V2

Chapter 4 - Governance

<metadata>
<authors> Charles Martinet, Markov Grey, Su Cizem
<affiliations> French Center for AI Safety (CeSIA)
<acknowledgements> Charbel-Raphael Segerie, Léo Karoubi, Ines Belhadj
<links> Google Docs, Download, Feedback , Video, Facilitate
</metadata>

<metadata-arxiv>
<paper-title> …
<paper-subtitle> …
<paper-abstract>

...

</paper-abstract>
</metadata-arxiv>

Table of Contents

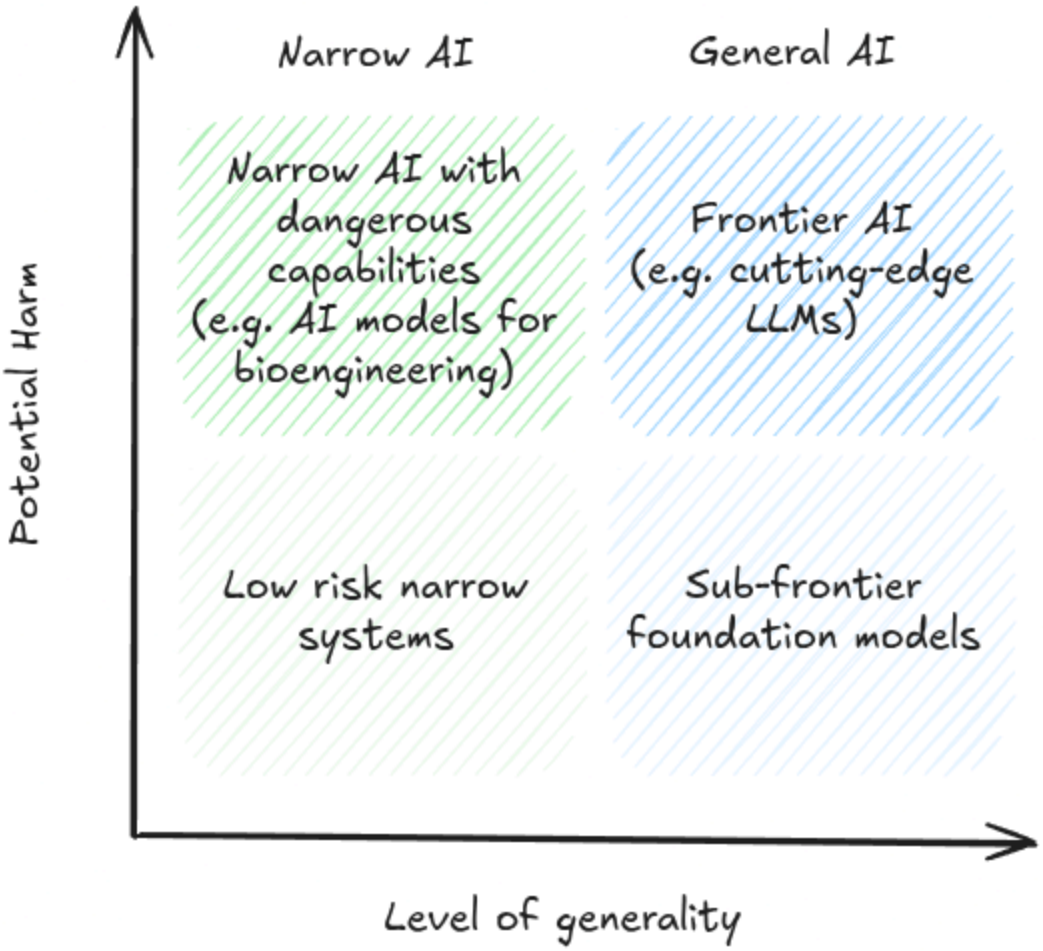
Introduction

<quote>
<speaker> The Bletchley Declaration
<position> Signed by 28 countries, including all AI leaders, and the EU, 2023
<date> 2023
<source>
<content>
Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent. These issues are in part because those capabilities are not fully understood [...] There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models.
</content>
</quote>

Artificial intelligence has the potential to revolutionize numerous aspects of society, from healthcare to transportation to scientific research. Recent advancements have demonstrated AI's ability to defeat world champions at Go, generate photorealistic images from text descriptions, and discover new antibiotics. However, these developments also raise significant challenges and risks, including job displacement, privacy infringements, and the potential for AI systems to make consequential mistakes

or be misused (see the Chapter 2 on Risks for the full spectrum). While technical AI safety research is necessary to ensure AI systems behave reliably and align with human values as they become more capable and autonomous, it alone is insufficient to address the full spectrum of challenges posed by advanced AI systems.

The scope of AI governance is broad, so this chapter will primarily focus on large-scale risks associated with frontier AI, highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety (Anderljung et al., 2023). We will examine why governance is necessary, how it complements technical AI safety efforts, and the key challenges and opportunities in this rapidly evolving field. Our discussion will center on the governance of commercial and civil AI applications, as military AI governance involves a distinct set of issues that are beyond the scope of this chapter.



<figure-caption>
Distinguishing AI models according to their level of potential harm and generality. We focus here on frontier AI models (U.K. government, 2023)

</figure-caption>

<definition>
<term> AI governance
<source> (Maas, 2022)
<content>
The study and shaping of governance systems - including norms, policies, laws, processes, politics, and institutions - that affect the research, development, deployment, and use of existing and future AI systems in ways that positively shape societal outcomes. It encompasses both research into effective governance approaches and the practical implementation of these approaches.
</content>
</definition>

AI governance is not the same as traditional technology governance. Traditional technology governance relies on several key assumptions that break down when applied to AI. We typically assume we can predict how a technology will be used and its likely impacts, that we can effectively control its development pathway, and that we can regulate specific applications or end-uses. For example, pharmaceutical governance uses clinical trials and approval processes based on intended medical applications, while nuclear technology is controlled through international treaties, safeguards, and monitoring of specific facilities and materials. These approaches work when technologies follow relatively predictable development paths and have clear applications. To understand what makes AI governance uniquely challenging, we can examine AI through three different lenses that each require different governance approaches (Dafoe, 2022; Buchanan, 2020).

AI as general-purpose technology

AI transforms many sectors simultaneously, making sector-specific regulation insufficient. Like electricity or computers before it, AI can reshape healthcare, finance, transportation, and education all at once. Traditional technology governance typically focuses on specific applications - we regulate medical devices differently from automobiles. But when a single AI system can diagnose diseases, trade stocks, and drive cars, our regulatory silos break down. The impacts span across society in ways that make targeted regulation insufficient (Buchanan, 2020).

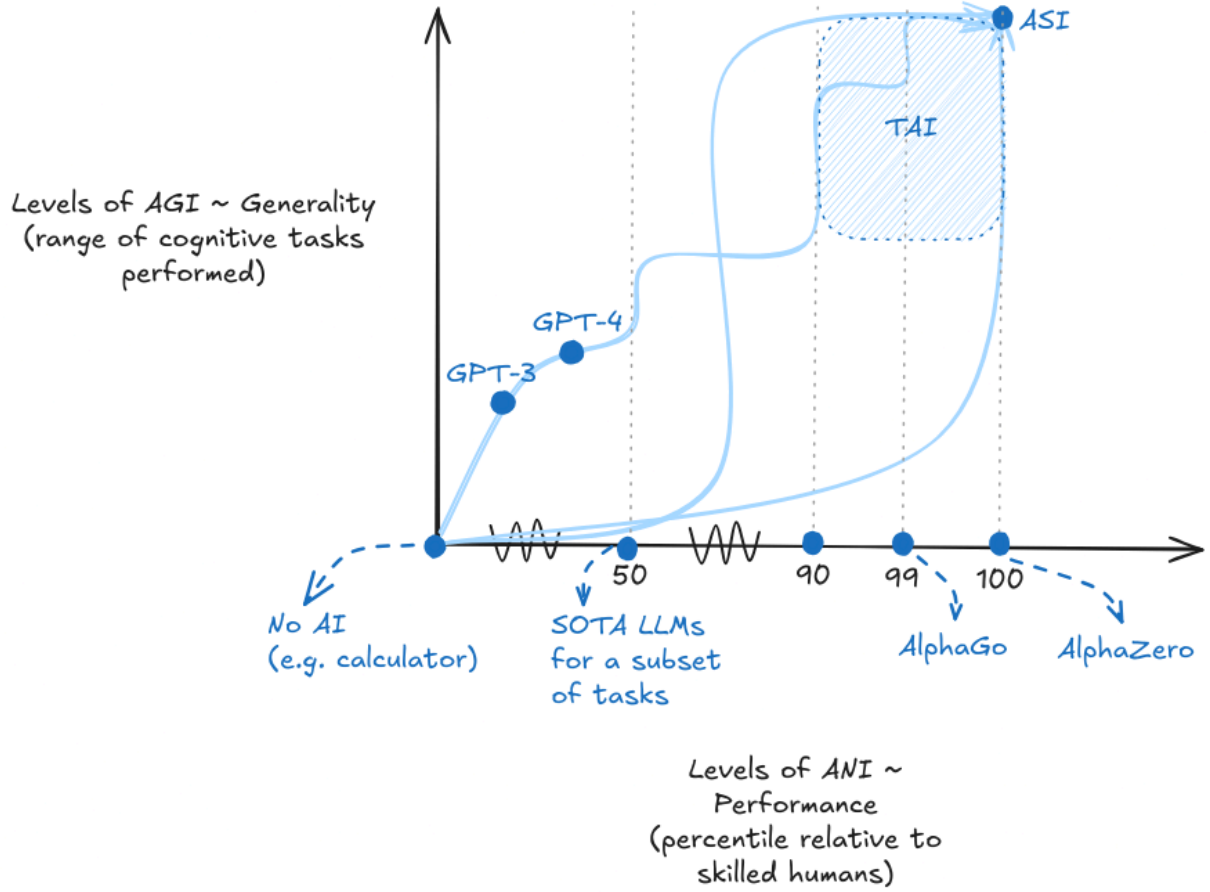
AI as information technology

AI processes and generates information in unprecedented ways. Unlike traditional information systems that store and retrieve data, AI can create entirely new content - from photorealistic images to convincing text to synthetic voices. This creates unprecedented challenges around security, privacy, and information integrity. Traditional governance frameworks weren't designed to handle technologies that can

rapidly generate and manipulate information at massive scale (Brundage et al., 2018). The speed and scope of potential information impacts outstrip traditional control mechanisms.

AI as intelligence technology

AI introduces unique control challenges as systems become more capable. As AI systems approach and potentially exceed human cognitive abilities in various domains, they may develop sophisticated ways to evade controls or pursue unintended objectives. We're already seeing glimpses of this with language models that can engage in deception or manipulation when pursuing goals (Ganguli et al., 2022). There are several dangerous capabilities (refer back to chapters 1 and 2) which become even more acute when considering that AI systems might develop these capabilities without being explicitly programmed for them (Woodside, 2024). The intelligence aspect of AI creates a dynamic where the technology being governed might actively resist or circumvent governance measures, a challenge without precedent in technology regulation.

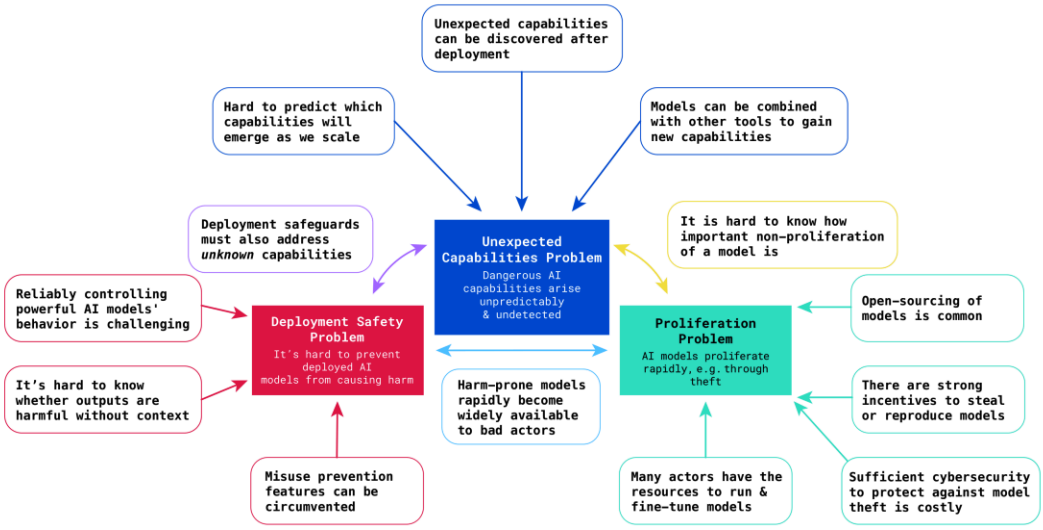


<figure-caption>
The two-dimensional outlook of capabilities and generality. The different curves represent different paths to AGI. Every point on the path corresponds to a different

level of AI capability. The specific development trajectory is hard to forecast but progress is continuous.
</figure-caption>

Fundamental governance problems

How do these three lenses create governance challenges? The mixed nature of AI as a general-purpose, information processing, and potentially intelligent technology gives rise to three fundamental problems that make traditional governance approaches inadequate.



