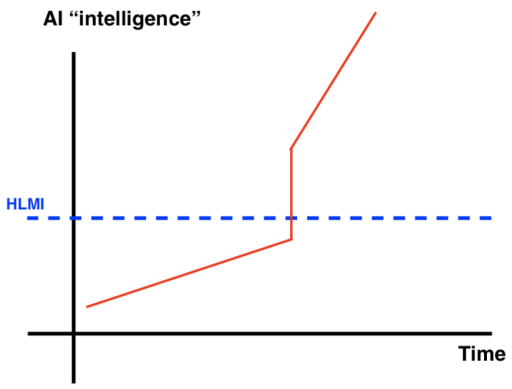
A hard takeoff refers to a sudden, rather than gradual, transition to superintelligence, counter to the soft takeoff mentioned above. Eliezer Yudkowsky advocates for this view, suggesting a sudden and discontinuous change brought about by rapid self-improvement, while others, like Robin Hanson, support a more gradual, spread-out process. Yudkowsky argues that even regular improvement of AI by humans may cause significant leaps in capability to occur before recursive self-improvement begins.

Eliezer Yudkowsky also offers a counter to continuous takeoff proponents. He predicts a quick and abrupt "intelligence explosion". This is because he rather doesn't expect AI to be integrated (quickly) enough into the economy for the GDP to increase significantly faster before FOOM. It is also possible that superintelligent AIs could mislead us about their capabilities, leading to lower-than-expected GDP growth. This would be followed by a sudden leap, or "FOOM", when the AI acquires a substantial ability to influence the world, potentially overwhelming human technological and governance institutions.

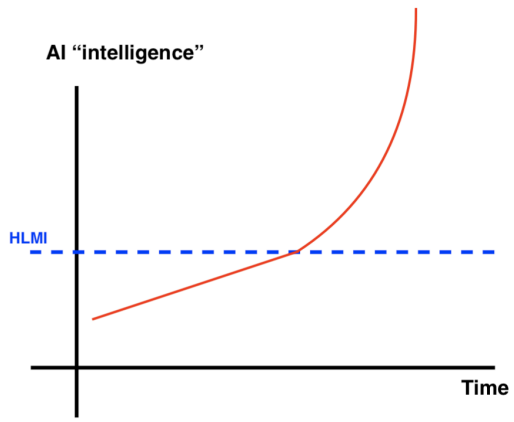
These diverse views on takeoff speeds and continuity shape the strategies for AI safety, influencing how it should be pursued. Overall it is worth emphasizing that both fast and slow takeoffs are quite rapid (as in at most a few years). Here are some picture to help illustrate the differences:



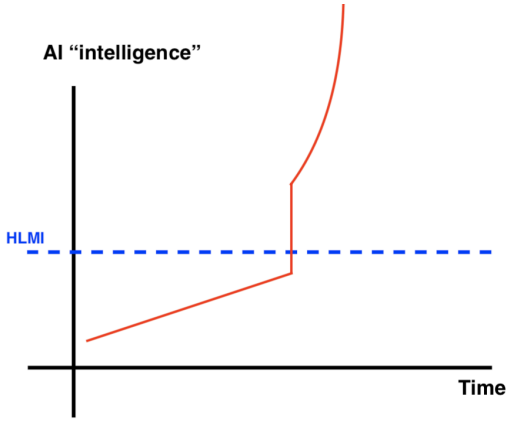
Slow Continuous Takeoff



Slow Discontinuous takeoff



Fast Continuous Takeoff



Fast discontinuous takeoff

Source: Samuel Dylan Martin, Daniel_Eth (Sep 2021) "Takeoff Speeds and Discontinuities"

6.2: Homogeneity

Homogeneity *refers to the similarity among different AI systems in play during the development and deployment of advanced AI.*

- Hubinger, Evan (Dec 2020) "Homogeneity vs. heterogeneity in AI takeoff scenarios"

