

Medium Level AI Governance Reading List
Aaron Scher | updated occasionally including through August 2026

Note: This serves as a list of “my recommendations for the most important mid-level AI governance readings”. If you have not first engaged with the more intro-level things such as [AISE](#), I recommend doing that first.

High priority

- Strategy readings (~9–16 hr)
 - Read the book [If Anyone Builds It Everyone Dies](#) by Yudkowsky and Soares. 7 hr.
 - Read [Nearcast-based "deployment problem" analysis](#) by Holden Karnofsky about the AI deployment problem. 1 hr.
 - Deep skim [Situational Awareness](#) by Leopold Ashenbrenner. 2 hr.
 - Read [Response to Aschenbrenner's "Situational Awareness"](#) by Rob Bensinger. 5 min.
 - Read [MIRI Technical Governance Team's research agenda](#). 1.75 hr.
 - Read [An International Agreement to Prevent the Premature Creation of Artificial Superintelligence](#). 1.5 hr.
 - There are also [some blog posts](#) explaining a few parts of the agreement in more detail. These are medium priority. 45 min.
 - You could also watch [the related lecture](#). 30 min.
 - Skim [A Narrow Path](#) by various. 1 hr.
 - Skim [AI 2027](#) and their research reports. 45 min.
 - [Skim AI 2040](#), Plan A. 30 min.
 - Skim [Plans A, B, C, and D for misalignment risk](#). 15 min.
 - Skim [What's the short timeline plan?](#) by Marius Hobbhan and its comments. 30 min.
 - Deep skim [Chips for Peace](#) by Cullen O'Keefe. 10 min.
- Other (~3 hr)
 - In [Mechanisms to Verify International Agreements About AI Development](#), read the Executive Summary, the [Background and Motivation](#) section, and the appendix section titled [Hypothetical Futures](#). 30 min.
 - Read summary and skim [Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment](#) (this is a more recent and better paper on the overall question of how to do verification). 45 min.
 - [Here](#) is a synthesis for the state of the field of technical AI verification. 10 min. without the works linked inside this doc, hours to skim them all.
 - [The Closing Window: How Governments Could Lose Their Ability to Restrain Advanced AI](#). 20 min.
 - [How to pace the US frontier](#). 25 min.

Medium priority

- Suggested AI governance readings (important work that people should be aware of) (~2 hr)
 - Skim [Securing AI Model Weights](#) about AI model weight security. 10 min.
 - Skim through the various pieces in [this series](#) about compute. 15 min.
 - Skim through [Open Problems in Technical AI Governance](#) laying out a bunch of open problems in technical AI governance. 15 min.
 - Skim [Towards Best Practices in AGI Safety and Governance](#) overviewing what AI governance experts want. 15 min.
 - Skim [Societal Adaptation to Advanced AI](#) which I think adds useful framings on top of the traditional “prevent harm” frame. 10 min.
 - Skim [The Semiconductor Supply Chain](#) about the chip supply chain. 15 min.
 - Skim [2024 State of the AI Regulatory Landscape](#). 15 min
 - Skim [Technical Requirements for Halting Dangerous AI Activities](#) MIRI TGT early work. 15 min.
 - Do a medium-depth read in order to understand the key ideas for the blog post [What Increasing Compute Efficiency Means for the Proliferation of Dangerous Capabilities](#) or the [accompanying paper](#). 10 min.
 - [How to Catch a GPU: A Taxonomy of Verification and Enforcement Mechanisms for International AI Agreements](#)
 - [Warning Shots for AI Existential Risk: Will society answer the wake-up calls it receives?](#)
 - [The Scramble: getting in position to pace the frontier](#)
- Familiarize yourself with what work exists already in AI governance land, broadly (~5+ hr)
 - Briefly look at all research on [GovAI’s website](#) (read all titles, skim/read a few that seem especially interesting). 45 min.
 - Briefly look at all blog posts, reports, datasets, and resources on [Epoch’s website](#) (read all titles, skim/read a few that seem especially interesting). 1.5 hr.
 - Briefly look at all blog posts and papers on [IAPS’s website](#) (read all titles, skim/read a few that seem especially interesting). 20 min.
 - Briefly skim through relevant [CSIS articles](#) (not that many are actually relevant). 15 min.
 - Briefly look at the content and resources on [Artificial Analysis’s website](#). 10 min.
 - Briefly skim stuff from [CNAS website](#) (not much). 5 min.
 - Briefly look at all blog posts and papers on this short list from [CSET](#) (they put out tons of stuff, so start here but feel free to expand; read all titles, skim/read a few that seem especially interesting). (I haven’t actually read these — eek). 1 hr.
 - Skim through the blog posts on [Lennart Heim’s](#) website (read all titles, skim/read a few that seem especially interesting). (These might be good examples of what the output of a shallow dive looks like). 20 min.
 - Look through this reading list from Lennart Heim about [Transformative AI and Compute](#). It’s now a bit old, but many of the things are still relevant. 20 min.
 - Skim through stuff from the [Center for Long-Term Resilience](#). 10 min.