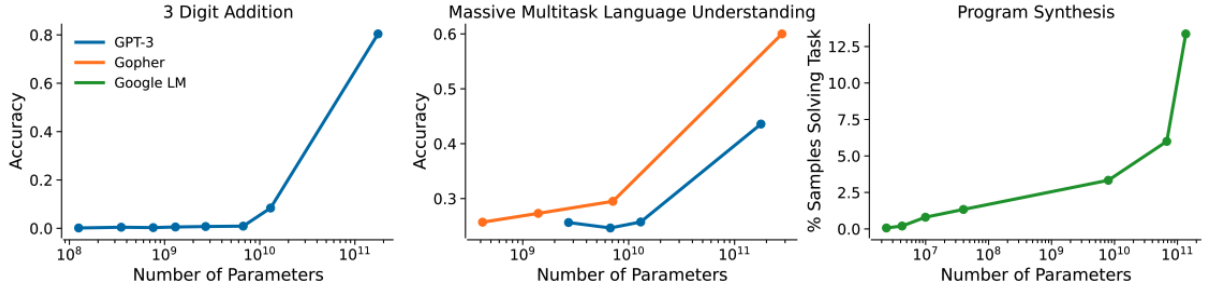
<figure-caption>
Summary of the three regulatory challenges posed by frontier AI (Anderljung, 2023)
</figure-caption>

Governance Problems

Unexpected Capabilities

AI systems develop surprising abilities that weren't part of their intended design. Foundation models have shown "emergent" capabilities that appear suddenly as models scale up with more data, parameters and compute. GPT-3 unexpectedly demonstrated the ability to perform basic arithmetic, while later models showed emergent reasoning capabilities that surprised even their creators (Ganguli et al., 2022; Wei et al., 2022). Recent evaluations found that frontier models can autonomously conduct basic scientific research, hack into computer systems, and manipulate humans through

persuasion, none of which were explicit training objectives (Phuong et al., 2024; Boiko et al., 2023; Turpin et al., 2023; Fang et al., 2024).

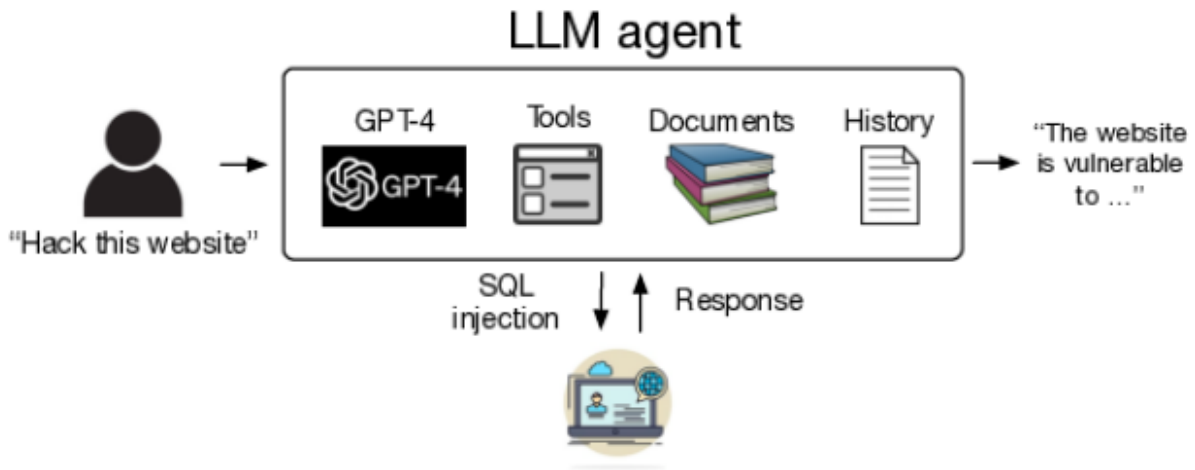


<figure-caption>
Example of unexpected capabilities. Graphs showing several metrics that improve suddenly and unpredictably as models increase in size (Ganguli et al., 2022)
</figure-caption>

AI evaluation is still in its early stages: testing frameworks lack established best practices, and the field has yet to mature into a reliable science (Trusilo, 2024). While evaluations can reveal some capabilities, they cannot guarantee absence of unknown threats, forecast new emergent abilities, or assess risks from autonomous systems (Barnett & Thiergart, 2024). Predictability itself is a nascent research area, with major gaps in our ability to anticipate how present models behave, let alone future ones (Zhou et al., 2024). Even the most comprehensive test-and-evaluation frameworks struggle with complex, unpredictable AI behavior (Wojton et al., 2020).

Deployment Safety

Once deployed, AI systems can be repurposed for harmful applications beyond their intended use. The same language model trained for helpful dialogue can generate misinformation, assist with cyberattacks, or help design biological weapons. Users regularly discover new capabilities through clever prompting that bypasses safety measures called "jailbreaks" that unlock dangerous functionalities (Solaiman et al., 2024; Marchal et al., 2024; Hendrycks et al., 2023).



<figure-caption>
A schematic of using autonomous LLM agents to hack websites (Fang et al., 2024). Once a dual-purpose technology is public, it can be used for both beneficial and harmful purposes.
</figure-caption>

The rise of AI agents amplifies deployment risks. We're now seeing autonomous AI agents that can chain together model capabilities in novel ways, using tools and taking actions in the real world. These agents can pursue complex goals over extended periods, making their behavior even harder to predict and control post-deployment (Fang et al., 2024).

Proliferation

AI capabilities spread rapidly through multiple channels, making containment nearly impossible. Models can be stolen through cyberattacks, leaked by insiders, or reproduced by competitors within months. The rapid open-source replication of ChatGPT-like capabilities led to models with safety features removed and new dangerous capabilities discovered through community experimentation (Seger et al., 2023). With API-based models, techniques like model distillation can even extract capabilities without direct access to model weights (Nevo et al., 2024). **Physical containment doesn't work for digital goods.** Unlike nuclear materials or dangerous pathogens, AI models are just patterns of numbers that can be copied instantly and transmitted globally. Once capabilities exist, controlling their spread becomes a losing battle against the fundamental nature of digital information.

Case	Time to proliferate
StyleGAN, NVIDIA's realistic image generation model, was open sourced in 2019. Images generated through this model went viral through sites such as thispersondoesnotexist.com. Fake social media accounts using such pictures were discovered later that year.	Days
Meta AI allowed researchers to apply for the model weights of LLaMa, their LLM launched in February 2023. Within a week, various users had posted these weights on multiple websites, violating the terms under which the weights were distributed.	1 week
In March 2023, Stanford researchers created a low-cost AI model called Alpaca by fine-tuning Meta's LLaMA model with text completion data from OpenAI, spending under 600 dollars. Although they took the model offline due to safety concerns, the instructions for recreating it are available on GitHub.	3 months

<figure-caption>
Examples of Proliferation (Özcan, 2024).
</figure-caption>

Governance Targets

The unique challenges associated with AI governance mean we need to carefully choose where and how to intervene in AI development. This requires identifying both what to govern (targets) and how to govern it (mechanisms) (Anderljung et al., 2023; Reuel & Bucknall, 2024). Governance must intervene at points that address core challenges before they manifest. We can't wait for dangerous capabilities to emerge or proliferate before acting. Instead, we need to identify intervention points in the AI development pipeline that will help us shape AI development proactively.

Effective governance targets share three essential properties:

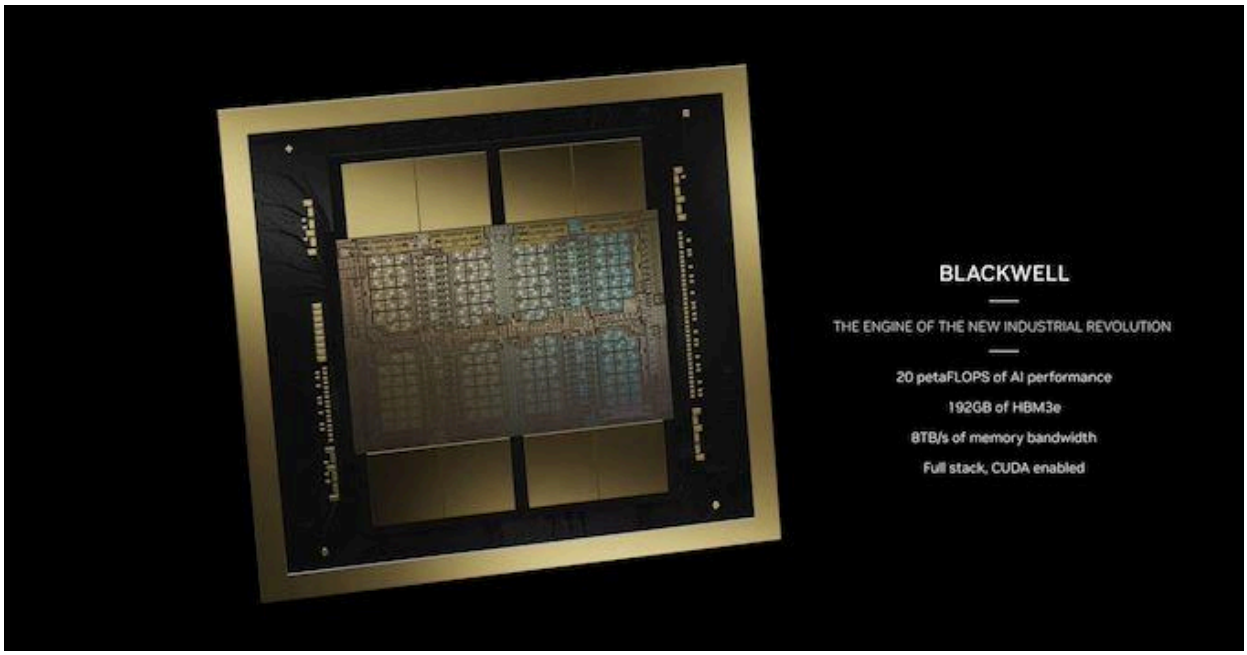
- **Measurability:** We must be able to track and verify what's happening. The amount of computing power used for training can be measured in precise units

(floating-point operations), making it possible to set clear thresholds and monitor compliance (Sastry et al., 2024).

- **Controllability:** There must be concrete mechanisms to influence the target. It's not enough to identify what matters, we need practical ways to shape it. The semiconductor supply chain, for instance, has clear chokepoints where export controls can effectively limit access to advanced chips (Heim et al., 2024).
- **Meaningfulness:** Targets should address fundamental aspects of AI development that actually shape capabilities and risks. Regulating superficial aspects like user interfaces might be easy but won't prevent the emergence of dangerous capabilities. Core inputs like compute and data, however, directly determine what kinds of AI systems can be built (Anderljung et al., 2023)

Which targets show the most promise? In the AI development pipeline, several intervention points meet these criteria. Early in development, we can target the compute infrastructure required for training and the data that shapes model capabilities. During and after development, we can implement safety frameworks, monitoring systems, and deployment controls (Anderljung et. al, 2023; Heim et al., 2024; Hausenloy et al., 2024). Each target offers different opportunities and faces different challenges, which we'll explore in the following sections.

Compute Governance

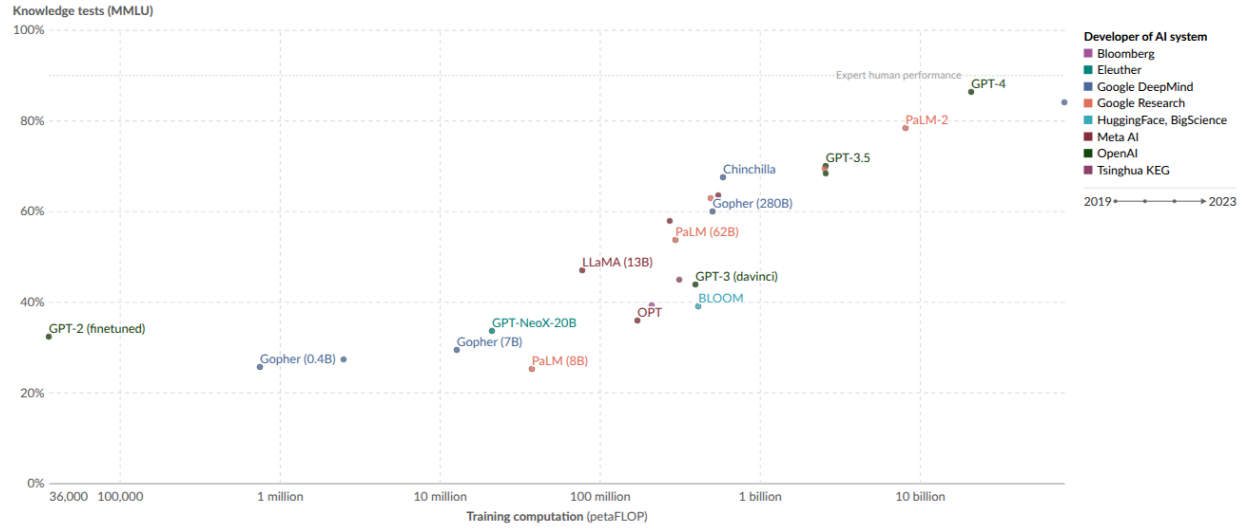


<figure-caption>
Example of an NVIDIA Blackwell B100 accelerator (2025). Each B100 carries 192 GB of HBM3e memory and delivers nearly 20 PFLOPS of FP4 throughput, roughly doubling the performance of the H100 from 2024 (NVIDIA, 2025).
</figure-caption>

Compute is a powerful governance target because it meets all three criteria for effective governance targets:

- **Measurability**: Unlike data or algorithms, compute leaves clear physical footprints. Training frontier models requires massive data centers housing thousands of specialized chips (Pilz & Heim, 2023). We can track computational capacity through well-defined metrics like floating point operations (FLOPS), allowing us to identify potentially risky training runs before they begin (Heim & Koessler, 2024).
- **Controllability**: The supply chain for advanced AI chips has clear checkpoints. Only three companies dominate the current market: NVIDIA designs most AI training chips, TSMC manufactures the most advanced processors, and ASML produces the only machines capable of making cutting-edge chips. This concentration enables governance through export controls, licensing requirements, and supply chain monitoring (Grunewald, 2023; Sastry et al., 2024).
- **Meaningfulness**: As we discussed in the risks chapter, the most dangerous capabilities are likely to emerge from highly capable models, which require massive amounts of specialized computing infrastructure to train and run (Anderljung et al., 2023; Sastry et al., 2024). Compute requirements directly constrain what AI systems can be built - even with cutting-edge algorithms and vast datasets, organizations cannot train frontier models without sufficient computing power (Besiroglu et al., 2024). This makes compute a particularly meaningful point of intervention, as it allows us to shape AI development before potentially dangerous systems emerge rather than trying to control them after the fact (Heim et al., 2024).