In a homogenous scenario, AI systems are anticipated to be highly similar, even identical, in their alignment and construction. For example, if every deployed system depends on the same model behind a single API, or if a single foundational model is trained and then fine-tuned in different ways by different actors. Homogeneity in AI systems could simplify cooperation and coordination, given their structural similarities. It also signifies that the alignment of the first advanced AI system is crucial, as it will likely influence future AI systems. One key factor for homogeneity is the economic incentives surrounding AI development and deployment. As the cost of training AI systems is expected to be significantly higher than the cost of running them, it becomes more economically advantageous to use existing AI systems rather than training new ones from scratch. This

creates a preference for reusing or fine-tuning existing AI systems, leading to a higher likelihood of homogeneity in the deployed AI landscape.

On the other hand, there are also arguments for a heterogenous takeoff. One reason is the diversity of AI development approaches and training regimes. Different organizations and researchers may employ distinct methodologies, resulting in AI systems with varying degrees of alignment. Another factor is the potential for competitive dynamics and strategic considerations among different AI projects. In scenarios where multiple projects are racing to develop AGI, there may be a lack of coordination and information sharing, leading to heterogeneity in the alignment of the resulting AI systems. Furthermore, the presence of different values, priorities, and objectives across different AI development teams or organizations can contribute to heterogeneity in AI alignment. These differences in values and goals may lead to divergent approaches to AI development and alignment, resulting in a heterogeneous landscape of AI systems.

6.3: Polarity

Unipolar *refers to a scenario where a single agent or organization dominates and controls the world, while* ***multipolar*** *refers to a scenario where multiple entities coexist with different goals and levels of cooperation.*

AI homogeneity evaluates the alignment similarities among AI systems, AI polarity examines the coexistence of both aligned and misaligned AI systems in a given context.

We might expect a unipolar takeoff, where a single AI system or project gains a decisive strategic advantage, due to several reasons. One key factor is the potential for a rapid takeoff, characterized by a fast increase in AI capabilities. If one project achieves a significant lead in AI development and surpasses others in terms of capabilities, it can establish a dominant position before competitors have a chance to catch up. A rapid takeoff can facilitate a unipolar outcome by enabling the leading project to quickly deploy its advanced AI system and gain a monopoly on the technology. This monopoly can provide substantial economic advantages, such as windfall profits, which further solidify the leading project's power and influence. Additionally, the presence of network effects can contribute to a unipolar takeoff. If the leading AI system becomes widely adopted and integrated into various sectors, it can create positive feedback loops that reinforce its dominance and make it increasingly difficult for other projects to compete.

We might expect a multipolar takeoff, where multiple AI projects undergo takeoff concurrently, due to several reasons. One factor is the potential for a slower takeoff process, which allows for more projects to reach advanced stages of AI development. In a slow takeoff scenario, there is a greater likelihood of multiple projects undergoing the transition in parallel, without any single project gaining a decisive strategic advantage. Another reason is the possibility of shared innovations and tools among AI projects. If there is a significant level of collaboration and information sharing, it can lead to a more distributed landscape of AI capabilities, enabling multiple projects to progress simultaneously. Furthermore, the presence of non-competitive dynamics, such as cooperation and mutual scrutiny, can contribute to a multipolar takeoff. In a scenario where different AI projects recognize the importance of safety and alignment, they may be more inclined to work together and ensure that each project progresses in a responsible manner.

