

Transformative AI and Compute

Reading list

Updated every so often ▾ by [Lennart Heim](mailto:lennart@heim.xyz) (lennart@heim.xyz)

(with lots of help and summaries from Konstantin Pilz)

Last updated November 2023 (minor updates in 2024)



By MidJourney

There could be significant progress in AI this century, leading to what has been described as [transformative AI](#) or even [Artificial General Intelligence](#) (AGI). This would involve substantial risks, as capable AI systems may be hard to control and could have devastating effects on society if used by malicious actors. Potentially, advanced forms of AI could constitute an [existential risk](#) to humanity. The topic, hence, deserves much more attention, both the technical aspects of how to create safe systems ([AI alignment](#)) as well as their governance ([AI governance](#)).

Computational resources (compute) refers to the computational infrastructure required to run AI training and inference and is, therefore, a necessity for AI.

Due to its unique properties (quantifiability, rivalrousness, excludability) and *state of affairs*, such as a concentrated supply chain, compute may be one of the most promising nodes to steer the development of beneficial and safe AI.

I put readings into the following buckets:

1. [Compute in the AI Production Function](#): *Why compute matters for AI*
 - Compute is one of the key drivers of AI progress.
2. [Compute Supply Landscape](#): *How chips are produced and used*
3. [Compute Governance](#): *How can we govern compute to achieve beneficial AI outcomes?*
 - Using compute as a governance node by either (a) monitoring, (b) restricting, or (c) promoting access.
4. [Compute-Based Transformative AI Forecasting](#)
 - How much compute we might need to achieve certain transformative and potentially dangerous capabilities
 - How much compute we will have in the future and which computing paradigms will dominate
5. [Others](#): Books, research questions, related topics, newsletters, podcasts, career advice, forecasting

This reading list is not meant as “*read all of this in this order*”, rather “*here are some selected papers and articles on topics related to AI and compute*”.

Overview

Pieces that cover all of the domains below:

- **“Computing Power and the Governance of AI”:** You can find a Twitter summary [here](#), a blog post summary [here](#), and the paper [here](#).
- Podcast: [Lennart Heim on the compute governance era and what has to come after](#), 80,000 Hours Podcast, 2023
 - A broad overview over the motivation behind compute governance, what has recently happened and many other commonly asked questions on the topic.
- [Computational Power and the Social Impact of Artificial Intelligence](#), Hwang, 2018
 - Some early work on the connection of compute and AI. Provides a good overview but is a bit outdated.
- Virtual Talk: [Introduction to Compute Governance](#), Heim, 2023
 - In this talk, I present the idea of using computational resources (short compute) as a node for AI governance. First, I will start by talking about recent events in compute and AI and how they relate to compute governance. I will then discuss the unique properties and state of compute affairs that make it a particularly governable node for AI governance and how this relates to the compute supply chain and other concepts. Subsequently, we will explore the idea of hardware-enabled mechanisms and how they can be used for providing assurances and other AI governance goals. Lastly, I will present our policy work and close with a summary.

Table of Contents

[Transformative AI and Compute](#)

[Overview](#)

[Table of Contents](#)

[1 Compute in the AI Production Function](#)

[1.1 Compute as an input to AI systems](#)

[1.2 Compute usage trends in AI](#)

[1.2.1 Inference](#)

[1.3 Scaling laws](#)

[1.4 Compute efficiency](#)

[2 Compute Supply Landscape](#)

[2.1 Semiconductor supply chain](#)

[2.1.1 Taiwan's dominance and implications for US security](#)

[2.2 Compute provision](#)

[2.1.1 High-Performance Computing \(HPC\) and AI](#)

[2.1.2 Data Centers](#)

[2.3 Trends in Computing Hardware](#)

[2.4 Where's all the compute?](#)

[2.4.1 China](#)

[2.5 What's next in computing?](#)

[3 Compute Governance](#)

[3.1 Motivation](#)

[3.2 Existing policy work relevant to Compute Governance](#)

[3.3 Promoting Access through Compute Funds](#)

[3.4 Measuring compute](#)

[3.4.1 Monitoring compute](#)

[3.4.2 Verification](#)

[3.5 Export Restrictions](#)

[3.5.1 US Export Restrictions from October 2022](#)

[3.6 Verifying Compute Use](#)

[3.7 Hardware Security](#)

[3.8 Compute Providers](#)

[3.10 Corporate Compute Governance](#)

[4 Compute-Based Transformative AI Forecasting](#)

[4.1 TAI forecasting](#)

[4.2 Forecasting compute](#)