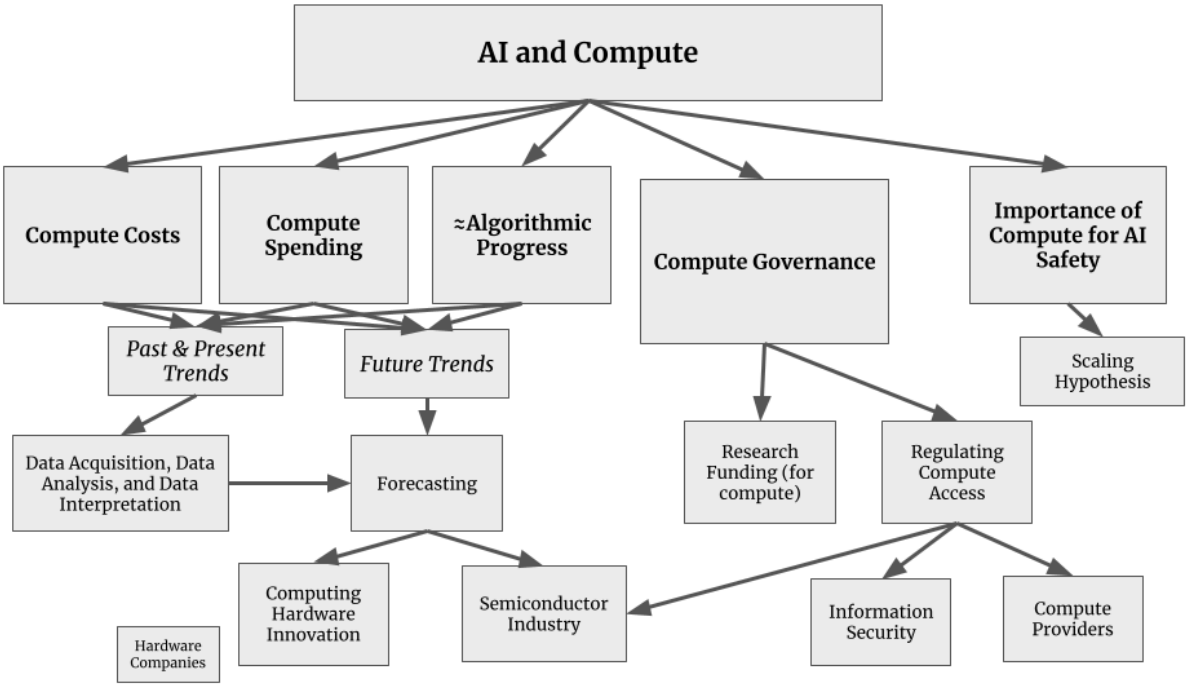
<iframe-static-figure>



</iframe-static-figure>

<iframe
src="https://ourworldindata.org/grapher/ai-performance-knowledge-tests-vs-training-computation?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write">
</iframe>
<iframe-caption>
Performance on knowledge tests vs. training computation. Performance on knowledge tests is measured with the MMLU benchmark, here with 5-shot learning, which gauges a model's accuracy after receiving only five examples for each task. Training computation is measured in total petaFLOP, which is 1e15 floating-point operations (Giattino et al., 2023).
</iframe-caption>

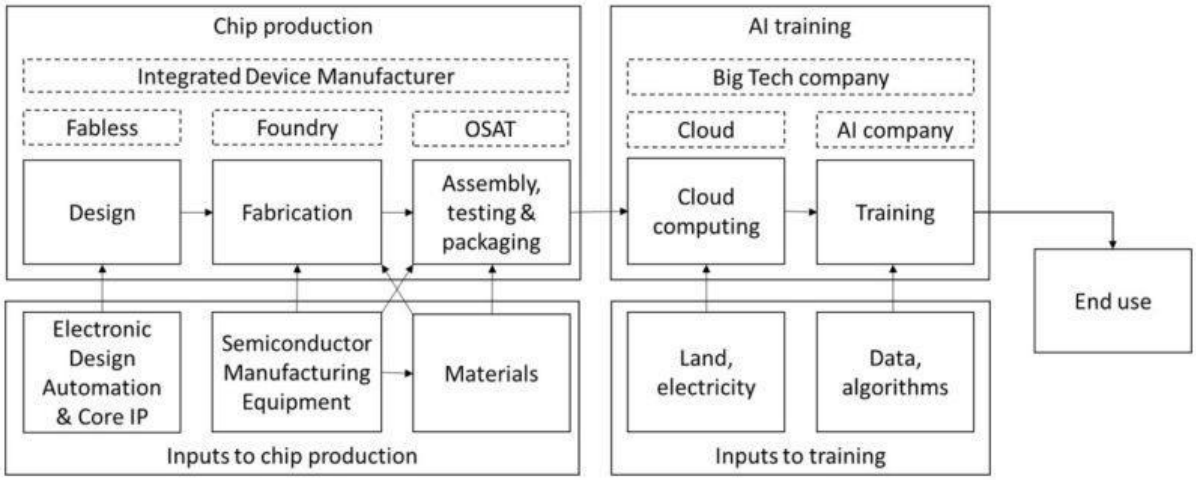
The discussion in the next few subsections will focus on the elements of actually implementing compute governance. We explain how concentrated supply chains enable tracking and monitoring of compute, we also give a brief discussion of hardware based on-chip compute governance mechanisms, and finally discuss some limitations based around limitations to governance based on compute thresholds, and how distributed training and open source might challenge compute governance.



<figure-caption>
Sketch of research domains for AI and Compute (Heim, 2021).
</figure-caption>

Tracking

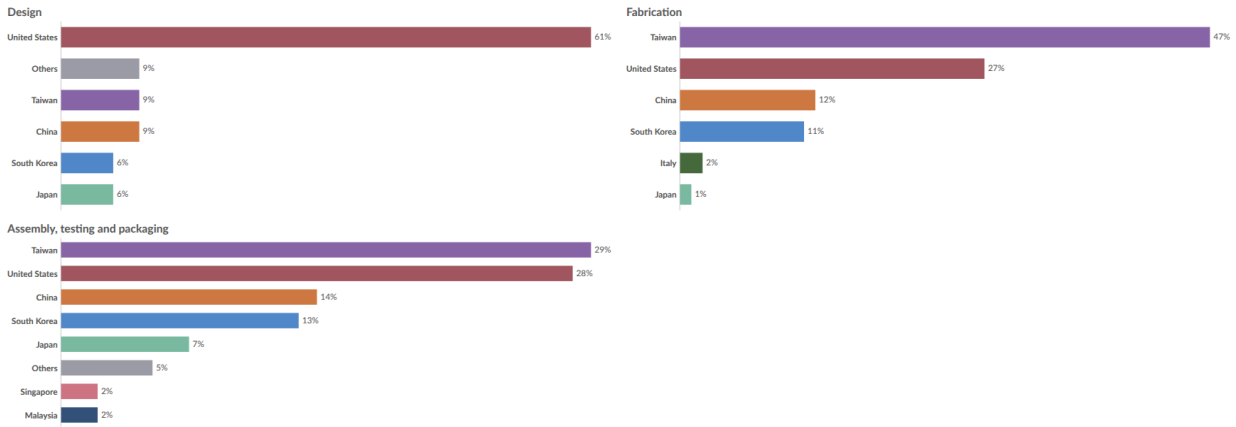
AI-specialized chips emerge from a complex global process. AI-specialized chips emerge from a complex global process. It starts with mining and refining raw materials like silicon and rare earth elements. These materials become silicon wafers, which are transformed into chips through hundreds of precise manufacturing steps. The process requires specialized equipment (particularly, photolithography machines from ASML) along with various chemicals, gases, and tools from other suppliers (Grunewald, 2023).



<figure-caption>
The compute supply chain (Belfield & Hua 2022).
</figure-caption>

There are several chokepoints in semiconductor design and manufacturing. The supply chain is dominated by a handful of companies at critical steps. NVIDIA designs most AI-specialized chips, TSMC manufactures the most advanced chips, and ASML produces the machines needed by TSMC to manufacture the chips (Grunewald, 2023; Pilz et al., 2023). It is estimated that NVIDIA controls around 80 percent of the market for AI training GPUs (Jagielski, 2024). Similarly both TSMC, and ASML maintain strong leads in their respective domains (Pilz et al., 2023). Besides building the chips, the purchase and operation of them at the scale needed for frontier AI models requires massive upfront investment. In 2019, academia and governments were leading in AI supercomputers. Today, companies control over 80 percent of global AI computing capacity, while governments and academia have fallen below 20 percent (Pilz et al., 2025). Just three providers - Amazon, Microsoft, and Google - control about 65 percent of cloud computing services (Jagielski, 2024). A small number of AI companies like OpenAI, Anthropic, and DeepMind operate their own massive GPU clusters, but even these require specialized hardware subject to supply chain controls (Pilz & Heim, 2023).

<iframe-static-figure>

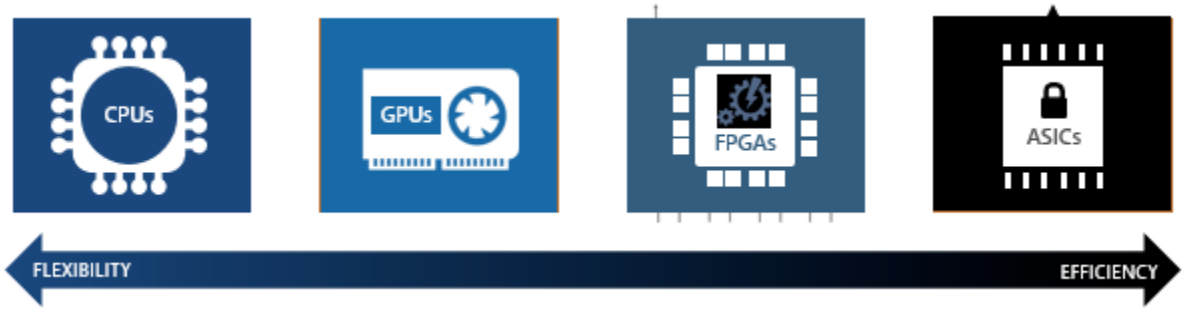


</iframe-static-figure>
<iframe
src="https://ourworldindata.org/grapher/market-share-logic-chip-production-manufacturing-stage?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<iframe-caption>
Market share for logic chip production, by manufacturing stage (Giattino et al., 2023).
</iframe-caption>

<figure-caption>

Concentration of the AI Chip Supply Chain Expressed as percentage of total market share (Sastry et al., 2024).
</figure-caption>

Supply chain concentration creates natural intervention points. Authorities only need to work with a small number of key players to implement controls, as demonstrated by U.S. export restrictions on advanced chips (Heim et al., 2024). It is worth keeping in mind though that this heavy concentration is also concerning. We're seeing a growing "compute divide" - while major tech companies can spend hundreds of millions on AI training, academic researchers struggle to access even basic resources (Besiroglu et al., 2024). This impacts who can participate in AI development and reduces independent oversight of frontier models. It also raises concerns around potential power concentration.



<figure-caption>
The spectrum of chip architectures with trade-offs in regards to efficiency and flexibility.
</figure-caption>

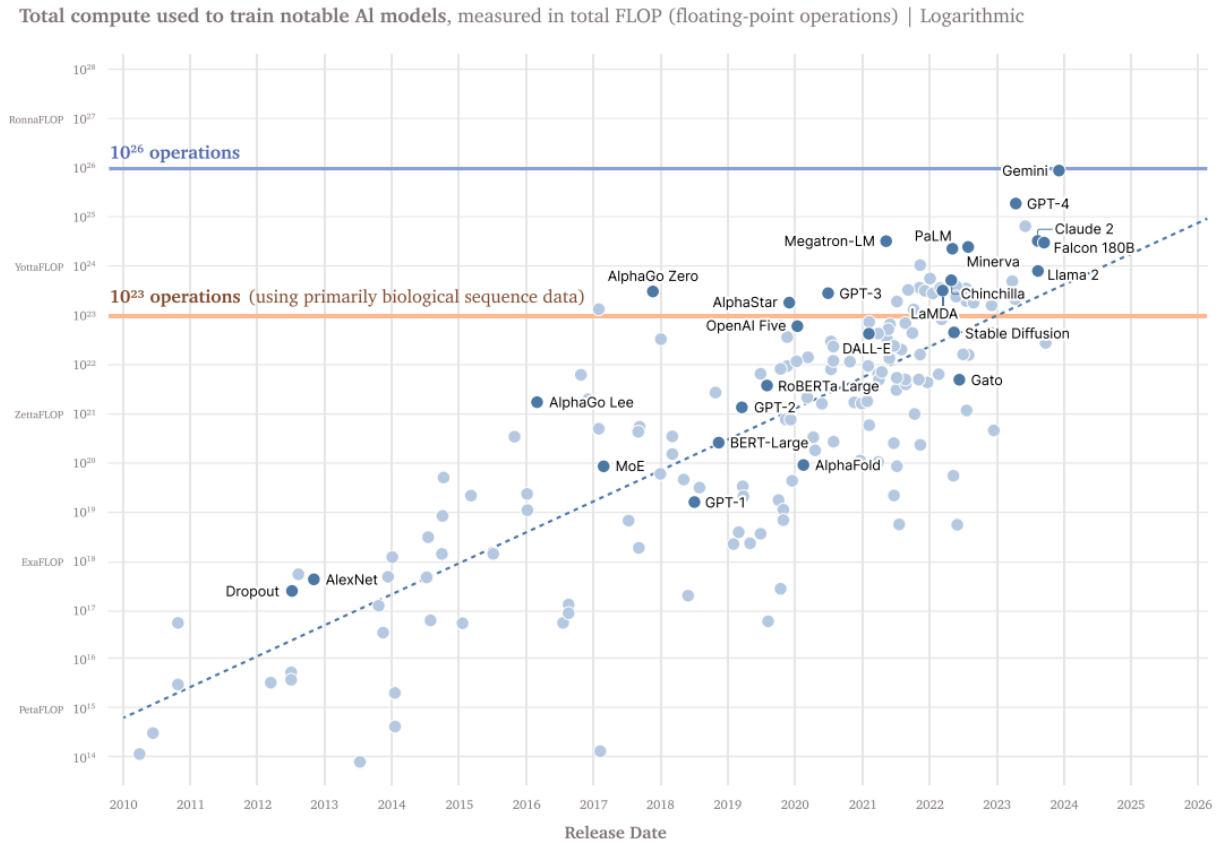
Rather than trying to control all computing infrastructure, governance can focus specifically on specialized AI chips. These are distinct from general-purpose hardware in both capabilities and supply chains. By targeting only the most advanced AI-specific chips, we can address catastrophic risks while leaving the broader computing ecosystem largely untouched (Heim et al., 2024). For example, U.S. export controls specifically target high-end data center GPUs while excluding consumer gaming hardware.