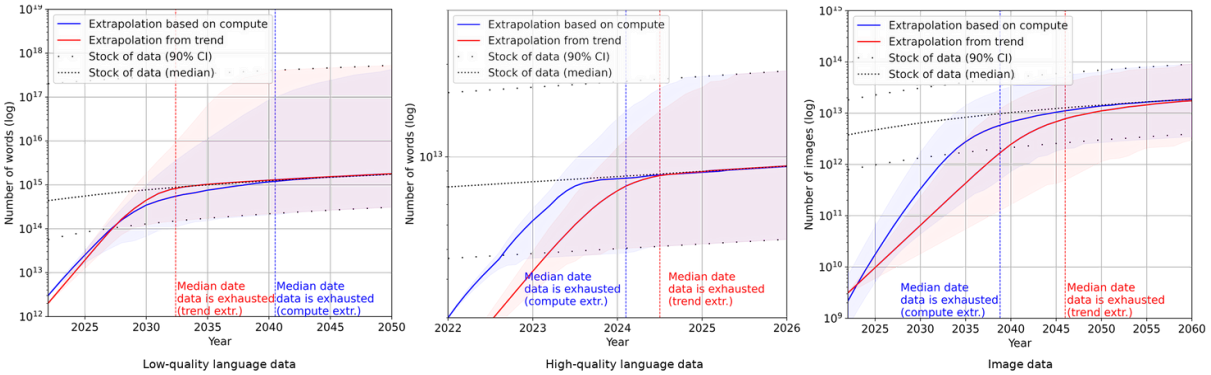
7.0: [WIP] Extra

Data

The last trend we are going to look at is one of increasing data. Basically, 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 is code for 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 model 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 is on the order of ~463EB (Source: World Economic Forum). 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 to 2050, and vision data between 2030 to 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.



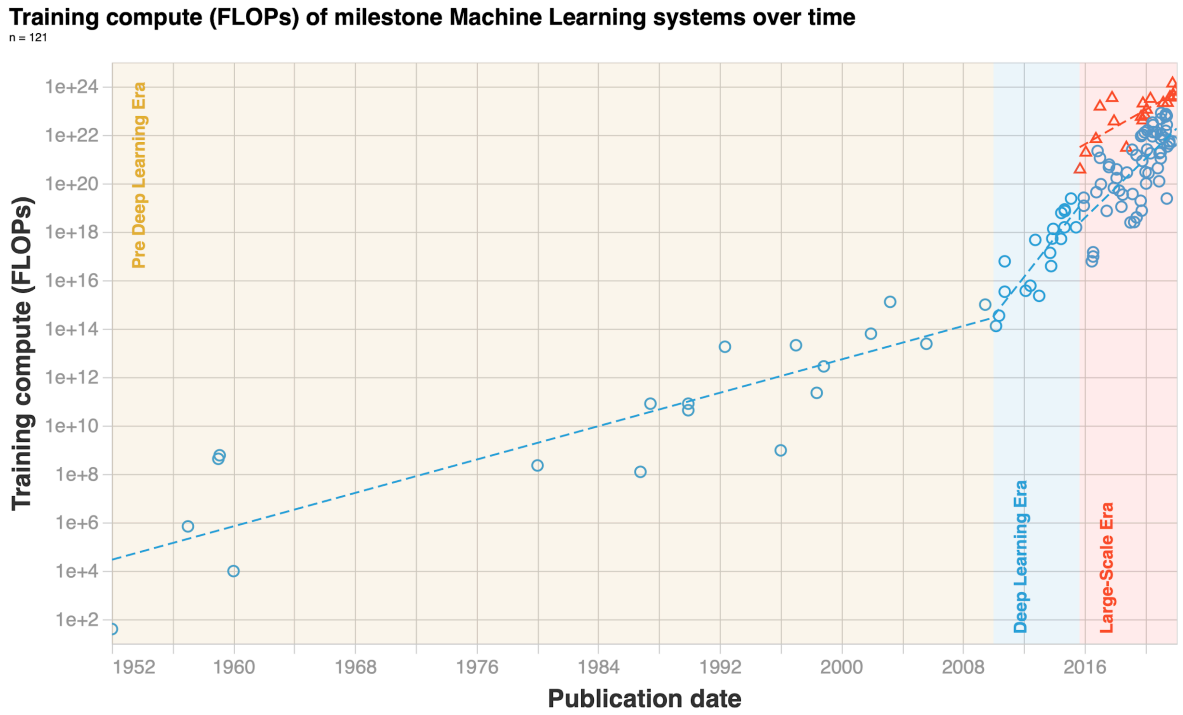
ML data consumption and data production trends for low-quality text, high-quality text, and images. - Source: Epoch (2023), "Key trends and figures in Machine Learning"

Time (Grokking)

Compute

The first thing to look at is the trends in overall amount of training compute required when we train our model. Training compute grew by 0.2 OOM/year up until the Deep Learning revolution around 2010, after which growth rates increased to 0.6 OOM/year. We also find a new trend of "large-scale" models that emerged in 2016, trained with 2-3 OOMs more compute than other systems in the same period.