[4.2.1 Interpreting AI compute trends and the limits](#)

[4.3 Takeoff scenarios and compute](#)

[4.4 Brain Comparisons and analogies](#)

[5 Others](#)

[5.1 Books](#)

[5.2 Research Questions](#)

[5.3 AI chip architecture](#)

[5.4 Compute and carbon emissions](#)

[5.5 Podcast Episodes](#)

[5.6 Career advice](#)

[5.7 Newsletters](#)

[5.8 Youtube Channels](#)

[5.9 Documentaries](#)

[5.10 Forecasting Platforms](#)

[5.11 Bonus: Information/Compute Security](#)

1 Compute in the AI Production Function


Pieces that cover the role of compute in developing more powerful AI systems

- [AI Chips: What They Are and Why They Matter](#), CSET, Khan, 2020
 - Modern AI training requires specialized hardware. Those AI chips outperform general-purpose chips (CPUs) in speed by ten to a thousand times for machine learning applications and are thus much more energy efficient. Due to their high cost, training a large model can cost up to \$1m. The main advantages of AI chips over CPUs are parallel computing and faster memory access. The report further outlines types of AI Chips and current trends.

1.1 Compute as an input to AI systems

- [The AI Triad and What It Means for National Security Strategy](#), CSET, Buchanan, 2020
 - *“Three components make deep learning happen: data, algorithms, and computing power. Together, I call these components the AI triad.”*
- [A Compute-Based Framework for Thinking About the Future of AI](#), Barnett, 2023
 - *“Compute is ultimately the most important driver of progress in AI, and AI will likely dramatically increase the world economic growth rate later this century. Progress in AI will likely become relatively predictable, allowing us to anticipate AI capabilities before they are fully formed.”*

1.2 Compute usage trends in AI

- [Compute Trends Across Three Eras of Machine Learning](#), Epoch, Sevilla et al., 2022
 - [Summary on Alignment Forum](#)
 - [Twitter summary](#)
 - Database:  Parameter, Compute and Data Trends in Machine Learning
- [AI and Compute](#), OpenAI, Amodei & Hernandez, 2020
- [Compute Trends — Comparison to OpenAI's AI and Compute](#), LessWrong, Heim et al., 2022
 - *Compares the two analyses above*
- [OpenAI overview of how conversational models are made](#) (Video, 42 min), Andrey Karpathy
 - Explains the current LLM development pipeline
- [The Limited Benefit Of Recycling Foundation Models](#), Matthew Barnett, 2023
 - *It seems unlikely that model recycling will result in more than a modest increase in AI capabilities.*
- [AI Capabilities Can Be Significantly Improved Without Expensive Retraining](#), Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas, 2023
 - Fine-tuning can enhance performance by up to 20x on certain benchmarks and takes <1% of the compute of the final training run

- [Extrapolating Performance In Language Modeling Benchmarks](#), David Owen, 2023
 - Attempts to predict future AI model performance based on training compute growth trends

1.2.1 Inference

- [Trading Off Compute in Training and Inference](#), Villalobos & Atkinson, 2023
- The majority of all ML compute today is likely used for inference
 - (Google estimated 60% ([Paterson et al., 2022](#)), NVIDIA estimated 80% to 90%, and AWS estimated that 90% of their workload was inference ([Patterson et al., 2021](#)))
- [Menghani, 2023](#) for an overview of methods to make AI models more compute-efficient

1.3 Scaling laws

Exploring the connection of compute and capabilities of AI systems.

- [Scaling Hypothesis](#) Gwern.net
 - Introduction to the Scaling Hypothesis, that AI systems will continue to get more powerful by increasing their training compute
- [Scaling Laws for Neural Language Models](#), Kaplan et al. (OpenAI), 2020
 - Studies the optimal allocation of a fixed compute budget and derives laws for compute scaling.
- [Training Compute-Optimal Large Language Models](#), Hofman et al. (DeepMind), 2022
 - Update towards data being relatively more important than Kaplan et al, 2020 above found
- [New Scaling Laws for Large Language Models](#), LessWrong, 1a3orn, 2022
 - Summary and comparing the papers above
- [Chinchillas Wild Implications](#), LessWrong, nostalgebraist, 2022
 - The new scaling laws proposed by DeepMind are a major update toward compute being relatively less important and data potentially being a bottleneck for building bigger systems.
- [Extrapolating GPT-N performance - AI Alignment Forum](#), Finnveden, 2020
- Scaling laws for other domains
 - [Scaling Laws for Autoregressive Generative Modeling](#), OpenAI, Henighan et al., 2020
 - [Will we run out of ML data? Evidence from projecting dataset size trends](#), Epoch, Villalobos et al., 2022
 - [Trends In Training Dataset Sizes](#), Pablo Villalobos and Anson Ho, 2022
- [Machine Learning Model Sizes And The Parameter Gap](#), Pablo Villalobos, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Anson Ho, and Marius Hobbhahn, 2022

1.4 Compute efficiency

Algorithmic improvements allow us to achieve similar capabilities for less compute. Therefore, monitoring them informs the prioritization of compute.