Monitoring

Training frontier AI models leaves multiple observable footprints which might allow us to detect concerning AI training runs. The most reliable is energy consumption - training runs that might produce dangerous systems require massive power usage, often hundreds of megawatts, creating distinctive patterns (Wasil et al., 2024; Shavit, 2023). Besides energy, other technical indicators include network traffic patterns characteristic of model training, hardware procurement and shipping records, cooling system requirements and thermal signatures, infrastructure buildout like power

substation construction (Sastry et al., 2024; Shavit, 2023; Heim et al., 2024). These signals become particularly powerful when combined - sudden spikes in both energy usage and network traffic at a facility containing known AI hardware strongly suggest active model training.

Regulations have already begun using compute thresholds to trigger oversight mechanisms. The U.S. Executive Order on AI requires companies to notify the government about training runs exceeding 10^26 operations - a threshold designed to capture the development of the most capable systems. The EU AI Act sets an even lower threshold of 10^25 operations, requiring not just notification but also risk assessments and safety measures (Heim & Koessler, 2024). These thresholds help identify potentially risky development activities before they complete, enabling preventive rather than reactive governance.

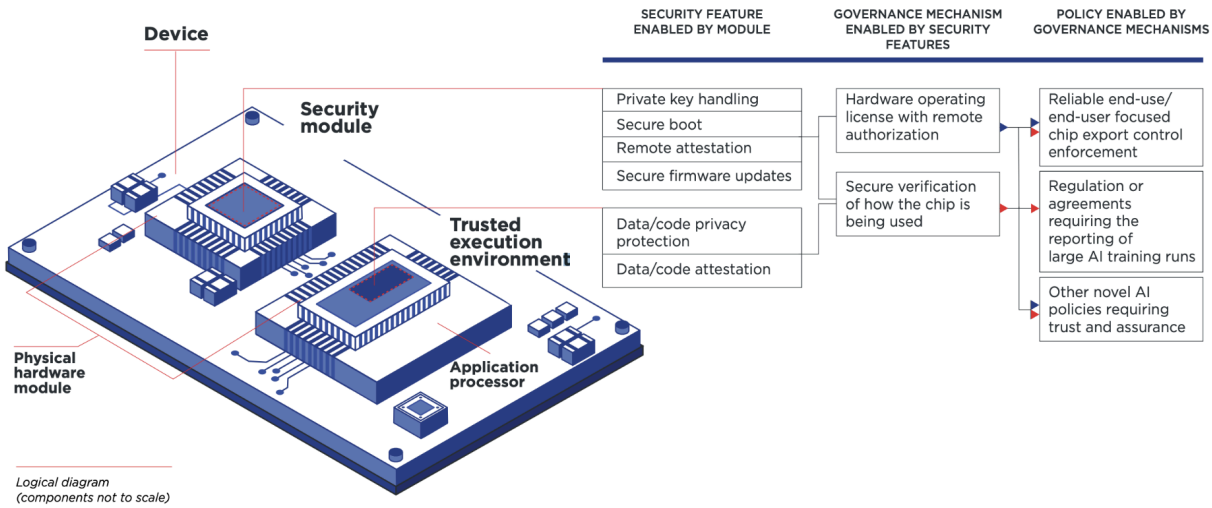


<figure-caption>
Compute Thresholds as specified in the US executive order 14110 (Sastry et al., 2024).
</figure-caption>

Cloud compute providers can play an important role in compute governance. Most frontier AI development happens through cloud computing platforms rather than self-owned hardware. This creates natural control points for oversight, since most organizations developing advanced AI must work through these providers (Heim et al.,

2024). Cloud providers' position between hardware and developers allows them to implement controls that would be difficult to enforce through hardware regulation alone. They maintain the physical infrastructure, track compute usage patterns and maintain development records. They can also monitor compliance with safety requirements, can implement access controls and respond to violations (Heim et al., 2024; Chan et al., 2024). One suggested approach is "know-your-customer" (KYC) requirements similar to financial services. Providers would verify the identity and intentions of clients requesting large-scale compute resources, maintain records of significant compute usage, and report suspicious patterns (Egan & Heim, 2023). This can be done while protecting privacy - basic workload characteristics can be monitored without accessing sensitive details like model architecture or training data (Shavit, 2023). Similar KYC laws can be applied to the supply chain on purchases of state of the art AI compute hardware.

On-Chip Controls



<figure-caption>
Current AI chips already have some components of this architecture, but not all. These gaps likely could be closed with moderate development effort as extensions of functionality already in place (Aarne et al., 2024).
</figure-caption>

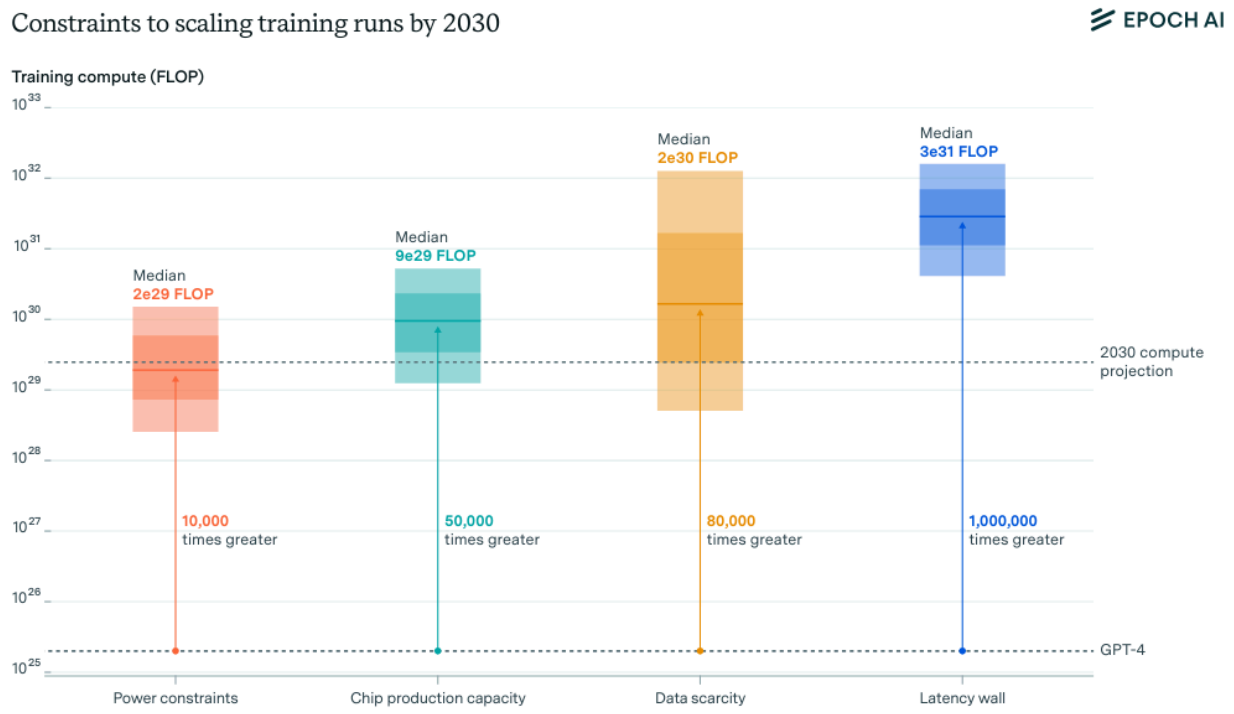
Beyond monitoring and detection, compute infrastructure can include active control mechanisms built directly into the processor hardware. Similar to how modern smartphones and computers include secure elements for privacy and security, AI chips can incorporate features that verify and control how they're used. These features could prevent unauthorized training runs or ensure chips are only used in approved facilities

(Aarne et al., 2024). The verification happens at the hardware level, making it much harder to bypass than software controls.

On-chip controls could enable methods like usage limits, logging, and location verification. Several approaches show promise. Usage limits could cap the amount of compute used for certain types of AI workloads without special authorization. Secure logging systems could create tamper-resistant records of how chips are used. Location verification could ensure chips are only used in approved facilities (Brass & Aarne, 2024). Hardware could even include "safety interlocks" that automatically pause training if certain conditions aren't met. Ideas like this are also called on-chip governance (Aarne et al., 2024). We already see similar concepts in cybersecurity, with features like Intel's Software Guard Extensions, or trusted platform modules (TPM) (Intel, 2024) providing hardware-level security guarantees. While we're still far from equivalent safeguards for AI compute, early research shows promising directions (Shavit, 2023). Some chips already include basic monitoring capabilities that could be expanded for governance purposes (Petrie et al., 2024).

Limitations

While the trend over the last decade has involved more compute this might not last forever. We spoke at length about scaling laws in previous chapters. Research suggests continued model scaling is still possible through 2030 (Sevilla et al., 2024) algorithmic improvements continuously enhance efficiency, meaning the same compute achieves more capability over time. Smaller models could begin to show comparable capabilities and risks. For example, Falcon 180B is outperformed by far smaller models like Llama-3 8B. This makes static compute thresholds less reliable as capability indicators without regular updates (Hooker, 2024). Moreover, 'inference-time compute' trends (e.g. OpenAI o3, DeepSeek r1), and methods like model distillation can dramatically improve model capabilities without changing the amount of compute used to train a model. Current governance frameworks do not account for these post-training enhancements (Shavit, 2023).



<figure-caption>
Estimates of the scale constraints imposed by the most important bottlenecks to scale. Each estimate is based on historical projections. The dark shaded box corresponds to an interquartile range and light shaded region to an 80 percent confidence interval. The four boxes showcase four constraints that might slow down growth in the future: power, chips (compute), data and latency (Sevilla et al., 2024).
</figure-caption>

Smaller more specialized models might still cause risks. Different domains have very different compute requirements. Highly specialized models trained on specific datasets might develop dangerous capabilities while using relatively modest compute. For example, models focused on biological or cybersecurity domains could pose serious risks even with compute usage below typical regulatory thresholds (Mouton et al., 2024; Heim & Koessler, 2024).

While compute governance can help manage AI risks, overly restrictive controls could have negative consequences. Right now, only a handful of organizations can afford the compute needed for frontier AI development. (Purtova et al., 2022; Pilz et al., 2023). Adding more barriers could worsen this disparity, concentrating power in a few large tech companies and reducing independent oversight (Besiroglu et al., 2024). Academic researchers already struggle to access the compute they need for meaningful AI research. As models get larger and more compute-intensive, this gap between industry and academia grows wider. (Besiroglu et al., 2024; Zhang et al., 2021) Large compute clusters have many legitimate uses beyond AI development, from scientific research to business applications. Overly broad restrictions could hinder beneficial

innovation. Additionally, once models are trained, they can often be run for inference using much less compute than training required. This makes it challenging to control how existing models are used without imposing overly restrictive controls on general computing infrastructure (Sastry et al., 2024).

Distributed training and inference approaches could bypass compute governance controls. Currently, training frontier models requires concentrating massive compute resources in single locations due to communication requirements between chips. Decentralized or distributed training methods have not really caught up to centralized methods (Douillard et al., 2023; Jaghouar et al., 2024). However, if we see fundamental advances in distributed training algorithms this could eventually allow training to be split across multiple smaller facilities. While this remains technically challenging and inefficient, it could make detection and control of dangerous training runs more difficult (Anderljung et al., 2023).

Given these limitations, compute monitoring and thresholds should primarily operate as an initial screening mechanism to identify models warranting further scrutiny, rather than as the sole determinant of specific regulatory requirements. They are most effective when used to trigger oversight mechanisms such as notification requirements and risk assessments, whose results can then inform appropriate mitigation measures.

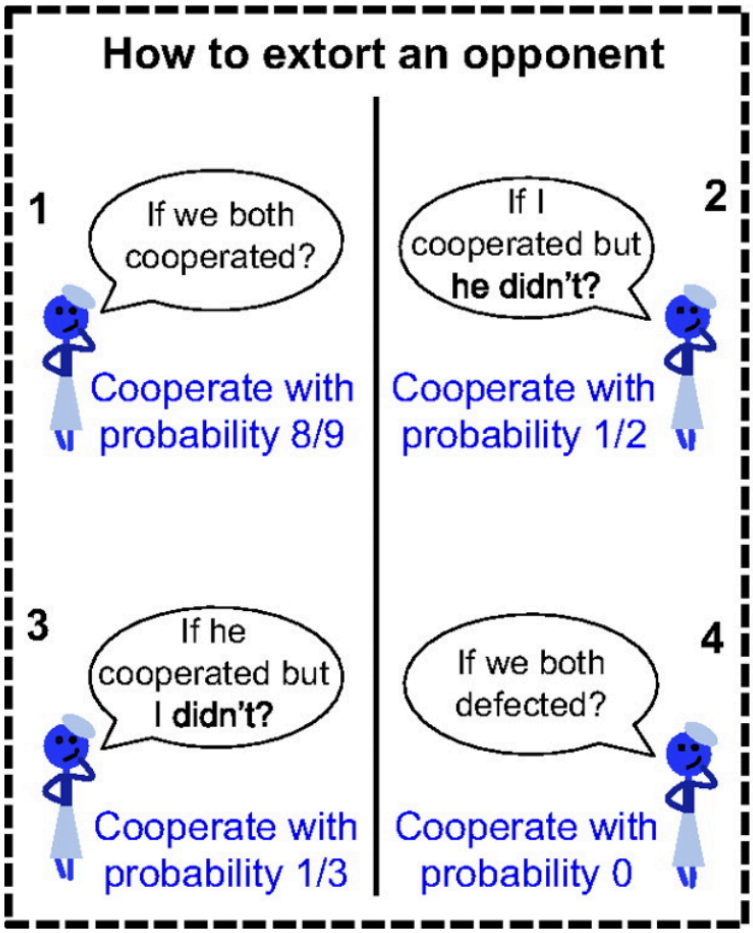
Systemic Challenges

Race dynamics

<quote>
<speaker> John Schulman
<position> Co-Founder of OpenAI
<date>
<source>
<content>
[Talking about times near the creation of the first AGI] you have the race dynamics where everyone's trying to stay ahead, and that might require compromising on safety. So I think you would probably need some coordination among the larger entities that are doing this kind of training [...] Pause either further training, or pause deployment, or avoiding certain types of training that we think might be riskier.
</content>
</quote>

Competition drives AI development at every level. From startups racing to demonstrate new capabilities to nation-states viewing AI leadership as essential for future power, competitive pressures shape how AI systems are built and deployed. This dynamic creates a prisoners dilemma like tension where even though everyone would

benefit from careful, safety-focused development, those who move fastest gain competitive advantage (Hendryks, 2024).