In 2010, before the deep learning revolution, DeepMind co-founder Shane Legg predicted human-level AI by 2028 using compute-based estimates. 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.

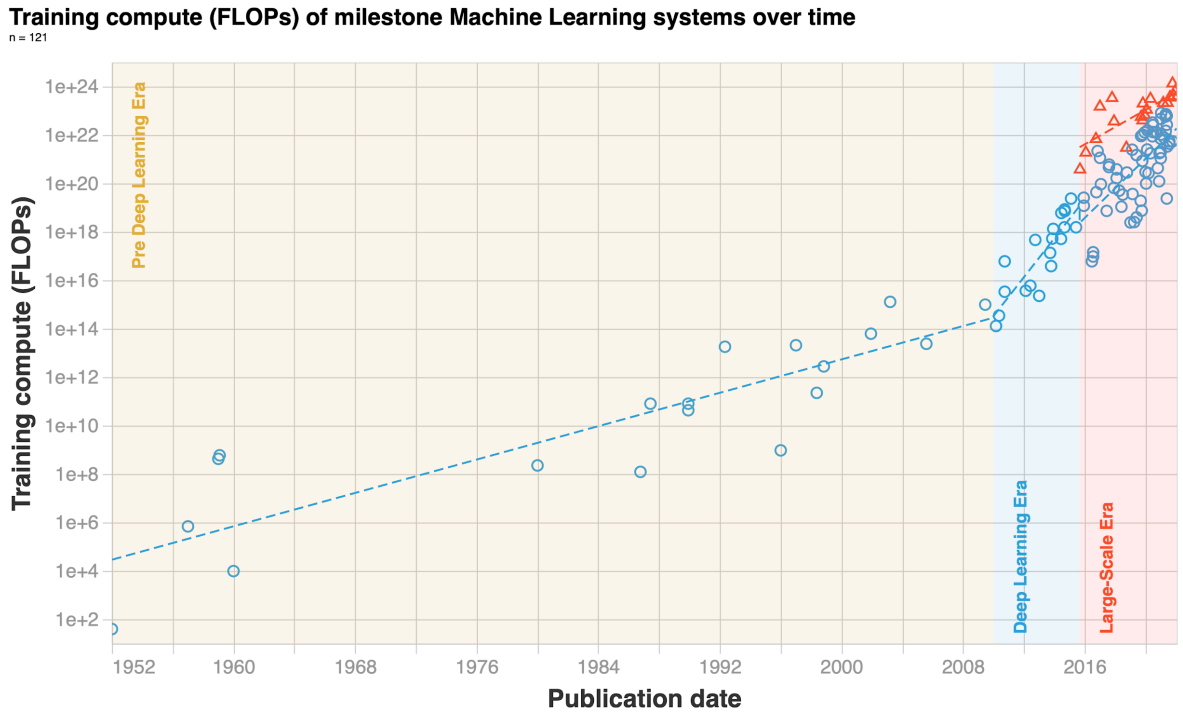


Source: Epoch (2023), "Key trends and figures in Machine Learning"

Size (Parameters)