- [AI and Efficiency](#), OpenAI, Hernandez & Brown, 2020
- [Revisiting algorithmic progress](#), Epoch, Erdil & Besiroglu, 2022
- [Algorithmic Progress In Computer Vision](#), Ege Erdil and Tamay Besiroglu, 2023
- [Increased Compute Efficiency and the Diffusion of AI Capabilities](#), Pilz, Heim, and Brown, 2023
 - *Increasing compute efficiency (encompassing both hardware and software improvements) has some counter-intuitive consequences. While it diffuses the ability to recreate capabilities, it also allows large compute investors to further improve their models, potentially allowing them to maintain a performance advantage. Society will need to prepare for the impact of widely available dangerous capabilities, though policymakers should still pay particular attention to large compute investors who discover dangerous capabilities first and are often in the best position to mitigate them.*

2 Compute Supply Landscape

- [Supply Chain Explorer](#), CSET, Emerging Technology Observatory, 2022
 - Interactive summary of the compute supply chain, great to get an appreciation for its complexity and learn about specific inputs
 - [A blog post about the main takeaways](#)
- [The Geopolitics Of AI Chips Will Define The Future Of AI](#), Forbes, Rob Toews

2.1 Semiconductor supply chain

With semiconductors, I usually refer to the production of chips — most notably the semiconductor supply chain.

- [Introduction to AI chip making in China: Relevant background, considerations, and forecasting questions](#), IAPS, Grunewald & Phenicie, 2023
- [Quick introduction graphics to the semiconductor industry](#), Twitter, Leandro, 2022
 - The first page covers types of chips and how they are made, the second continues with the production, discusses Moore's law, and gives a brief introduction to the industry's history
- [The Huge Endeavor to Produce a Tiny Microchip](#), NYT, Clark, 2022
 - Short introduction to chip production with a lot of photos. Fabs (factories producing chips) are sophisticated facilities with large cleanrooms with controlled

conditions for producing microscopically precise chips. Construction costs billions and needs various specialized equipment.

- [The Semiconductor Supply Chain](#), CSET, Khan, 2020
 - Summarizes each component of the semiconductor supply chain and where the United States and its allies possess the greatest leverage
- [China, Semiconductors, and the Push for Independence - Part 1](#), Substack, Li, 2021
- [Decoupling in Strategic Technologies](#), CSET, Hwang & Weinstein, 2022
- [ASML's Secret: An exclusive view from inside the global semiconductor giant | VPRO...](#), Documentary, 2022
 - ASML is the only company worldwide capable of producing the lithography machines needed to fabricate advanced chips at scale. They continuously invest in developing the next generation of lithography machines, that could be used to make chips that are even more efficient.
- [How Taiwan became the indispensable economy](#), Nikkei Asia, 2023
- [Mapping the Semiconductor Supply Chain: The Critical Role of the Indo-Pacific Region](#), CSET, Thadani & Allen, 2023
- [What Goes On Inside a Semiconductor Wafer Fab](#) and other videos by Asianometry for an overview of semiconductor manufacturing

2.1.1 Taiwan's dominance and implications for US security

- [Silicon Triangle - The United States, Taiwan, China and Global Semiconductor Security](#), Diamond et al., 2023
 - US view of dependence on Taiwan and threat from China
- [Supply Chain Interdependence and Geopolitical Vulnerability - The Case of Taiwan and High-End Semiconductors](#), Martin et al. 2023, RAND
 - Calling for the US to reduce reliance on Taiwan

2.2 Compute provision

2.1.1 High-Performance Computing (HPC) and AI

- [Reinventing High-Performance Computing: Challenges and Opportunities](#), Reed et al., 2022
 - [Twitter summary](#)
 - I think this paper provides a decent overview of HPC and upcoming challenges (among others for AI).
- [A Roadmap for Big Model](#) (withdrawn, older version [here](#)), Yuan et al, 2022
 - Paper from China laying out a roadmap for creating big models. [Section 4](#) lays out some important considerations for computing and training such “big models”. While this paper has been accused of plagiarism, I still think it provides a decent overview of the challenges for AI compute.