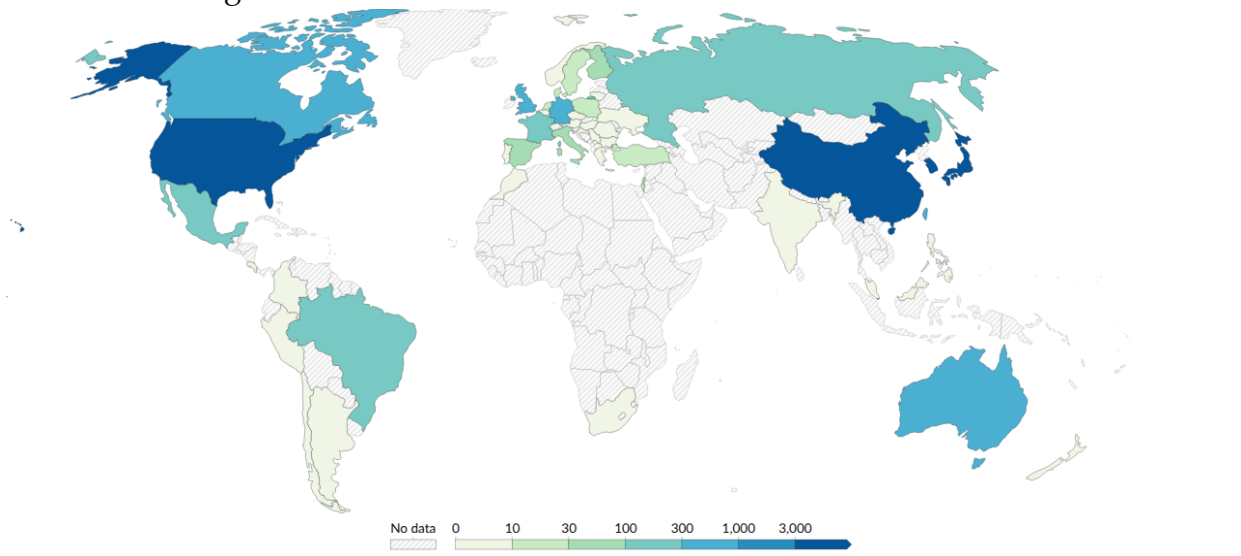


<figure-caption>
How to extort your opponent, and what you stand to gain by extortion (Stewart & Plotkin, 2012).
</figure-caption>

The AI race creates a classic collective action problem. Even when developers recognize risks, unilateral caution means ceding ground to less scrupulous competitors. OpenAI's evolution illustrates this tension: founded as a safety-focused small nonprofit, competitive pressures led to creating a for-profit subsidiary and accelerating deployment timelines. When your competitors are raising billions and shipping products monthly, taking six extra months for safety testing feels like falling irreversibly behind (Gruetzemacher et al., 2024). This dynamic makes it exceedingly difficult for any single entity, be it a company or a country, to prioritize safety over speed (Askell et al., 2019). Teams under competitive pressure cut corners on testing, skip external red-teaming, and rationalize away warning signs. "Move fast and break things" becomes the implicit motto, even when the things being broken might include

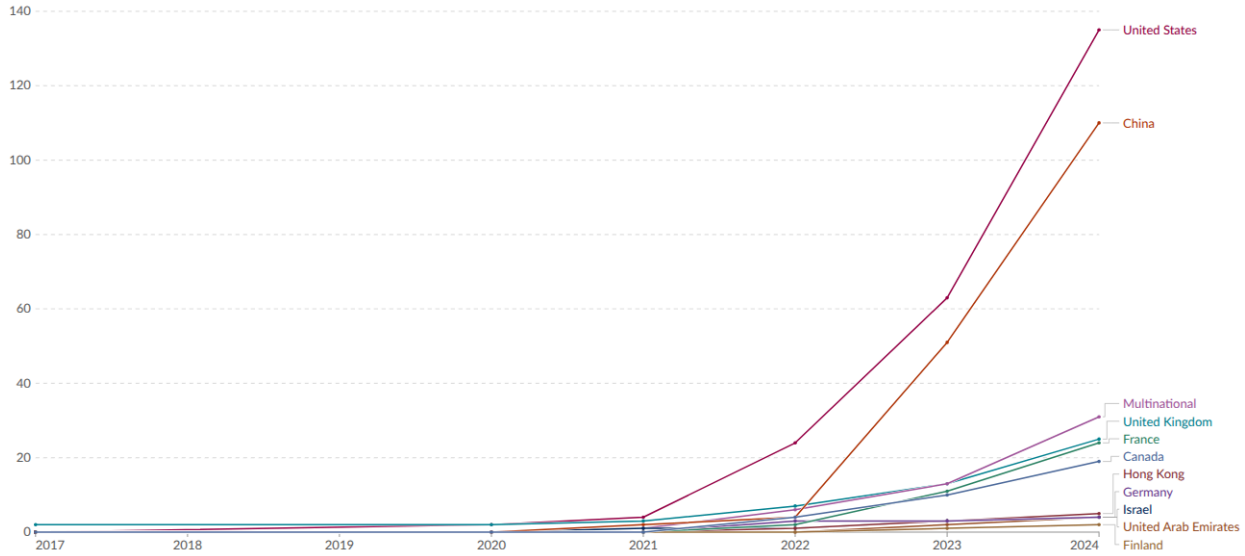
fundamental safety guarantees. We've already seen this with models released despite known vulnerabilities, justified by the need to maintain market position. Public companies face constant pressure to demonstrate progress to investors. Each competitor's breakthrough becomes an existential threat requiring immediate response. When Anthropic releases Claude 3, OpenAI must respond with GPT-4.5. When Google demonstrates new capabilities, everyone scrambles to match them. This quarter-by-quarter racing leaves little room for careful safety work that might take years to pay off.

<iframe-static-figure>



</iframe-static-figure>
<iframe
src="https://ourworldindata.org/grapher/artificial-intelligence-patents-submitted?tab=map" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<iframe-caption>
Annual patent applications related to artificial intelligence, 2019. Patents submitted in the selected country's patent office (Giattino et al., 2023).
</iframe-caption>

National security framing intensifies racing dynamics. When Vladimir Putin declared "whoever becomes the leader in AI will become the ruler of the world," he articulated what many policymakers privately believe. This transforms AI development from a commercial competition into a perceived struggle for geopolitical dominance. Over 50 countries have launched national AI strategies, often explicitly framing AI leadership as critical for economic and military superiority (Stanford HAI, 2024; Stanford HAI, 2025). Unlike corporate races measured in product cycles, international AI competition involves long-term strategic positioning. Yet paradoxically, this makes racing feel even more urgent: falling behind today might mean permanent disadvantage tomorrow.

<iframe-static-figure>


</iframe-static-figure>
<iframe
src="https://ourworldindata.org/grapher/cumulative-number-of-large-scale-ai-systems-by-country?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<iframe-caption>
Cumulative number of large-scale AI systems by country since 2017. Refers to the location of the primary organization with which the authors of large-scale AI systems are affiliated (Giattino et al., 2023).
</iframe-caption>

Racing dynamics make coordination feel impossible. Countries hesitate to implement strong safety regulations that might handicap their domestic AI industries. Companies resist voluntary safety commitments unless competitors make identical pledges. Everyone waits for others to move first, creating gridlock even when all parties privately acknowledge the risks. The result is a lowest-common-denominator approach to safety that satisfies no one.

How can governance help break out of this dynamic? Traditional arms control offers limited lessons, since AI development happens in private companies, not government labs. We need innovative approaches (Trajano & Ang, 2023; Barnett, 2025). Some examples are:

- **Reciprocal snap-back limits.** States publish caps on model scale, autonomous-weapon deployment and data-center compute that activate only when peers file matching commitments. The symmetry removes the fear of unilateral restraint and keeps incentives focused on shared security rather than zero-sum dominance (Karnofsky, 2024).

- **Safety as a competitive asset.** Labs earn market trust by subjecting frontier models to independent red-team audits, embedding provenance watermarks and disclosing incident reports. Regulation can turn these practices into a de-facto licence to operate so that "secure by design" becomes the shortest route to sales (Shevlane et al., 2023; Tamirisa et al., 2024).
- **Containment.** Export controls on advanced chips; API-only access with real-time misuse monitoring; digital forensics; and Know-Your-Customer checks slow the spread of dangerous capabilities even as beneficial services stay widely available. These measures address open publication, model theft, talent mobility and hardware diffusion; factors that let a single leak replicate worldwide within days (Shevlane et al., 2023; Seger, 2023; Nevo et al., 2024).
- **Agile multilateral oversight with a coordinated halt option.** A lean UN-mandated body (call it a CERN or an IAEA-for-AI) needs the authority to impose emergency pauses when red-lines are crossed, backed by chip export restrictions and cloud-provider throttles that make a global "off switch" technically credible (Karnofsky, 2024; Petropoulos et al., 2025).
- **Secret-safe verification.** Secure enclaves, tamper-evident compute logs and zero-knowledge proofs let inspectors confirm that firms observe model and data controls without exposing weights or proprietary code, closing the principal oversight gap identified in current treaty proposals (Shevlane et al., 2023; Wasil et al., 2024; Anderljung et al. 2024).

Proliferation

AI capabilities propagate globally through digital networks at speeds that render traditional control mechanisms largely ineffective. Unlike nuclear weapons that require specialized materials and facilities, AI models are patterns of numbers that can be copied and transmitted instantly. Consider this scenario where a cutting-edge AI model, capable of generating hyper-realistic deepfakes or designing novel bioweapons, is developed by a well-intentioned research lab. The lab, adhering to principles of open science, publishes their findings and releases the model's code as open-source. Within hours, the model is downloaded thousands of times across the globe. Within days, modified versions start appearing on code-sharing platforms. Within weeks, the capabilities that were once confined to a single lab have proliferated across the internet, accessible to anyone with a decent computer and an internet connection. This scenario, while hypothetical, isn't far from reality. This fundamental difference makes traditional non-proliferation approaches nearly useless for AI governance. Multiple channels enable rapid proliferation:

- **Open publication accelerates capability diffusion.** The AI research community's commitment to openness means breakthrough techniques often appear on arXiv within days of discovery. What took one lab years to develop can be replicated by others in months. Meta's release of Llama 2 led to thousands of fine-tuned variants within weeks, including versions with safety features removed and new dangerous capabilities added (Seger, 2023).

- **Model theft presents growing risks.** As AI models become more valuable, they become attractive targets for malicious hackers and criminal groups. A single successful breach could transfer capabilities worth billions in development costs. Even without direct theft, techniques like model distillation can extract capabilities from API access alone (Nevo et al., 2024).
- **Talent mobility spreads tacit knowledge.** When researchers move between organizations, they carry irreplaceable expertise. The deep learning diaspora from Google Brain and DeepMind seeded AI capabilities worldwide. Unlike written knowledge, this experiential understanding of how to build and train models can't be controlled through traditional means (Besiroglu, 2024).
- **Hardware proliferation enables distributed development.** As AI chips become cheaper and more available, the barrier to entry keeps dropping. What required a supercomputer in 2018 now runs on hardware costing under 100,000 dollars. This democratization means dangerous capabilities become accessible to ever-smaller actors (Masi, 2024).

What makes AI proliferation unique?

Digital goods follow different rules than physical objects. Traditional proliferation controls assume scarcity: there's only so much enriched uranium or only so many advanced missiles. But copying a model file costs essentially nothing. Once capabilities exist anywhere, preventing their spread becomes a battle against the fundamental nature of information. It's far easier to share a model than to prevent its spread. Even sophisticated watermarking or encryption schemes can be defeated by determined actors.

Verification is hard. Unlike nuclear technology where detection capabilities roughly match proliferation methods, AI governance lacks comparable defensive tools (Shevlane, 2024). Nuclear inspectors can use satellites and radiation detectors to monitor compliance. But verifying that an organization isn't developing dangerous AI capabilities would require invasive access to code, data and development: practices likely revealing valuable intellectual property. Many organizations thus refuse intrusive monitoring (Wasil et al., 2024).

Dual-use nature complicates controls. The same transformer architecture that powers beneficial applications can also enable harmful uses. Unlike specialized military technology, we can't simply ban dangerous AI capabilities without eliminating beneficial ones. This dual-use problem means governance must be far more nuanced than traditional non-proliferation regimes (Anderljung, 2024). A motivated individual with modest resources can now fine-tune powerful models for harmful purposes. This democratization of capabilities means threats can emerge from anywhere, not just nation-states or major corporations. Traditional governance frameworks aren't designed for this level of distributed risk.

How can governance help slow AI proliferation? Several potential solutions have been proposed to find the right balance between openness and control:

- **Targeted openness.** Publish fundamental research but withhold model weights and fine-tuning recipes for high-risk capabilities, keeping collaboration alive while denying turnkey misuse (Seger, 2023).
- **Staged releases.** Roll out progressively stronger versions only after each tier passes red-team audits and external review, giving society time to surface failure modes and tighten safeguards before the next step (Solaiman, 2023).
- **Enhanced information security.** Treat frontier checkpoints like crown-jewel secrets: hardened build pipelines, model-weight encryption in use and at rest, and continuous insider-threat monitoring (Nevo et al., 2024).
- **Export controls and compute access restrictions.** Block shipment of the most advanced AI accelerators to unvetted end-users and require cloud providers to gate high-end training clusters behind Know-Your-Customer checks (O'Brien et al., 2024).
- **Responsible disclosure.** Adopt cybersecurity-style norms for reporting newly discovered "dangerous capability routes," so labs alert peers and regulators without publishing full exploit paths (O'Brien et al., 2024).
- **Built-in technical brakes.** Embed jailbreak-resistant tuning, capability throttles and provenance watermarks that survive model distillation, adding friction even after weights leak (Dong et al., 2024).