- Skim through stuff from [MIRI's Technical Governance Team](#) including the [blog page](#). 10 min.
- Skim through stuff from [Miles Brundage](#). 15 min.
- Skim through stuff from [AI Impacts](#). 15 min.
- Let's give you stuff from [Forethought](#). 15 min.
- Skim through stuff from [Amodo Design](#) about AI verification. 15 min.
- Skim [this map](#) of other organizations in the AI governance space (out of date), more general [AI safety map](#). 10 min.
- **Fill in any gaps in your knowledge that seem really important to fill in (many hr). Consider talking to ChatGPT or Claude, googling, doing other obvious things to resolve your confusion.**

Low priority

These are things that I don't particularly endorse but think it's worth being aware of

- [Toward a Global Regime for Compute Governance: Building the Pause Button](#)
- [Pacing the Frontier A Framework & Research Agenda](#)

Important topics not covered above

- AGI timelines
 - [This comment](#).
 - TO DO
- AGI threat models, difficulty of alignment, overview of methods for improving AI safety
 - [ifanyonebuildsit.com](#)
 - TO DO

BELOW THIS POINT ARE MORE PROJECT-SPECIFIC READING LISTS (that I don't necessarily endorse)

China related

- High priority. China-related readings (**most relevant for China-related projects**). Skim through (but obviously feel free to dig deeper) the following (~3 hr):
 - Jeff Ding's [ChinAI](#) newsletter (i.g., skim the last 10 issues). 30 min.
 - This surprisingly outdated [piece](#) about China's LLM landscape (maybe an intermediate project for this SPAR project would be to do an updated version?). 5 min.
 - The [Governing Through the Cloud](#) paper. 10 min.
 - I haven't read most of the things linked in this [old syllabus](#), but maybe skim a couple to get a general sense of what work people used to do. 10 min.
 - Skim [Preventing AI Chip Smuggling to China](#). 10 min.
 - Deep Skim [Countering AI Chip Smuggling Has Become a National Security Priority](#), a good example of open source intelligence research, I think. 10 min.
 - Deep Skim [Introduction to AI chip making in China](#). 20 min.
 - Deep Skim [The PRC's Domestic Approach](#). 20 min.
 - Deep Skim [The PRC's Efforts Abroad](#). 20 min.
 - Skim [A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports](#) about the 2023 US chip export control updates. 10 min.
 - Skim [The State of Chinese AI](#). 10 min.
 - Skim through [Matt Sheehan's research](#) on China. 10 min.
 - Skim the podcast (transcript) [Sihao Huang on the risk that US-China AI competition leads to war](#) about China's AI development. 30 min.
- Lower priority. Other China readings (~5 hr)
 - This [reading list](#) has many things. I have not read the vast majority of them. Consider spending 3 hr reading various things from the list that seem interesting to you. 3 hr.
 - Look around the writings from [Concordia AI](#). 30 min.
 - Read [The Illusion of China's AI Prowess](#). 20 min.
 - Read [Lennart Heim's tweet thread](#) about Jan 2025 export control updates. 10 min.
 - [China's Advanced AI Research](#) CSET report. 1 hr.
 - [Why China isn't about to leap ahead of the West on compute](#). 15 min.
 - [ArtificialAnalysis Country Analysis](#). 5 min.
 - [DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race](#). 30 min.