- [Exascale Computing Technology Challenges](#), Shalf et al., 2011
 - As clock speed increases become more difficult due to physical limits of shrinking transistor size, parallelizing on the chip could deliver further performance gains. The paper explores connected challenges.
- [Techniques for training large neural networks](#), OpenAI, 2022
 - Covers parallelization, a major challenge when setting up training runs.

2.1.2 Data Centers

- [Compute at scale: A broad investigation into the data center industry](#), Pilz & Heim, 2023
 - Provides an overview of how the industry works, what its key inputs are and draws some conclusions for AI governance, emphasizing data centers' importance in the AI compute supply chain and their role in training and deploying advanced AI systems. It outlines key features of data centers, such as power consumption, cooling requirements, and infrastructure, while also discussing market size and growth, major players, and potential bottlenecks.
 - See [An assessment of data center infrastructure's role in AI governance](#), Pilz, 2023 for an interpretation of the key findings and some helpful models for how to think about data centers.

2.3 Trends in Computing Hardware

- [Moore's Law, AI, and the pace of progress](#), LessWrong, Veedrac, 2021
 - Some insights into recent hardware trends and why they (don't) matter for AI.
- [The big question of how small chips can get | Financial Times](#), Bradshaw & Gross, 2023
- Benchmarks (*not necessarily a "reading" but rather a resource*)
 - Benchmarks are useful to measure progress in computing hardware. Find below some commonly used benchmarks for HPC and ML-specific workloads. Note that [various caveats apply](#).
 - [TOP500](#) / Graph500 / Green500
 - The TOP500 benchmark is not an ML benchmark and uses different computation schemes (FP64 etc.). Some of them have GPUs, some don't. One should expect new systems to have GPUs and be usable for AI workloads.
 - [ML Commons Training HPC](#)
 - Benchmark suite for ML training
 - [Trends in GPU price-performance](#), Epoch, Hobbhahn & Besiroglu, 2022
- [Predicting GPU performance](#), Epoch, Hobbhahn & Besiroglu, 2022
- [Trends in Machine Learning Hardware](#), Hobbhahn et al., 2023
 - Updated overview of efficiency increases in AI hardware. Performance doubles every 2.3 years, and memory bandwidth only every 4.1 years.

2.4 Where's all the compute?

- [The world's distribution of computation \(initial findings\) - Machine Intelligence Research Institute](#), MIRI, Naik, 2014
 - *[Note that I would be interested in an updated investigation of this. Got a couple of ideas on how to go about it.]*
- [Compute at Scale](#), Pilz & Heim, 2023 - Section “Locations of data centers”

2.4.1 China

- [White Paper on China's Computing Power](#), ChinAI, 2023

2.5 What's next in computing?

- [There's plenty of room at the Top: What will drive computer performance after Moore's law?](#), Leiserson et al. 2020
 - Although Moore's law has been slowing down in the last years, further advances in computing could be gained through better algorithms, code, and specialized hardware.
- [The Hardware Lottery](#), Hooker, 2020
 - Our current hardware paradigms are dictating which AI innovations succeed.
- [The decline of computers as a general purpose technology](#), Communications of the ACM, Thompson, 2021
 - We will see more specialized chips in the future — first specialized ones for AI and then even for different AI workloads (such as inference and training).
- [Beyond CMOS: the Future of Semiconductors](#), IEEE IRDS, 2018
- [The Accelerator Wall: Limits of Chip Specialization](#), Fuchs & Wntzlaff, 2019

3 Compute Governance

The previous readings laid out the technical foundation of AI and compute. Compute Governance focuses on compute as a governance lever by either monitoring, restricting, or promoting access.

3.1 Motivation

- [AI Governance: Opportunity and Theory of Impact](#), Allen Dafoe (Centre for the Governance of AI), 2020
 - General motivation for governing advancing AI systems to limit risks to society coming from different aspects of the technology
- [Section 6 in Compute Governance and Conclusions - Transformative AI and Compute \[3/4\]](#), Heim, 2021
 - Compute is a uniquely governable input into AI systems because it is specialized, expensive, highly centralized and requires physical space and energy to be

deployed. This means restrictions on compute can be enforced more easily than restrictions on data or algorithms.