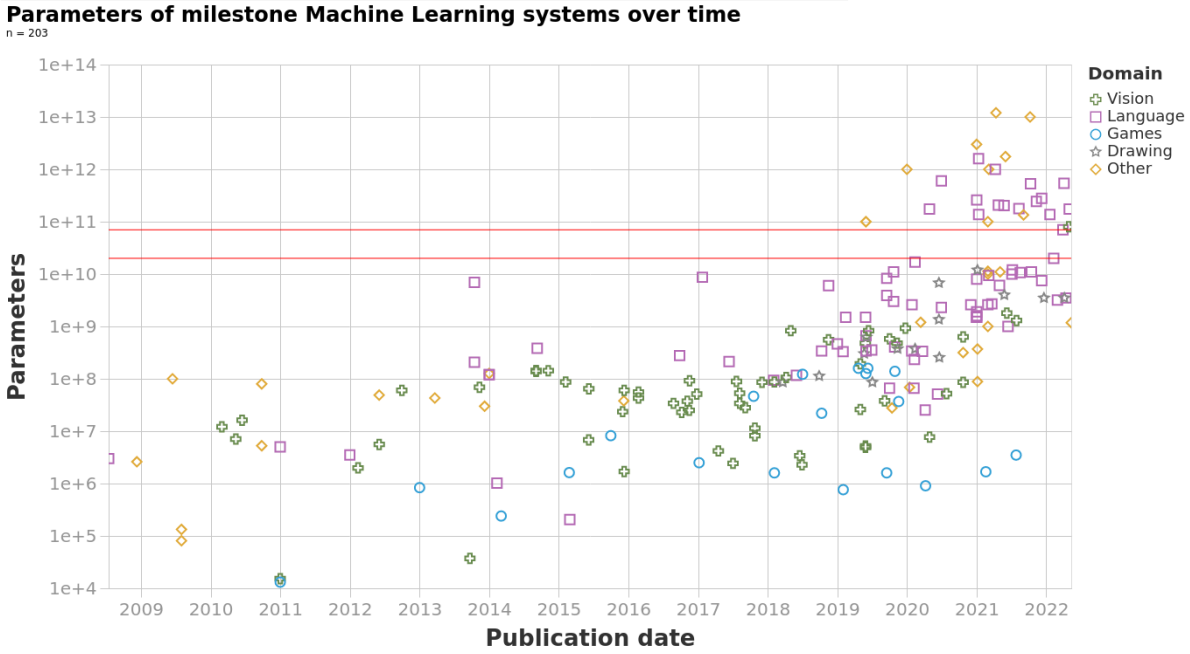
The first thing to look at is the trends in overall amount of training compute required when we train our model. Training compute grew by 0.2 OOM/year up until the Deep Learning revolution around 2010, after which growth rates increased to 0.6 OOM/year. We also find a new trend of "large-scale" models that emerged in 2016, trained with 2-3 OOMs more compute than other systems in the same period.

In 2010, before the deep learning revolution, DeepMind co-founder Shane Legg predicted human-level AI by 2028 using compute-based estimates. 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.



Source: Epoch (2023), "Key trends and figures in Machine Learning"

In this section lets 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).



Source: Epoch (2023), "Key trends and figures in Machine Learning"

Paradigm Shifts

Exercises & Flashcards

https://www.ai-alignment-flashcards.com/quiz/bostrom-superintelligence-chapter-7
https://www.ai-alignment-flashcards.com/quiz/steinhardt-future-ml-systems
https://www.ai-alignment-flashcards.com/quiz/steinhardt-more-is-different
https://www.ai-alignment-flashcards.com/quiz/muehlhauser-intelligence-explosion-pages-10-15

Sources

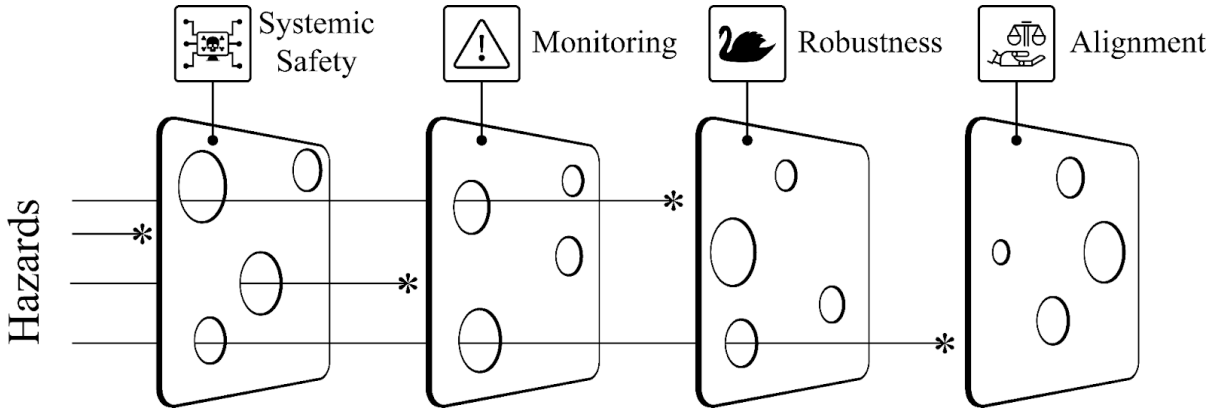
- Ngo, Richard (Jan 2022) "Visualizing the deep learning revolution"