Consumer hardware related

- Quick list of relevant stuff, not necessarily endorsed.
- [Epoch AI Hardware data](#)
- [Supply chain concentration](#)
- [SemiAnalysis](#)
- [Epoch where are the GPU clusters](#)
- [NVIDIA's Deep learning performance data](#)
- [NVIDIA's HPC performance data](#)
- [AI 2027 Chip forecasts](#)
- [MLPerf benchmarks](#)
 - <https://mlcommons.org/benchmarks/inference-datacenter/>
 - <https://mlcommons.org/benchmarks/inference-edge/>
- [Thread about distillation](#)
- [Thread with a bunch of new AI models and technical reports and similar](#). Relevant highlights:
 - [Google DM guide](#) on scaling models (esp on TPUs).
 - [Huggingface guide](#) on scaling models
 - [This paper](#) from Oak Ridge National Lab about large-scale training with AMD GPUs seems good for generally understanding infra.
- [Cost Effective Deployment of DeepSeek R1 with Intel® Xeon® 6 CPU on SGLang](#)
- [H2:Towards Efficient Large-Scale LLM Training on Hyper-Heterogeneous Cluster over 1,000 Chips](#)
- [Artificial Analysis Hardware Benchmarking](#)
- [Reddit: Consumer hardware landscape for local LLMs June 2025](#)
- [Frontier AI performance becomes accessible on consumer hardware within a year](#)
- [Towards Large Scale Training on Apple Silicon](#)
- [Workshop: Tiny Titans: The next wave of On-Device Learning for Foundation Models \(TTODLer-FM\)](#)
- [Distributed and Decentralised Training: Technical Governance Challenges in a Shifting AI Landscape](#)
- [Llama.cpp](#)
- [**LLM Inference Benchmarking Cheat-Sheet for Hardware Reviewers**](#)
- [LLM Performance on Consumer Hardware](#)
- [Are Consumer GPUs a Problem for US Export Controls?](#)
- [A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports](#)
- [Efficiently Scaling Transformer Inference](#)
- [Inference Economics of Language Models](#)
- [Performance of llama.cpp with Vulkan #10879](#)
- [LLM Inference Economics from First Principles](#)
- [Inside NVIDIA GPUs: Anatomy of high performance matmul kernels](#)
- [STATE OF AI REPORT 2025](#)
- [HexiScale: Accommodating Large Language Model Training over Heterogeneous Environment](#)
- huggingface.js/packages/tasks/src/hardware.ts

- [Intelligence per Watt: Measuring Intelligence Efficiency of Local AI](#)
- [Opportunistic Expert Activation: Batch-Aware Expert Routing for Faster Decode Without Retraining](#)
- [What share of all chips are high-end data center AI chips?](#)
- [InferenceMAX™: Open Source Inference Benchmarking](#)
- [Basic facts about GPUs](#)
- [How much computing capacity exists in GPUs and TPUs in Q1 2023?](#)
- [Which GPU\(s\) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning](#)
- [Openbenchmarking Machine Learning](#)