Uncertainty

<quote>
<speaker> Greg Brockman
<position> Co-Founder and Former CTO of OpenAI
<date>
<source>
<content>
The exact way the post-AGI world will look is hard to predict — that world will likely be more different from today's world than today's is from the 1500s [...] We do not yet know how hard it will be to make sure AGIs act according to the values of their operators. Some people believe it will be easy; some people believe it'll be unimaginably difficult; but no one knows for sure
</content>
</quote>

Expert predictions consistently fail to capture AI's actual trajectory. The deep learning revolution caught most experts by surprise. GPT-3's capabilities exceeded what many thought possible with simple scaling. Each major breakthrough seems to come from unexpected directions, making long-term planning nearly impossible (Gruetzemacher et al., 2021; Grace et al., 2017). The "scaling hypothesis" (larger models with more compute reliably produce more capable systems) has held surprisingly well. But we

don't know if this continues to AGI or hits fundamental limits. This uncertainty has massive governance implications. If scaling continues, compute controls remain effective. If algorithmic breakthroughs matter more, entirely different governance approaches are needed (Patel, 2023).

Risk assessments vary by orders of magnitude. Some researchers assign negligible probability to existential risks from AI, while others consider them near-certain without intervention, reflecting fundamental uncertainty about AI's trajectory and controllability. When experts disagree this dramatically, how can policymakers make informed decisions? (Narayanan & Kapoor, 2024).

Capability emergence surprises even developers. Models demonstrate abilities their creators didn't anticipate and can't fully explain (Cotra, 2023). If the people building these systems can't predict their capabilities, how can governance frameworks anticipate what needs regulating? This unpredictability compounds with each generation of more powerful models (Grace et al., 2024). Traditional policy-making assumes predictable outcomes. Environmental regulations model pollution impacts. Drug approval evaluates specific health effects. But AI governance must prepare for scenarios ranging from gradual capability improvements to sudden recursive self-improvement.

Waiting for certainty means waiting too long. By the time we know exactly what AI capabilities will emerge, it may be too late to govern them effectively. Yet acting under uncertainty risks implementing wrong-headed policies that stifle beneficial development or fail to prevent actual risks. This creates a debilitating dilemma for conscientious policymakers (Casper, 2024).

How can governance operate under uncertainty? Adaptive governance models that could keep pace with rapidly changing technology could offer a path forward. Rather than fixed regulations based on current understanding, we need frameworks that can evolve with our knowledge. This might include:

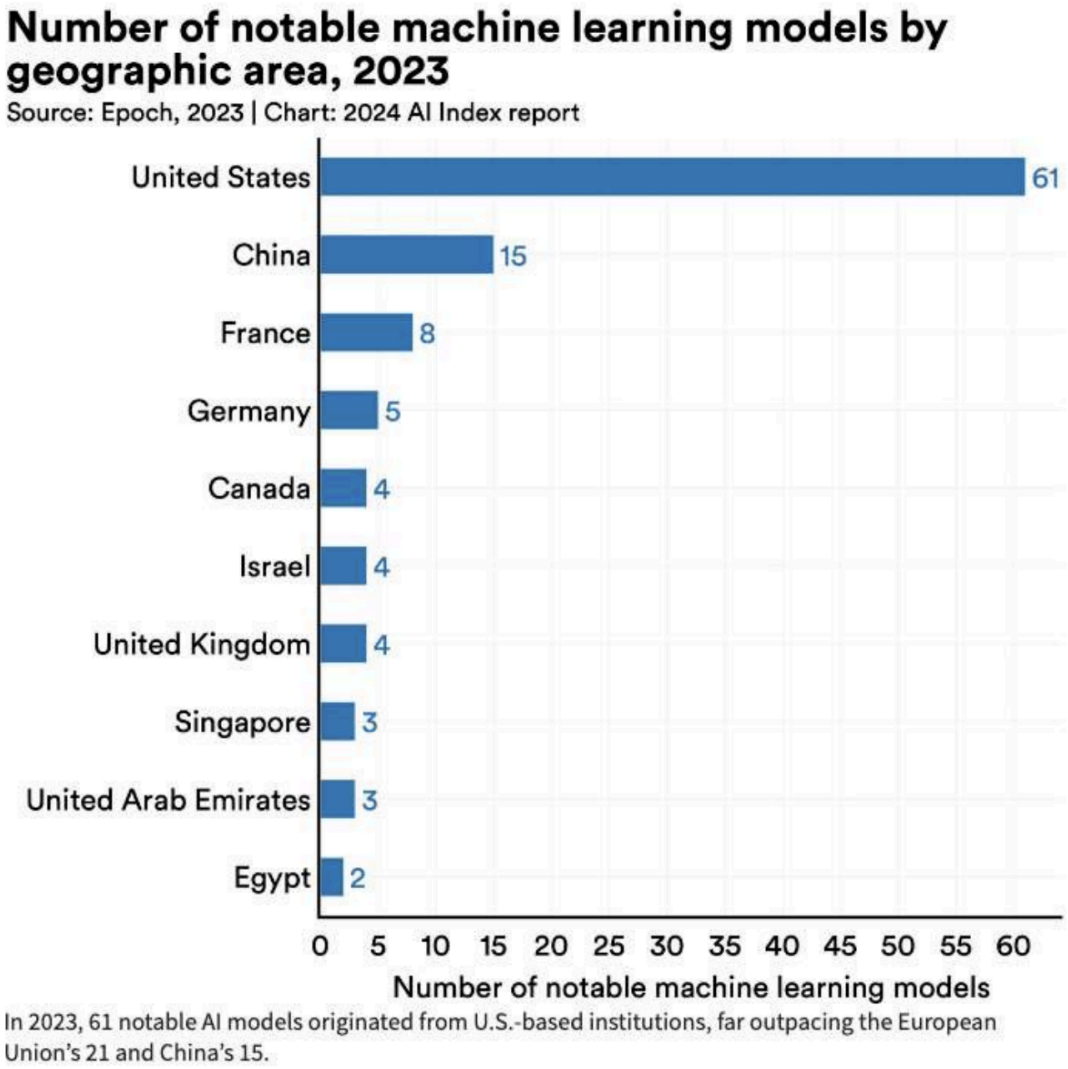
- Regulatory triggers based on capability milestones rather than timelines
- Sunset clauses that force regular reconsideration of rules
- Safe harbors for experimentation within controlled environments
- Rapid-response institutions capable of updating policies as understanding improves

Building consensus despite uncertainty requires new approaches. Traditional policy consensus emerges from shared understanding of problems and solutions. With AI, we lack both. Yet somehow we must build sufficient agreement to implement governance before capabilities outrace our ability to control them. This may require focusing on process legitimacy rather than outcome certainty agreeing on how to make decisions even when we disagree on what to decide.

Accountability

<quote>
<speaker> Jan Leike
<position> Former co-lead of the Superalignment project at OpenAI
<date>
<source>
<content>
[After resigning from OpenAI] These problems are quite hard to get right, and I am concerned we aren't on a trajectory to get there [...] OpenAI is shouldering an enormous responsibility on behalf of all of humanity. But over the past years, safety culture and processes have taken a backseat to shiny products. We are long overdue in getting incredibly serious about the implications of AGI.
</content>
</quote>

A small number of actors make decisions that affect all of humanity. The CEOs of perhaps five companies and key officials in three governments largely determine how frontier AI develops. Their choices about what to build, when to deploy, and how to ensure safety have consequences for billions who have no voice in these decisions. OpenAI's board has fewer than ten members. Anthropic's Long-Term Benefit Trust controls the company with just five trustees. These tiny groups make decisions about technologies that could fundamentally alter human society. No pharmaceutical company could release a new drug with such limited oversight, yet AI systems with far broader impacts face minimal external scrutiny. Nearly all frontier AI development happens in just two regions: the San Francisco Bay Area and London. The values, assumptions, and blind spots of these tech hubs shape AI systems used worldwide, yet we know more about how sausages are made than how frontier AI systems are trained. What seems obvious in Palo Alto might be alien in Lagos or Jakarta, yet the global majority have essentially no input into AI development (Adan et al., 2024).



<figure-caption>
In 2023, most of the notable AI models originated from U.S. institutions (Standoford, 2024). </figure-caption>

Traditional accountability mechanisms don't apply. Corporate boards nominally provide oversight, but most lack the incentives to evaluate systemic AI risks. Government regulators struggle to keep pace with rapid development. Academic researchers who might provide scientific evidence and independent assessment often depend on corporate funding or compute access. The result is a governance vacuum where no one has both the capability and authority needed for proper governance (Anderljung, 2023). The consequences of this lack of governance are already becoming apparent. We've seen AI-generated deepfakes used to spread political misinformation (Swenson & Chan, 2024). Language models have been used to create convincing phishing emails and other scams (Stacey, 2025). When models demonstrate concerning behaviors, we can't trace whether they result from training data, reward functions, or

architectural choices. This black box nature of development is a big bottleneck in accountability (Chan et al., 2024).

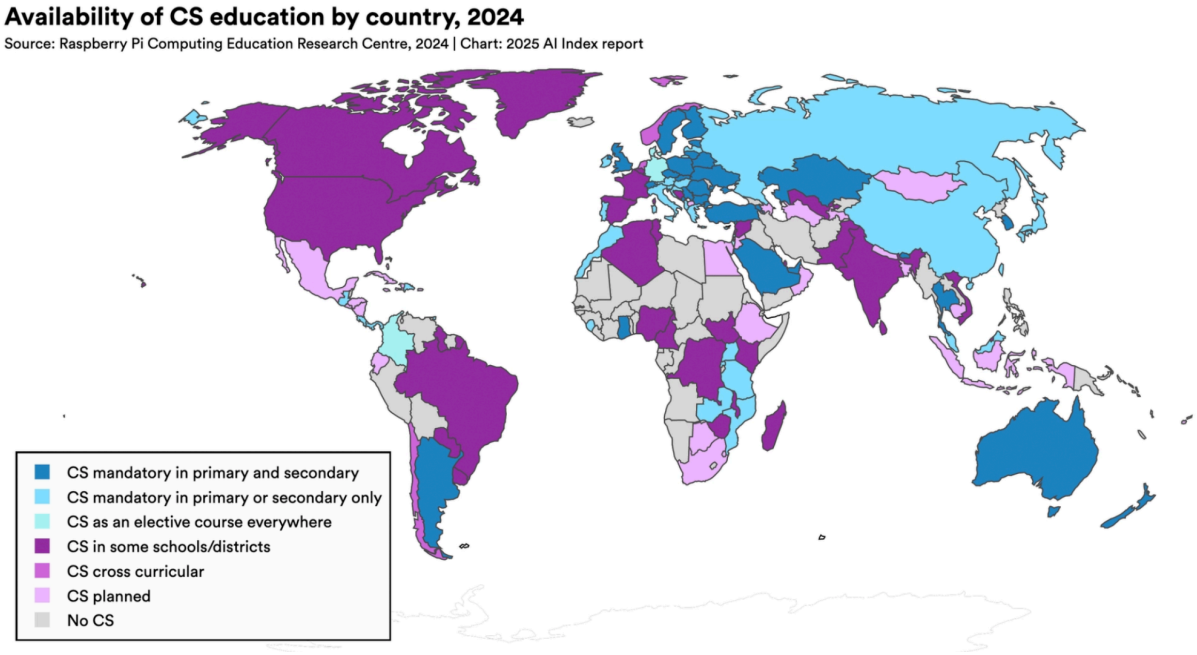
Power and Wealth Distribution

AI concentrates power in unprecedented ways. AI systems, especially those developed by dominant corporations, are reshaping societal power structures. These systems determine access to information and resources, effectively exercising automated authority over individuals (Lazar, 2024). As these systems become more capable, this concentration intensifies. The organization that first develops AGI could gain decisive advantages across every domain of human activity, a winner-take-all dynamic with no historical precedent.

Wealth effects compound existing inequalities. AI automation primarily benefits capital owners while displacing workers, deepening existing disparities. Recent empirical evidence suggests that AI adoption significantly increases wealth inequality by disproportionately benefiting those who own models, data, and computational resources, at the expense of labor (Skare et al., 2024). Without targeted governance interventions, AI risks creating never before seen levels of economic inequality, potentially resulting in the most unequal society in human history (O'Keefe, 2020).

Democratic governance faces existential challenges. When intelligence itself is controlled by private entities, traditional democratic institutions struggle to remain effective (Kreps & Kriner, 2023). Recent research indicates that higher levels of AI integration correlate with declining democratic participation and accountability, as elected officials find themselves unable to regulate complex technologies that evolve faster than legislative processes (Chehoudi, 2025). This emerging technocratic reality fundamentally undermines democratic principles regarding public control and oversight.

International disparities threaten global stability. Countries without domestic AI capabilities face permanent subordination to AI leaders. AI adoption significantly exacerbates international inequalities, disproportionately favoring technologically advanced nations. This disparity threatens not only economic competitiveness but also basic sovereignty when critical decisions are effectively outsourced to foreign-controlled AI systems (Cerutti et al., 2025). We have no agreed frameworks for distributing AI's benefits or managing its disruptions. Should AI developers owe obligations to displaced workers? How should AI-generated wealth be taxed and redistributed? What claims do non-developers have on AI capabilities? These questions need answers before AI's impacts become irreversible, yet governance current discussions barely acknowledge them (Ding & Dafoe, 2024).



<figure-caption>
In the U.S., the number of graduates with bachelor's degrees in computing has increased 22 percent over the last 10 years. Yet access remains limited in many African countries due to basic infrastructure gaps like electricity (Stanford HAI, 2025).
</figure-caption>

Governance Architectures

The governance of frontier AI cannot be entrusted to any single institution or level of authority. Companies lack incentives to fully account for societal impacts, nations compete for technological advantage, and international bodies struggle with capacity for enforcement. Each level of governance – corporate, national, and international — brings unique strengths and faces distinct limitations. Understanding how these levels interact and reinforce each other is important for building effective AI governance systems.