- [Compute and Antitrust](#), Verfassungsblog, Haydn Belfield & Shin-Shin Hua, 2022
 - *We argue that the antitrust and regulatory literature to date has failed to pay sufficient attention to compute, despite compute being a key input to AI progress and services, the potentially substantial market power of companies in the supply chain, and the advantages of compute as a ‘unit’ of regulation in terms of detection and remedies.*
- [Compute Accounting Principles Can Help Reduce AI Risks](#), Jackson et al., 2023
- [Arms Control for Artificial Intelligence - Texas National Security Review](#), Lamberth & Scharre, 2023
- [How We Can Regulate AI](#), Avital Balwit, 2023
 - *“The chips used to train the most advanced AIs are scarce, expensive, and trackable — giving regulators a path forward.”*

3.2 Existing Policy Work Relevant to Compute Governance

- [UK’s Future of Compute Review](#) (2022)
 - The UK government calls for proposals to establish a national compute strategy
 - [Future of compute review - submission of evidence](#), CLTR, Whittlestone et al., 2022
 - Suggests enabling more access to compute for academia and beginning to monitor (AI) compute.
 - [GovAI Response to the UK Future of Compute Review - Call for Evidence](#), Heim & Anderljung, 2022
 - Suggests structured access to AI compute favoring research focused on interpretability and safety and ensuring that irresponsible actors do not get access.
 - [Response to the UK’s Future of Compute Review: A missed opportunity to lead in compute governance](#), CLTR & GovAI, and others, 2022
- [US National AI Research Resource \(NAIRR\) Request for Information](#) (2022)
 - Relevant to any discussion around giving academics more compute and setting up (inter-)national computing centers.
 - [Submission to the Request for Information \(RFI\) on Implementing Initial Findings and Recommendations of the NAIRR Task Force](#), GovAI, Heim & Anderljung, 2022
 - Suggests giving academia access to pre-trained models via an API while ensuring research is conscious of safety concerns and favoring actors doing safety-relevant research.
- [A blueprint for building national compute capacity for artificial intelligence](#), OECD AI, 2023
- [Biden’s Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence](#), White House, 2023
 - Implements compute thresholds

3.3 Promoting Access through Compute Funds

As machine learning models get increasingly larger, research requires access to large amounts of compute. Consequently, there have been calls to establish compute funds to provide subsidized access for academia.

- [White Paper | Building a National AI Research Resource](#), Stanford HAI, Ho et al., 2021
 - Proposes a US national project to give academia access to compute to ensure it can work on state-of-the-art AI models.
- [Compute Funds and Pre-trained Models](#), GovAI, Anderljung et al., 2022
 - Suggests giving academia access to large AI models via structured access
 - [Twitter summary](#)
 - Based on the above recommendation, we submitted to the US National AI Research Resources (NAIRR) task force (see section above)
- [The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?](#), Besiroglu, Bergerson, Michael, Heim, Luo, and Thompson, 2023
 - *“We show that a compute divide has coincided with a reduced representation of academic-only research teams in compute intensive research topics, especially foundation models. We argue that, academia will likely play a smaller role in advancing the associated techniques, providing critical evaluation and scrutiny, and in the diffusion of such models. Concurrent with this change in research focus, there is a noticeable shift in academic research towards embracing open source, pre-trained models developed within the industry.”*

3.4 Measuring compute

Measuring compute is crucial for understanding the AI capabilities of relevant actors. It’s also a prerequisite to regulating or allocating compute.

- [Large-scale computing: the case for greater UK coordination](#), UK Government, 2021
 - Describes the economic value of computing infrastructure and suggests more public investment
- Measuring national compute capacity for Artificial Intelligence (AI): existing measurement tools and preliminary findings, OECD, 2022 (*not declassified yet*)
 - Related blog posts
 - [Measuring compute capacity: a critical step to capturing AI's full economic potential](#), OECD.AI, Strier et al., 2022
 - Most countries don’t know how much compute they have. As compute is a key driver for innovation, they will need to measure it to make informed decisions. To stay competitive, countries will need to invest into national compute resources.
 - [Reducing the carbon emissions of AI](#), OECD.AI, Patterson, 2022

- Suggests reducing carbon emissions of AI training by supporting further efficiency gains in algorithms and hardware and incentivizing the use of green energy.

3.4.1 Monitoring compute

- [Why and How Governments Should Monitor AI Development](#), Whittlestone & Clark, 2021
 - Governments already face challenges in relation to governing AI. These will get more severe as systems advance. To make informed decisions and notice trends early, they should set up mechanisms for monitoring different aspects of AI developments as well as their impacts.
- Pieces mentioning and discussing compute monitoring
 - [Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims](#), Brundage et al. (OpenAI), 2020
 - 4.2: The lack of standards is an obstacle to more widespread reporting of compute used for training AI systems. AI labs should start reporting their system specifications in more detail to encourage a wide adoption of the procedure.
 - [Open Problems in Cooperative AI](#), DeepMind, Dafoe et al., 2020

3.4.2 Verification

Verifying and proving the amount of compute one has.