Verification Reading List

- Overviews
 - Naci's April 2026 overview of the space:
 - ▢ **The state of technical AI verification**
 - Somebody else's reading list: ▢ **[Shared] Verification Reading List**
 - In [Mechanisms to Verify International Agreements About AI Development](#) (listed above), read the Executive Summary, the [Background and Motivation](#) section, and the appendix section titled [Hypothetical Futures](#). 30 min.
 - **Read summary and skim [Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment](#) (listed above). 45 min.**
 - [An International Agreement to Prevent the Premature Creation of Artificial Superintelligence](#) (listed above) (not exactly an overview, but useful I expect). 1 hr.
 - [The Compute Verification Post](#). 30 min.
- Foundational verification readings
 - [Secure, Governable Chips](#)
 - [What does it take to catch a Chinchilla?](#)
 - [Hardware-Enabled Governance Mechanisms](#)
 - [Tools for Verifying Neural Models' Training Data](#)
 - [Computing Power and the Governance of Artificial Intelligence](#) (listed above)
- Other important verification readings according to me
 - [Detecting Hidden ML Training With Zero-Overhead Telemetry](#)
 - [Research Note: The Fundamentals and Feasibility of Secure Network Taps for Verifying AI Datacenter Use](#)
- Other verification work (misc., at some point these should be categorized more)
 - [Verification for International AI Governance](#)
 - [Verification methods for international AI agreements](#)
 - [Nuclear Arms Control Verification and Lessons for AI Treaties](#)
 - (from above) Read [Article VII](#) including precedent and commentary in [An International Agreement to Prevent the Premature Creation of Artificial Superintelligence](#)
 - [Hardware-Enabled Mechanisms for Verifying Responsible AI Development](#)

- [Timing and Memory Telemetry on GPUs for AI Governance](#)
- [Haiku to Opus in Just 10 bits: LLMs Unlock Massive Compression Gains](#)
- [Can governments quickly and cheaply slow AI training?](#)
- Be aware of [flexHEG](#), [Flexible Hardware-Enabled Guarantees for AI Compute Plan A](#) and supplements
- [100,000 H100 Clusters: Power, Network Topology, Ethernet vs InfiniBand, Reliability, Failures, Checkpointing](#)
- [Securing deep frontier AI evaluations](#)
- [Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation](#)
- [Guaranteeable Memory: An HBM-Based Chiplet for Verifiable AI Workloads](#)
- [Securing AI Inference at Scale](#)
- [Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies](#)
- [Requests for discussion on AI infrastructure security and AI verification](#)
- [June 2026 A System Overview for Near-Term, Low-Trust AI Compute Verification](#)
- [Petrie and Drotos Off Switch \(in progress?\)](#)
- [Memory Wipes - Performance Analysis](#)
- [Hardware-Level Governance of AI Compute: A Feasibility Taxonomy for Regulatory Compliance and Treaty Verification](#)
- [Suppressing Side Channels in an Untrusted Data Center via Retrofitted Defenses](#)
- [A taxonomy of physical side channels for AI security and verification](#)
- [Software-Based Memory Erasure with relaxed isolation requirements: Extended Version](#)
- [Verifying Restrictions on Frontier AI Research](#)
- [Verifiable constraints on frontier training via proofs of compartmentalization](#)
- [An International AI Slowdown Is Ready Whenever Politicians Are](#)
- [Internationalising AI Verification](#)
- [Zero knowledge verification for frontier AI training is possible](#)
- [Privacy-Preserving AI Verification via Minimal Information Disclosure](#)
-
- Orgs doing verification work or very-verification-related work that are worth being aware of (mid 2026):
 - [Amodo Design](#)
 - [Lucid Computing](#)
 - [RAND](#)
 - [MIRI Technical Governance Team](#)
 - [Tinfoil](#)
 - [Attestable](#)
 - [Confidential AI](#)
 - [Singapore AI Safety Hub \(SASH\)](#)
 - [AI Futures Project](#) (high-level plans)
 - [Intelligence Security Laboratories](#)
 - Verifiable Compute Foundation (getting started right now I think)
 - Some of the frontier AI companies (not sure how public they are about this)

- Field coordination / funding, maybe some research
 - [FLI](#)
 - [CAIS](#)
 - [Coefficient Giving](#)
 - [ARIA](#)
 - [SEF](#)
- Honorable mentions (not sure if there's ongoing work)
 - [GovAI](#)
 - [IAPS](#)
 - [CNAS](#)
 - Anthropic (1-6 people)
 - OpenAI (0-2 people)
- (There are various small research groups or individuals working on this as well, such as MATS projects and some academics. My guess is these smaller efforts are like 50% of the effort)