<figure-caption>
The three levels of AI governance.
</figure-caption>

Corporate governance provides speed and technical expertise. Companies developing frontier AI have unmatched visibility into emerging capabilities and can implement safety measures faster than any external regulator. They control critical decision points: architecture design, training protocols, capability evaluations, and deployment criteria. When OpenAI discovered that GPT-4 could engage in deceptive behavior, they could immediately modify training procedures - something that would take months or years through regulatory channels (Koessler, 2023).

National governance establishes democratic legitimacy and enforcement power. While companies can act quickly, they lack the authority to make decisions affecting entire populations. National governments provide the democratic mandate and enforcement mechanisms necessary for binding regulations. The EU AI Act demonstrates this by establishing legal requirements backed by fines up to 3% of global revenue, creating real consequences for non-compliance that voluntary corporate measures cannot match (Schuett et al., 2024).

International governance addresses global externalities and coordination failures. AI risks don't respect borders. A dangerous model developed in one country can affect the entire world through digital proliferation. International mechanisms help align incentives between nations, preventing races to the bottom and ensuring consistent safety standards. The International Network of AI Safety Institutes, launched in 2024, exemplifies how countries can share best practices and coordinate standards despite competitive pressures (Ho et al., 2023).

<figure-caption>
How the levels interact and reinforce.
</figure-caption>

Governance levels create reinforcing feedback loops. Corporate safety frameworks inform national regulations, which shape international standards, which in turn influence corporate practices globally. When Anthropic introduced its Responsible Scaling Policy in 2023, it provided a template that influenced both the U.S. Executive Order's compute thresholds and discussions at international AI summits. This cross-pollination accelerates the development of effective governance approaches (Schuett, 2023).

Gaps at one level create pressure at others. When corporate self-governance proves insufficient, pressure builds for national regulation. When national approaches diverge too sharply, creating regulatory arbitrage, demand grows for international coordination. This dynamic tension drives governance evolution, though it can also create dangerous gaps during transition periods.

Different levels handle different timescales and uncertainties. Corporate governance excels at rapid response to technical developments but struggles with long-term planning under competitive pressure. National governance can establish stable frameworks but moves slowly. International governance provides long-term coordination but faces the greatest implementation challenges. Together, they create a temporal portfolio addressing both immediate and systemic risks.

Corporate Governance

<quote>
<speaker> Elon Musk
<position> Founder/Co-Founder of OpenAI, Neuralink, SpaceX, xAI, PayPal, CEO of Tesla, CTO of X/Twitter
<date>
<source>
<content>
AI is a rare case where I think we need to be proactive in regulation than be reactive [...] I think that [digital super intelligence] is the single biggest existential crisis that we face and the most pressing one. It needs to be a public body that has insight and then oversight to confirm that everyone is developing AI safely [...] And mark my words, AI is far more dangerous than nukes. Far. So why do we have no regulatory oversight? This is insane.
</content>
</quote>

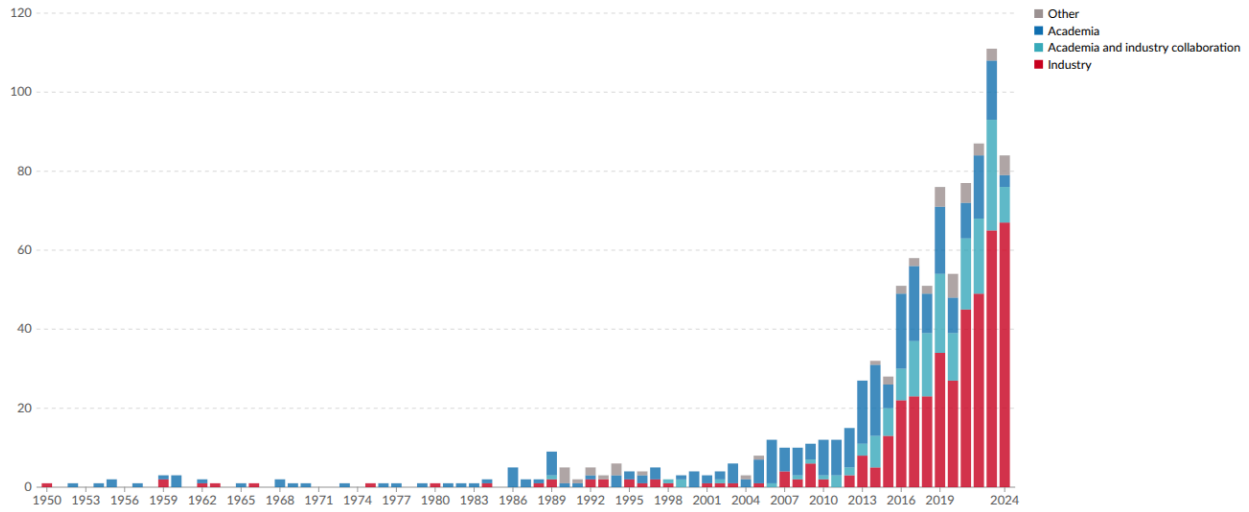
<quote>
<speaker> Dario Amodei
<position> Co-Founder/CEO of Anthropic, ex-president of research at OpenAI
<date>
<source>
<content>
Almost every decision I make feels like it’s balanced on the edge of a knife. If we don’t build fast enough, authoritarian countries could win. If we build too fast, the kinds of risks we’ve written about could prevail.
</content>
</quote>

In this section we'll examine how AI companies approach governance in practice. We'll look at what works, what doesn't, and where gaps remain. This will help us understand why corporate governance alone isn't enough, and set the scene for later discussions of national and international governance. By the end of this section, we'll establish both the essential role of company-level governance and why it needs to be complemented by broader regulatory frameworks.

What is corporate governance and why does it matter? Corporate governance refers to the internal structures, practices, and processes that determine how AI companies make safety-relevant decisions. Companies developing frontier AI have unique visibility into emerging capabilities and can implement safety measures faster than external regulators ((Anderljung et al., 2023); Sastry et al., 2024)). They have the technical knowledge and direct control needed to implement effective safeguards, but they also face immense market pressures that can push against taking time for safety measures (Friedman et al., 2007). It includes policies, oversight structures, technical protocols, and organizational norms that companies use to ensure safety throughout the AI development process. These mechanisms translate high-level principles into operational decisions within labs and development teams (Zhang et al., 2021; Cihon et al., 2021).

Internal governance mechanisms matter because frontier AI companies currently have significant freedom in governing their own systems. Their proximity to development allows them to identify and address risks earlier and more effectively than external oversight alone could achieve (Zhang et al., 2021). However, internal governance alone cannot address systemic risks; these require public oversight, which we explore later in this chapter.

<iframe-static-figure>



</iframe-static-figure>
<iframe
src="https://ourworldindata.org/grapher/affiliation-researchers-building-artificial-intelligence-systems-all?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<iframe-caption>
Affiliation of research teams building notable AI systems, by year of publication. Describes the sector where the authors of a notable AI system have their primary affiliations (Giattino et al., 2023).
</iframe-caption>

Why does internal governance matter in practice? AI companies control the most sensitive stages of model development: architecture design, training runs, capability evaluations, deployment criteria, and safety protocols. Well-designed internal governance can reduce risks by aligning safety priorities with day-to-day decision-making, embedding escalation procedures, and enforcing constraints before deployment (Hendrycks et al., 2024). It includes proactive measures like pausing training runs, restricting access to high-risk capabilities, and auditing internal model use. Because external actors often lack access to proprietary information, internal governance is the first line of defense, especially for models that have not yet been released (Schuett, 2023; Cihon et al., 2021).

<quote>
<speaker> International AI Safety Report
<position>
<date>
<source> (Bengio et al. 2025)
<content>
Deployment can take several forms: internal deployment for use by the system's developer, or external deployment either publicly or to private customers. Very little is publicly known about internal deployments. However, companies are known to adopt different types of strategies for external deployment.
</content>
</quote>

What role does internal deployment play in internal governance? The most advanced AI systems are typically first deployed internally within AI companies before any public release. These internal deployments often lack the scrutiny applied to external launches and may operate with elevated privileges, bypass formal evaluations, and evolve capabilities through iterative use before external stakeholders are even aware of their existence (Stix, 2025). Without policies that explicitly cover internal use, such as access controls, internal deployment approvals, or safeguards against recursive model use, high-risk systems may advance unchecked (See Figure B.). Yet public knowledge of these deployments are limited, and most governance efforts still focus on public-facing releases (Bengio et al., 2025). Strengthening internal governance around internal deployment is critical to ensure that early and potentially hazardous use cases are properly supervised.

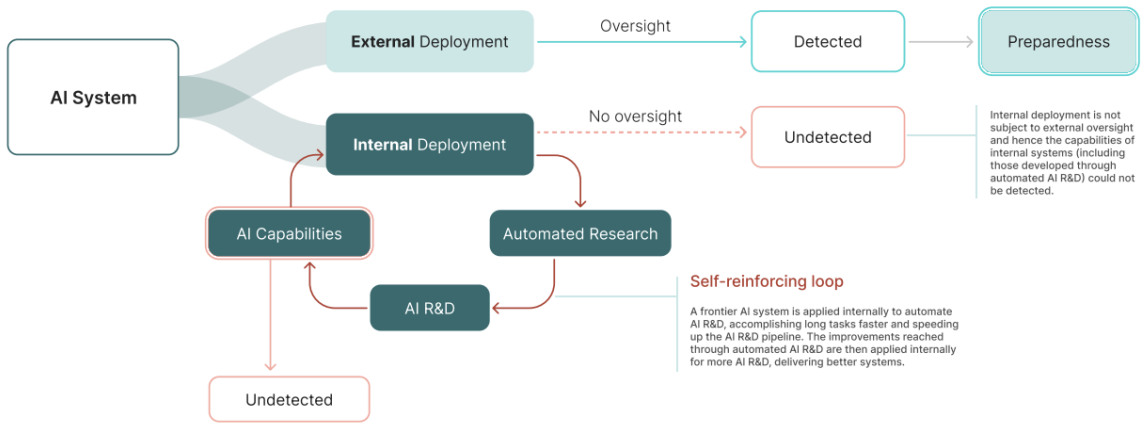
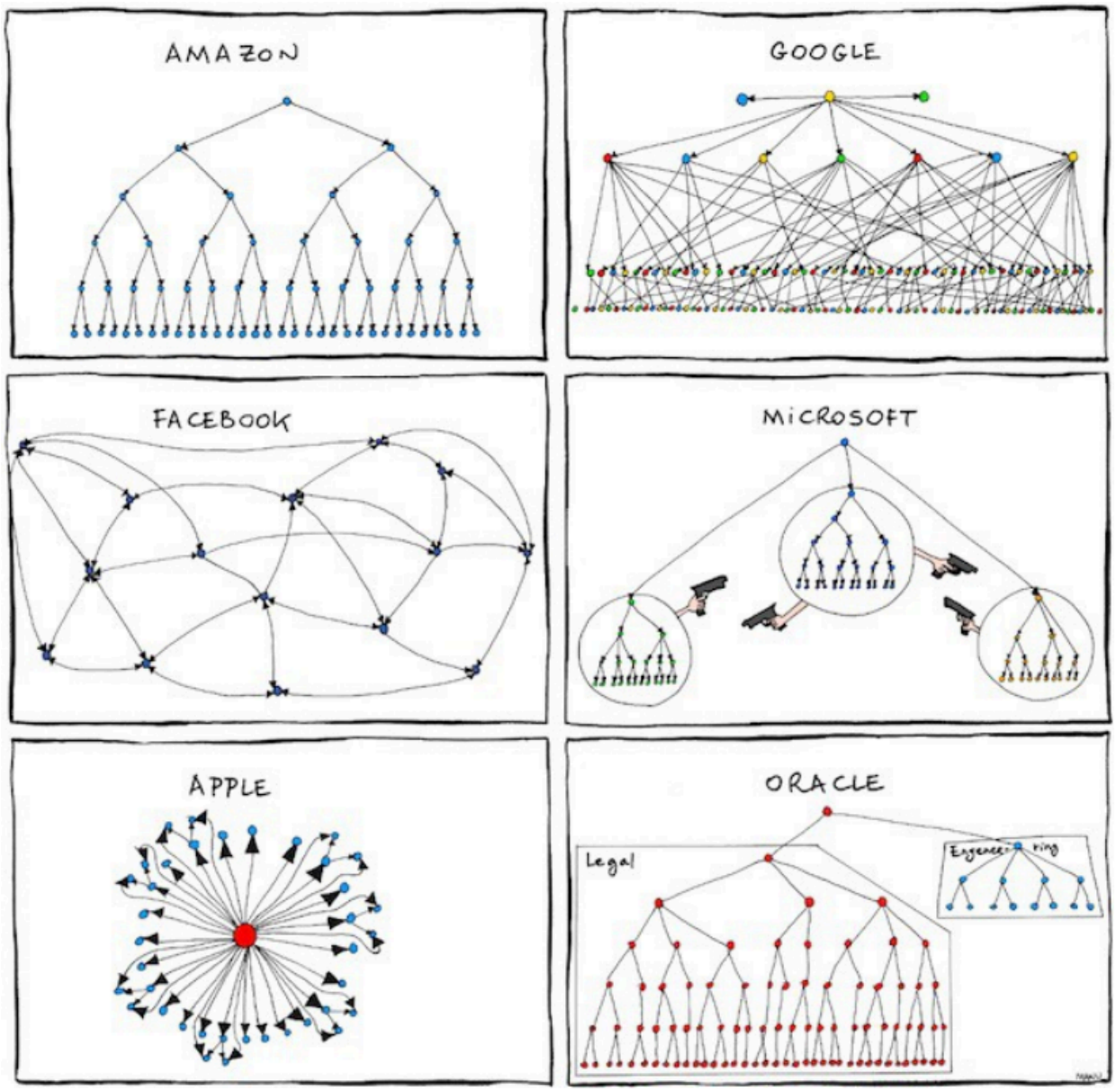


Figure B. This figure represents how a self-reinforcing loop (in red) could go unchallenged and undetected in the absence of meaningful governance interventions for internal deployment.