- [Joint Verification Experiments](#), Arms Control Wonk, Krepon, 2011
 - [requires explanation]

3.5 Export Controls

For the moment, this is mostly focusing on the semiconductor supply chain.

- All hardware and compute publications by CSET:
 - [Banned in DC - Examining Government Approaches to Foreign Technology Threats](#), CSET, Corrigan et al., 2022
 - [Securing Semiconductor Supply Chains](#), CSET, Khan, 2021
- [Slowing Moore's Law: How It Could Happen](#), Gwern, Branwen, 2017
- [Digitizing Export Controls: A Trade Compliance Technology Stack?](#), Center for Strategic and International Studies (CSIS), Reinsch & Benson, 2021
- [Microsoft and OpenAI's comment to BIS](#), O'Neal & Clark, 2020
 - advocates for "smart" export controls
- [CoCom's daughter?](#), CSET, Weinstein & Wolf, 2022 discusses the potential for new multilateral export controls.
 - [Associated presentation](#)

3.5.1 US Export Restrictions from October 2022

The Biden administration restricted the export of advanced AI hardware and tools and software required to build it to China (and Russia) in September 2022.

- [Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items: Supercomputer and Semiconductor End Use: Entity List Modification](#), Bureau of Industry and Security and Department of Commerce, 2022
- [Implementation of Additional Export Controls: Certain Advanced Computing Items: Supercomputer and Semiconductor End Use: Updates and Corrections](#), Bureau of Industry and Security and Department of Commerce, 2023
- [Choking Off China's Access to the Future of AI](#), CSIS, Allen, 2022
- Media summaries:
 - SemiAnalysis: [China and USA Are Officially At Economic War – Technology Restriction Overview](#)
 - NYT: [Biden Administration Clamps Down on China's Access to Chip Technology](#)
- South China Morning Post: [Tech war: US chip restrictions could cost 0.6 per cent of China's GDP and weigh on yuan, according to Barclays report](#)
- [Freeze-in-Place: The Impact of US Tech Controls on China](#), Rhodium Group, Goujon et al., 2022
- [New Chip Export Controls and the Sullivan Tech Doctrine with Kevin Wolf](#), China Talk, Schneider and Zhang, 2022
- [A Bombshell For U.S.-China Tech Relations](#), NOEMA, Jordan Schneider, 2022
- [The AI Lockout](#), The Wire China, 2023 ([pdf version](#))
- [Japan and the Netherlands Announce Plans for New Export Controls on Semiconductor Equipment](#), CSIS, Allen & Benson, 2023
- [Chinese Firms Are Evading Chip Controls](#), Tim Fist, Lennart Heim, and Jordan Schneider, 2023

3.6 Hardware-enabled mechanisms

- [Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing](#), Aerne et al., 2024
- [Hardware-Enabled Governance Mechanisms](#), RAND, Kulp et al, 2024
 - The authors introduce the concept of hardware-enabled governance mechanisms (HEMs), which have the potential to help achieve U.S. artificial intelligence (AI) governance goals. They explore the export control–related policy objectives that HEMs can support, analyze the threats that HEMs might face, examine the attack vectors that might compromise them, and describe protection measures that could be taken to counter those attacks.

3.7 Verifying Compute Use

- [What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring](#), Yondav Shavit, 2023
 - Hardware-based mechanisms on AI accelerators could periodically log hashed snapshots of the model weights in their memory. These could be used to verify claims about training specifications, allowing governments or international organisations to enforce rules on responsible AI development.
 - One of the most thorough proposals of a compute monitoring regime
- [Tools for Verifying Neural Models' Training Data](#), Dami Choi, Yonadav Shavit, and David Duvenaud, 2023
 - “We introduce the concept of a “Proof-of-Training-Data”: any protocol that allows a model trainer to convince a Verifier of the training data that produced a set of model weights.”
- [Proof-of-Learning: Definitions and Practice](#), Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot, 2021