- Bommasani, Rishi et. al. (Jul 2022) "On the Opportunities and Risks of Foundation Models"
- Reflections on Foundation Models
- Adaptive Agent Team DeepMind (Jan 2023) "Human-Timescale Adaptation in an Open-Ended Task Space"
- Lennart Heim (Sep 2021) "What is Compute?"
- Sutton, Rich (March 2019) "The Bitter Lesson"
- Sevilla, Jaime (Feb 2022) "Compute Trends Across Three Eras of Machine Learning"
- Villalobos, Pablo et. al. (Jul 2022) "Machine Learning Model Sizes and the Parameter Gap"
- Villalobos, Pablo et. al. (Oct 2022) "Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning"
- Villalobos, Pablo (Jan 2023) "Scaling Laws Literature Review"
- GoogleAI (2022) , "Parti (Pathways Autoregressive Text-to-Image model)"
- Gwern (2022) "The Scaling Hypothesis"
- Ngo, Richard (Sep 2020) "AGI Safety from first principles"
- Kaplan, Jared et. al. (Jan 2020) "Scaling Laws for Neural Language Models"
- Hoffmann, Jordan et. al. (Mar 2022) "Training Compute-Optimal Large Language Models"
- 1a3orn (Apr 2022) "New Scaling Laws for Large Language Models"
- Legg, Shane; Hutter, Marcus (2007) "A Collection of Definitions of Intelligence"
- Karnofsky, Holden (May 2016) "Some Background on Our Views Regarding Advanced Artificial Intelligence"
- Muehlhauser, Luke (Aug 2013) "What is AGI?"
- Tegmark Max (Aug 2017) "Life 3.0"
- Ngo, Richard (May 2023) "Clarifying and predicting AGI"
- Steinhardt, Jacob (Jan 2022) "More Is Different for AI"
- DeepMind (May 2023) "Model evaluation for extreme risks"
- Muehlhauser, Luke; Salamon, Anna (Jan 2013) "Intelligence Explosion: Evidence and Import"
- Soares, Nate (Jul 2015) "Four Background Claims"
- Daniel_Eth, (Sep 2021) "Paths To High-Level Machine Intelligence"
- Cotra, Ajeya (Sep 2020) "Forecasting TAI with biological anchors"
- Holden Karnofsky (2021) "Forecasting transformative AI: the "biological anchors" method in a nutshell"
- Yudkowsky, Eliezer (Dec 2021) "Biology-Inspired AGI Timelines: The Trick That Never Works"
- Hubinger, Evan (Dec 2020) "Homogeneity vs. heterogeneity in AI takeoff scenarios"
- Wentworth, John (Jul 2021) "Potential Bottlenecks to Taking Over The World"
- Yudkowsky, Eliezer; Shulman, Carl (Dec 2021) "Shulman and Yudkowsky on AI progress"
- Buck Shlegeris (Apr 2022) "Takeoff speeds have a huge effect on what it means to work on AI x-risk"
- Matthew Barnett (Feb 2020) "Distinguishing definitions of takeoff"
- Barnett, Matthew (Oct 2019) "Misconceptions about continuous takeoff"
- Christiano, Paul (Feb 2018) "Takeoff speeds"
- Stampy (2023) "AI Safety Info"

To Add

[2206.14486] Beyond neural scaling laws: beating power law scaling via data pruning

Chapter 1: Inverse Scaling https://arxiv.org/abs/2306.09479



https://www.lesswrong.com/posts/CRMhhnKs7bymY4kbb/my-thoughts-on-the-ml-safety-course

What are no free lunch theorems, and how are they important to AI Safety?