Verification: Network tap related things

- [A system overview for near-term, low-trust AI compute verification](#)
- [Fingerprinting All AI Cluster I/O Without Mutually Trusted Processors](#)
- [Research Note: The Fundamentals and Feasibility of Secure Network Taps for Verifying AI Datacenter Use](#)
- [The Tray as a Bandwidth Boundary: Using a DPU as a Bandwidth Limiter](#)
- [Network Taps - A First Test](#)
- [Network Traffic Hashing](#)
- [Verifying LLM Inference to Detect Model Weight Exfiltration](#)
- [DiFR: Inference Verification Despite Nondeterminism](#)
- [Example Schemes for Verifying High-Stakes AI Agreements](#)
- [Memory Wiping - A First Test](#)
- [Bit-Exact AI Inference Verification Without Performance Tradeoffs](#)
-

Verification: Interconnect Limits related things

- See the Network-tap related stuff
- In [Mechanisms to Verify International Agreements About AI Development](#) (from above), read **Read the appendix about [Interconnect Limits](#)**
- [De-risking Interconnect Limits for AI Verification](#)
- [Aggressive Compression Enables LLM Weight Theft](#)
- Distributed training literature (non comprehensive)
 - [DiLoCo: Distributed Low-Communication Training of Language Models](#)
 - [Streaming DiLoCo with overlapping communication: Towards a Distributed Free Lunch](#)
 - [Decoupled DiLoCo for Resilient Distributed Pre-training](#)
 - [SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient](#)

- [Introducing the distributed training interactive simulator](#)
- [INTELLECT-1 Technical Report](#)
- Lit review in Verification report: <https://arxiv.org/pdf/2506.15867#page=126>
- [Covenant-72B: Pre-Training a 72B LLM with Trustless Peers Over-the-Internet](#)
- [Does Distributed Training Undermine Compute Governance?](#)
- [Rigel Base: Reaching Llama-3.2 Quality with <1% of its Compute](#)

Verification: TEEs

- [On TEEs for Privacy-Preserving Monitoring in AI Governance](#)
- [Attestable Audits: Verifiable AI Safety Benchmarks Using Trusted Execution Environments](#)
- [Secure Enclaves for AI Evaluation](#)
- [Vulnerabilities in Partial TEE-Shielded LLM Inference with Precomputed Noise](#)
- [NDAI Agreements](#)
- [Agentic Witnessing: Pragmatic and Scalable TEE-Enabled Privacy-Preserving Auditing](#)
- [PAL*M: Property Attestation for Large Generative Models](#)
- [Guardain: Protecting Emerging Generative AI Workloads on Heterogeneous NPU](#)
- [Cove: Compositional Multi-Party Confidential Workflows for Verifiable AI Governance](#)
- [Roy Rinberg - Inference Verification \(in a TEE\)](#)
- [Enabling Verifiably-Scoped Monitoring through Large Language Models and Trusted Compute](#). Also [Auditor-in-a-Box: Tools for Third-Party Auditing](#)
- [Double Blind Evals: Resolving the Dual Confidentiality Dilemma in AI Safety Auditing](#)

Verification: Telemetry classification

- [Detecting Hidden ML Training With Zero-Overhead Telemetry](#)
- [Wave: Leveraging Architecture Observation for Privacy-Preserving Model Oversight](#)
- [Inspector Agents](#)
- [Workload Identification with Physical Side Channels for AI Governance](#)
- (Lots of other relevant literature)

Verification: Finding compute

- [Tracking Hyperscale AI Data Center Growth with Satellite Imagery](#)
- [Detecting Compute Structuring in AI Governance is likely feasible](#)
-

How to do research well

 [How To Do Research Well Reading List \[Public, Shareable\]](#)