3.8 Hardware Security

- [AI Verification: Mechanisms to Ensure AI Arms Control Compliance](#), CSET, Mittelsteadt, 2021
- The Malicious Use of Artificial Intelligence: [Forecasting, Prevention, and Mitigation](#), maliciousaireport.com, Brundage et al., 2018
 - See secure hardware recommendations
- [AI Research Considerations for Human Existential Safety](#), Critch & Krueger, 2022
 - See secure hardware recommendations
- [Structured Access: An Emerging Paradigm for Safe AI Deployment](#), Shevlane, 2022
 - See discussion of hardware-level mechanisms to control access to models
- Some specific technical papers on the above mechanisms:
 - [Hardware-Assisted Intellectual Property Protection of Deep Learning Models](#), IEEE, Chakraborty et al., 2020
 - [Deep-Lock: Secure Authorization for Deep Neural Networks](#), Alam et al., 2020
 - [AdvParams: An Active DNN Intellectual Property Protection Technique via Adversarial Perturbation Based Parameter Encryption](#), Xue et al., 2021
- [Enabling IC Traceability via Blockchain Pegged to Embedded PUF](#), Islam & Kundu, 2019

3.9 Compute Providers

Compute providers are at the end of the supply chain and provide access to compute as a service. How could they be governed?

- [Structured access: an emerging paradigm for safe AI deployment](#), Shevlane

- [Blog post](#) summarizing it
- [Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers](#), Egan & Heim, 2023
 - Introduces how KYC practices can contribute to safety and accountability for AI developers

3.10 Corporate Compute Governance

What could actors providing compute, such as AI hardware companies, do for AI safety?

- [soon]

3.11 International Compute Governance

- [International Institutions for Advanced AI](#), Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, and Duncan Snidal, 2023

4 Compute-Based Transformative AI Forecasting

4.1 TAI forecasting

- [Updates and Lessons from AI Forecasting](#), Steinhardt, 2021
- [More Is Different for AI](#), Alignment Forum, Steinhardt, 2022
- [AI Timelines: Where the Arguments, and the "Experts," Stand](#), Cold Takes, Karnofsky, 2021
 - Overview of current attempts at AI forecasting
 - [Grokking "Forecasting TAI with biological anchors"](#), LessWrong, Ho, 2022
 - Well written and visualized summary of bio-anchors report

4.2 Forecasting compute

Forecasting how much compute we might need and how much we will have.

- [Draft report on AI timelines](#) Alignment Forum, Cotra, 2021
 - Split into two aspects: (I) how much compute we might need based on biological anchors, and (II) how much compute might be available.
- [Forecasting Compute - Transformative AI and Compute \[2/4\]](#), EA Forum, Heim, 2021
- [Projecting compute trends in Machine Learning](#), Epoch, AI Alignment Forum, Besiroglu et al., 2022
- [Compute Trends Across Three Eras of Machine Learning](#), Sevilla et al., 2022

4.2.1 Interpreting AI compute trends and the limits

- [Interpreting AI compute trends](#), AI Impacts, Carey, 2018

- [Reinterpreting “AI and Compute”](#), AI Impacts, Garfinkel, 2018
- [The Computational Limits of Deep Learning](#), Thompson, 2020
 - I don't agree with their exact numbers and how they draw the comparison, however, I think the main message still holds.
- [How Much Longer Can Computing Power Drive Artificial Intelligence Progress?](#), CSET, Lohn & Musser, 2022
 - Note that their model does assume that computing cost stays constant. Also, they assume a 3.4-month doubling time which is not up to date anymore.
- [This can't go on\(?\) - AI Training Compute Costs](#), Heim, 2023
- [Trends In The Dollar Training Cost Of Machine Learning Systems](#), Ben Cottier, 2023
 - The cost of the largest final training run of ML systems has recently grown about 10x every two years.

4.3 Takeoff scenarios and compute

- [Inference cost limits the impact of ever larger models](#), LessWrong, Mindermann, 2021
- Definition of [Hardware overhang](#) on AI Impacts

4.4 Brain Comparisons and analogies

- [Brain Efficiency: Much More than You Wanted to Know](#), LessWrong, Canell, 2022
- [How Much Computational Power Does It Take to Match the Human Brain](#), Open Philanthropy, Carlsmith, 2021

5 Others

Anything that doesn't fit into the above categories.

5.1 Books

- [Chip War](#), Christopher Miller, 2022
 - Gives a broad overview of the history of semiconductors and the conflicts around them, shedding light on many contemporary issues and the geopolitics of chips.

5.2 Research Questions

- [Some AI Governance Research Ideas](#), Anderljung & Carlier, 2021
- [AI Governance Needs Technical Work](#), EA Forum, Baker, 2022
- [TODO: add updated doc]

5.3 AI chip architecture

If you want to have a more in-depth understanding of AI accelerators, here's a great series. My best guess is that this is probably not relevant for most projects. However, some understanding might be helpful in categorizing new innovations.