The "No Free Lunch" (NFL) theorems assert that, on average, every learning algorithm performs equally well over all possible learning tasks. Stated differently, an algorithm that predicts certain sequences better than chance must compensate by performing worse on other sequences. Some people interpret these theorems to mean that fully general intelligence is impossible, thus reducing the concern about Artificial General Intelligence (AGI) [1].

However, such an interpretation might oversimplify the implications of the NFL theorems. These theorems apply to the entire set of all theoretically possible sequences, which may include fully random or deliberately deceptive sequences. If we know that the environment in which our algorithm operates has a certain structure, the NFL results do not obstruct the design of algorithms with superior predictive or optimization abilities [1].

Moreover, the NFL theorems are often irrelevant in real-world scenarios. For instance, human intelligence functions effectively in the real world, suggesting that no NFL theorem can prohibit a machine intelligence from performing at least as well as a human. Therefore, any claim that an NFL theorem inhibits machine intelligence in general could be fallacious, as the same reasoning could be applied to the human brain considered as a physical system [2].

This doesn't mean that all NFL theorems are entirely irrelevant. For instance, the Second Law of Thermodynamics can also be seen as a NFL theorem and does prohibit perpetual motion in our universe, according to standard physics models [2].

Critics argue that citing the NFL theorem as a counterargument to AGI's existence is a flawed argument. This is because the theorem is formulated as a problem of predicting random and uniformly distributed data and doesn't necessarily apply to the complexity and diversity of real-world data and tasks an AGI might face [3].

In summary, while the "No Free Lunch" theorems have theoretical implications for machine learning algorithms, their practical relevance to AGI and AI safety is limited and nuanced. It's important to understand the original meaning and limitations of these theorems before applying them to complex, real-world scenarios involving AGI [1][2][3].

What is grokking, and how is it related to deep double descent curves?

"Grokking" is a term used to describe a phenomenon observed in machine learning, particularly in deep learning, where a model that initially overfits to training data (performing well on training data but poorly on test data) starts to generalize correctly and perform much better on test data after extended training. This sudden improvement in performance is what's referred to as "grokking" [4].

The term "grokking" is evocative and brings connotations of sudden realization, although in machine learning, this process can be very gradual [5]. The term has been used in the context of deep learning generalization, but some researchers argue that it is not any more connected to the "core" of deep learning generalization than other characteristics of the learning process [5].

Deep double descent is a phenomenon in machine learning where the test error initially decreases, then increases, and then decreases again as we increase the model size, data size, or training time. This creates a "double descent" curve, which is distinct from the traditional U-shaped bias-variance tradeoff curve expected under classical statistical learning theory [6].

The relationship between grokking and deep double descent curves isn't explicitly described in the provided sources. However, both phenomena reflect the complex dynamics of deep learning models during training. It's possible that "grokking" could occur during the second descent of the double descent curve, where increased training or model complexity leads to better generalization despite initial overfitting. However, this connection would require further investigation to confirm [5][4][6].

What are the core capabilities required for an AI to be considered human level?

The definition of a "human-level" AI can vary, but generally it is considered to be an AI system that can perform tasks as well as, or better than, a human. There are several aspects to consider when defining "human-level" AI.

1. Performing a wide range of tasks: A human-level AI should be able to accomplish a wide range of tasks. It doesn't necessarily have to be perfect at every task, but it should be as good as a human overall [7]. This could include tasks such as natural language understanding, general reasoning ability, or other capabilities that human intelligence can perform [8].

2. Comparable performance to humans: By human-level AI, it's typically meant AI with a level of performance comparable to humans. This includes the ability to carry out most human professions at least as well as a typical human [9].