- [AI Accelerators — Part I: Intro](#), Medium, Fuchs, 2022.
 - Youtube interview and summary: [All about AI Accelerators: GPU, TPU, Dataflow, Near-Memory, Optical, Neuromorphic & more \(w/ Author\)](#)
- [Tutorial on Hardware Accelerators for Deep Neural Networks](#), by Emer et al., MIT/ Nvidia
- Memory [to explain]
 - ...
- Interconnect [to explain]
 - [An Overview of Efficient Interconnection Networks for Deep Neural Network Accelerators](#), IEEE, Nabavinejad et al., 2020

5.4 Compute and carbon emissions

When talking about compute many people are concerned about its carbon footprint. Unfortunately, there have been some wrong numbers going around, somewhat exaggerating the effects on climate change. Carbon emissions also provide some insights into the economics and distribution of AI workloads in data centers.

- [The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink](#), Paterson et al., 2022
 - Predecessor paper: [Carbon Emissions and Large Neural Network Training](#), Patterson et al., 2021
- [The Carbon Emissions of Big Tech](#), Rodrigo Navarro, 2023, ElectronicsHub
 - Claims most emissions come from semiconductor manufacturing rather than data center operations.

5.5 Podcast Episodes

- [Nvidia: The Dawn of the AI Era - Acquired](#) (175 minutes)
- [Danny Hernandez on forecasting and the drivers of AI progress](#), 80,000 Hours
- [TSMC Deep Dive with Asionanometry](#), Compounding Curiosity, 2022
 - Some interesting insights into Taiwan and its TSMC cult.
- [The Asianometry Podcast](#) (audio version of YouTube channel linked below)
 -  [Semiconductors: Everything You Wanted to Know](#)
- [Markus Anderljung on AI Policy](#), The Inside View, 2022
 - Covers some of our work at GovAI on compute governance.
- [Lennart Heim \(that's me!\) on AI Compute 101: The Geopolitics of Giant models](#), ChinaTalk, 2023

- [AI Compute 101: The Geopolitics of GPUs](#)
- [Compute and the Future of US-China Relations](#)
- [How the rise of AI will affect Taiwan's semiconductor industry](#), Startup Island, 2023
 - Good summary of current trends in semiconductors, with strong focus on Taiwan
- ChinaTalk
 - [Huawei, SMIC, and Future of Export Controls: SemiAnalysis and Fabricated Knowledge in Conversation](#) SEP 18, 2023
 - Many interesting episodes on export restrictions in fall 2022
 - [New Chip Export Controls and the Sullivan Tech Doctrine with Kevin Wolf](#)
 - [Choking Off China's AI Access - by Jordan Schneider](#)
 - [Export Controls, Xi's S&T Dreams, and "Technological Vassaldom"](#)

5.6 Career advice

- [What does it mean to become an expert in AI Hardware?](#), EA Forum, Phenicie 2021
- [Expert in AI hardware - Career review](#), 80,000 Hours, 2021

5.7 Newsletters

- [SemiAnalysis](#)
- [The Asianometry Newsletter](#)
- [Fabricated Knowledge](#)
- [Semi-Literate](#)
- Sometimes touches on hardware/compute
 - [ChinAI Newsletter](#)
 - [CSET Newsletter](#)

5.8 Youtube Channels

- [TechTechPotato](#)
- [Asianometry](#)
- [Stanford MLSys Seminars - YouTube](#)
 - [Notes on AI Hardware - Benjamin Spector | Stanford MLSys #88 \(youtube.com\)](#)
- Misc:
 - [AI Hardware w/ Jim Keller \(youtube.com\)](#)

5.9 Documentaries

[The race for semiconductor supremacy | FT Film](#) (Sep 2023)

5.10 Forecasting Platforms

- [INFER Pub has many forecasting questions on microelectronics](#) in regards to China and the US
- Metaculus has many questions [related to compute](#) and [TAI timelines](#)
- [Chinese AI Chips | Metaculus](#)

5.11 Bonus: Information/Compute Security

I think information security is related to compute governance. To put it simply, we won't be able to restrict access to compute if you can just hack into the cluster and misuse it.

- [Information security in high-impact areas - Career review](#), 80,000 Hours, 2023
- [Information security considerations for AI and the long term future](#), EA Forum, Ladish & Heim, 2022
- Podcast: [Nova DasSarma on why information security may be critical to the safe development of AI systems](#), 80,000 Hours, 2022