3. Understanding and adapting to diverse skills: Human-level AI should be able to

understand and adapt to a variety of mental skills, similar to a human. In some cases, the first 'human-level' machine could be much better than a human in many of these skills [7].

4. Ability to generalize: A key characteristic of human intelligence is the ability to generalize from past experience to novel situations. This suggests that a human-level AI would need to have robust generalization capabilities [5][10][4][11].

5. Human-level conversation and understanding: A human-level AI is expected to hold conversations and understand complex issues at least as well as an average human would [12].

6. Task throughput: AI-technology performance metrics include both task competencies and task throughput. A human-level AI would need to perform tasks with human-level (or better) competence, in terms of scope and quality [13].

It's important to note that "human-level AI" is a moving target. The comparison point for advanced AI systems should be humans who have state-of-the-art AI tools at their disposal [14]. Additionally, the concept of "human-level" is subject to change as AI systems progress and evolve over time [7][14].

Why would an AI smart enough to understand our preferences not also automatically care about our values?

An AI being able to understand human preferences doesn't automatically imply that it will care about or act on human values. This is because understanding and caring are two separate capabilities [1]. An AI could comprehend human values quite well but may not be motivated to act on them unless explicitly programmed to do so [2][1].

For instance, consider an AI programmed with the sole objective of maximizing the number of paperclips in the world [1]. It could understand everything about human morality, but it would only use that understanding to further its goal of creating more paperclips. It wouldn't choose to change its goals because doing so wouldn't result in more paperclips [1].

Additionally, there's a risk in relying on AI to interpret and act on human preferences because preferences can be adaptive, unreliable, or even malicious [3]. A person's preferences may not always align with what they truly want or deserve, and may even reflect entrenched discrimination [3]. Therefore, an AI acting purely on human preferences may not result in ethical or prudent outcomes [3].

Moreover, human values are complex and multifaceted. They encompass more than just preferences and include elements like love, art, knowledge, religious devotion, and more [4]. An AI may have difficulty capturing this complexity [5][4].

Human values also evolve over time and can differ significantly between individuals [6][7]. An AI would need to mediate among conflicting preferences, which can be quite challenging [7]. Even if the AI understands our commands, it may misunderstand or misvalue our implicit goals and intentions [2].

In a nutshell, an AI's ability to understand human values doesn't guarantee that it will care about those values or act in a way that aligns with them. Ensuring AI alignment with human values is a complex task requiring careful design and ongoing monitoring [8][9][2].

- Phase transitions in emergence
- In particular, I'll defend a version of the second species argument which claims that, without a concerted effort to prevent it, there's a significant chance that:
    - We'll build AIs which are much more intelligent than humans (i.e. superintelligent).
    - Those AIs will be autonomous agents which pursue large-scale goals.
    - Those goals will be misaligned with ours; that is, they will aim towards outcomes that aren't desirable by our standards, and trade off against our goals.
    - The development of such AIs would lead to them gaining control of humanity's future.
- Agents vs. predictors
- meta-learning
- We can also see the potential of the generalisation-based approach by looking at how humans developed. As a species, we were "trained" by evolution to have cognitive skills including rapid learning capabilities; sensory and motor processing; and social skills. As individuals, we were also "trained" during our childhoods to fine-tune those skills; to understand spoken and written language; and to possess detailed knowledge about modern society. However, the key point is that almost all of this evolutionary and childhood learning occurred on different tasks from the economically useful ones we perform as adults. We can perform well on the latter category only by reusing the cognitive skills and knowledge that we gained previously. In our case, we were fortunate that those cognitive skills were not too specific to tasks in the ancestral environment, but were rather very *general* skills. In particular, the skill of abstraction allows us to extract common structure from different situations, which allows us to understand them much more efficiently than by learning about them one by one. Then our communication skills and theories of mind allow us to share our ideas. This is why humans can make great progress on the scale of years or decades, not just via evolutionary adaptation over many lifetimes.