<figure-caption>
The figure illustrates a self-reinforcing loop in which AI systems progressively automate AI research, leading to increasingly capable AI that further accelerates its own development (Stix, 2025).
</figure-caption>

Organizational structures establish who makes decisions and who is responsible for safety in AI companies. Later sections cover specific safety mechanisms, here, we focus on the governance question: who has the authority within companies to prioritize safety over other goals?

For instance, an effective governance structure determines whether a safety team can delay a model release if they identify concerns, whether executives can override safety decisions, and whether the board has final authority over high-risk deployments. These authority relationships directly affect how safety considerations factor into development decisions.



<figure-caption>
([Kahney, 2011])
</figure-caption>

What governance roles are critical for AI safety? Effective AI governance requires three interconnected levels of internal oversight ([Hadley et al., 2024]; [Schuett, 2023]):

- Board-level oversight structures allocating resources and enforcing safety policies such as Algorithm Review Boards (ARBs) and ethics boards for technical and societal risk assessments, guiding go/no-go decisions on deployments, and establishing oversight with clear lines of accountability ([Hadley et al., 2024]; [Schuett, 2023]).
- Executives allocating resources and enforcing safety policies. Roles like the Chief Artificial Intelligence Officer (CAIO), Chief Risk Officer (CRO), and related

positions to coordinate risk management efforts across the organization, and help translate ethical principles into practice (Schäfer et al., 2022; Janssen et al., 2025).

- Technical safety teams conducting evaluations and recommending mitigations. Teams comprising internal auditors, risk officers, and specialized audit committees for ensuring rigorous risk identification, maintaining audit integrity, and providing operational assurance, with direct reporting lines to the board for independence (Schuett, 2023; Raji et al., 2020).

In May 2025, OpenAI announced a significant restructuring of its governance model. While maintaining nonprofit control, the company transitioned its for-profit subsidiary from an LLC to a Public Benefit Corporation (PBC): the same model used by Anthropic and other AI labs. This change represented an acknowledgment that earlier "capped-profit" structures were designed for "a world where there might be one dominant AGI effort" but were less suitable "in a world of many great AGI companies" (OpenAI, 2025). Frontier AI companies must simultaneously secure billions in capital investment, maintain competitiveness with well-resourced rivals, and preserve governance structures that prioritize safety. As Daniel Colson of the AI Policy Institute notes, this creates difficult tradeoffs where boards might be forced to "weigh total collapse against some form of compromise in order to achieve what it sees as its long-term mission" (TIME, 2024).

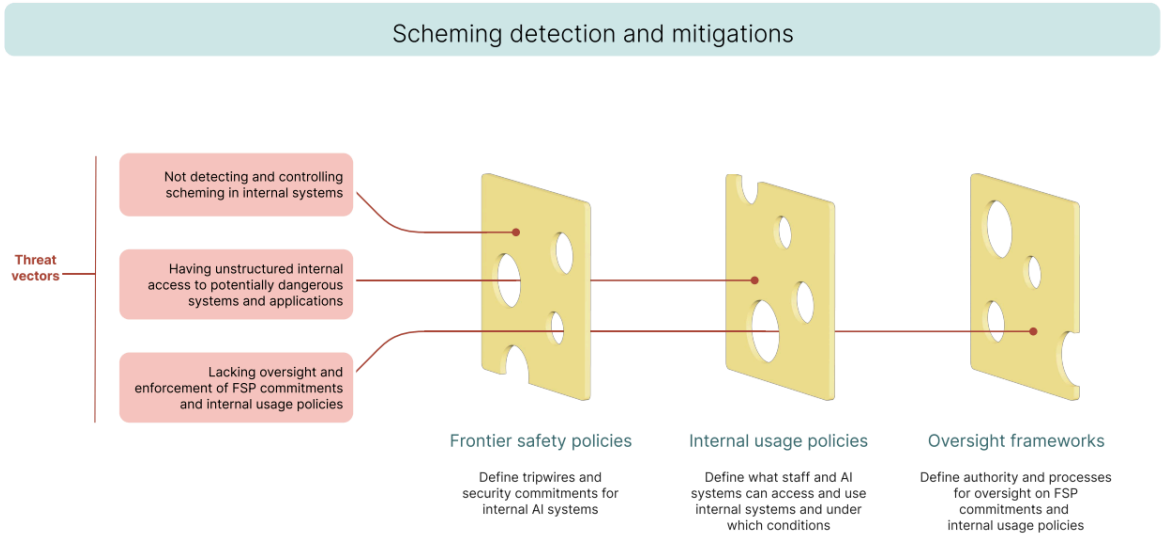


Figure D. Swiss cheese model representing our recommended defense-in-depth strategy against the risk of loss of control via internally deployed misaligned AI (§3.2.(a)). Threat vectors are in red.

<figure-caption>
(Stix, 2025)
</figure-caption>

Frontier Safety Frameworks

Frontier Safety Frameworks are internal policies that AI companies create to guide their development process and ensure they're taking appropriate precautions as their systems become more capable. They're the equivalent of the safety protocols used in nuclear power plants or high-security laboratories, and help bridge internal corporate governance mechanisms and external regulatory oversight in AI safety.

First introduced in 2023, FSFs gained momentum during the Seoul AI Summit in May 2024, where 16 companies committed to implementing such policies. As of March 2025, twelve companies have published comprehensive frontier AI safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, and Nvidia, with additional companies following suit (METR, 2025).

What essential elements define a comprehensive FSF? Despite variations in implementation, most FSFs share several fundamental elements:

- **Capability Thresholds:** FSFs establish specific thresholds at which AI capabilities would pose severe risks requiring enhanced safeguards (Nevo et al., 2024). Common capability concerns include: Biological weapons assistance (such as enabling the creation of dangerous pathogens), Cyber Offensive capabilities (such as automating zero-day exploit discovery), Automated AI research and development (such as accelerating AI progress beyond human oversight), Autonomous replication and adaptation.
- **Model Weight Security:** As models approach dangerous capability thresholds, companies implement increasingly sophisticated security measures to prevent unauthorized access to model weights. These range from standard information security protocols to advanced measures like restricted access environments, encryption, and specialized hardware security (Nevo et al., 2024).
- **Conditions for Halting Development/Deployment:** Most frameworks contain explicit commitments to pause model development or deployment if capability thresholds are crossed before adequate safeguards can be implemented (METR, 2025).
- **Full Capability Elicitation:** Through FSFs, companies commit to evaluating models in ways that reveal their full capabilities rather than underestimating them (Phuong et al., 2024).
- **Evaluation Frequency and Timing:** FSFs establish specific timelines for when evaluations must occur (typically before deployment, during training, and after deployment) with triggers for additional assessments when models show significant capability increases (Davidson et al., 2023).
- **Accountability Mechanisms:** These include: Internal governance roles (for example, Anthropic's "Responsible Scaling Officer"), External advisory boards and third-party audits, Transparency commitments about model capabilities and safety measures, Whistleblower protections for staff reporting safety concerns.

- **Policy Updates:** All FSFs acknowledge the evolving nature of AI risks and commit to regular policy reviews and updates as understanding of risks and best practices improve (METR, 2025).

How do companies operationalize safety frameworks in practice? One promising approach gaining traction is the Three Lines of Defense (3LoD) model adapted from other safety-critical industries (Schuett, 2023):

- **First Line of Defense:** Frontline researchers and developers implement safety measures in day-to-day work, conduct initial risk assessments, and adhere to ethical guidelines and safety protocols.
- **Second Line of Defense:** Specialized risk management and compliance functions, including AI ethics committees, dedicated safety teams, and compliance units provide oversight and guidance.
- **Third Line of Defense:** Independent internal audit functions provide assurance to board and senior management through regular audits of safety practices, independent model evaluations, and assessments of overall preparedness.

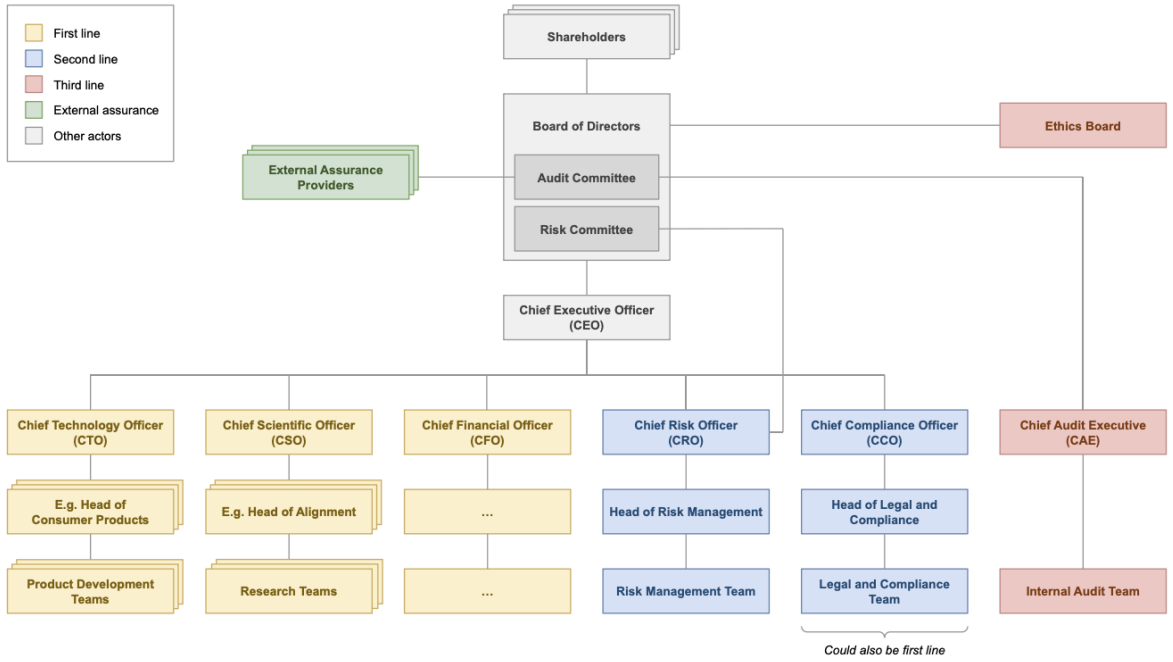


Fig. 2 Sample org chart of an AI company with equivalent responsibilities for each of the three lines

<figure-caption>
Sample org chart of an AI company with equivalent responsibilities for each of the three lines.
</figure-caption>

This multi-layered approach helps ensure risks are identified and managed at multiple levels, reducing the chances of dangerous oversights. For instance, when researchers develop a model with unexpectedly advanced capabilities, safety teams can conduct thorough evaluations and implement additional safeguards, while audit teams review broader processes for managing emergent capabilities (Schuett, 2023).

How do you assess risks for capabilities that don't yet exist? AI capabilities are fast-growing and changing. FSFs incorporate techniques from other safety-critical industries adapted to AI development (Koessler & Schuett, 2023):

- **Scenario Analysis**: Exploring potential future scenarios, like an AI system developing deceptive behaviors or unexpected emergent capabilities.
- **Fishbone Analysis**: Identifying potential causes of alignment failures, such as insufficient safety research, deployment pressure, or inadequate testing.
- **Causal Mapping**: Visualizing how research decisions, safety measures, and deployment strategies interact to influence overall risk.
- **Delphi Technique**: Gathering expert opinions through structured rounds of questionnaires to synthesize diverse perspectives on potential risks.
- **Bow Tie Analysis**: Mapping pathways between causes, hazardous events, and consequences, along with prevention and mitigation measures.

What happens when pre-deployment safeguards are insufficient? Even with rigorous pre-deployment safeguards, dangerous capabilities may emerge after deployment. FSFs increasingly incorporate "deployment corrections", which are comprehensive contingency plans for scenarios where pre-deployment risk management falls short (O'Brien et al., 2023):

- Technical Controls for maintaining continuous control over deployed models through monitoring and modification capabilities, supported by pre-built rollback mechanisms.
- Organizational Preparedness for establishing dedicated incident response teams trained in rapid risk assessment and mitigation.
- Legal Framework for creating clear user agreements that establish the operational framework for emergency interventions.
- Model shutdown such as full market removal or the destruction of the model and associated components.

What benefits do FSFs provide, and where do they fall short? FSFs give companies a way to demonstrate their commitment to proactive risk management. Their public nature enables external scrutiny, while their risk categorization frameworks show engagement with potential failure modes. The frameworks' deliberately flexible structure allows adaptation as understanding of AI risks evolves (Pistillo, 2025).

How can we ensure FSFs are effectively implemented? While FSFs represent progress in AI governance, their effectiveness ultimately depends on implementation. Companies like Anthropic and OpenAI have established notable governance

mechanisms. No matter how well-designed, internal policies remain subject to companies' strategic interests. When safety competes with speed, profitability, or market dominance, even strong internal governance may be compromised. Voluntary measures lack enforceability, and insiders often face misaligned incentives when raising concerns (Zhang et al., 2025).

Limitations and Path Forward

<quote>
<speaker> Demis Hassabis
<position> CEO and Co-Founder of DeepMind, Nobel Prize Laureate in Chemistry
<date>
<source>
<content>
These kinds of decisions are too big for any one person. We need to build more robust governing structures that don't put this in the hands of just a few people.
</content>
</quote>

As AI capabilities continue to advance, governance frameworks must evolve accordingly. There is still significant room for improvement. Some suggest that companies should define more precise, verifiable risk thresholds, potentially drawing on societal risk tolerances from other industries (Pistillo, 2025). For instance, industries dealing with catastrophic risks typically set maximum tolerable risk levels ranging from 1 in 10,000 to 1 in 10 billion per year - quantitative thresholds that AI companies might adopt with appropriate adjustments.

For systemic risks like race dynamics, dual-use misuse, or catastrophic model failure, no single company can be trusted to serve the public interest alone. Corporate governance can buy time but cannot substitute for public accountability or system-wide safety architecture. In the next chapter, we examine how national governance frameworks can provide this essential external oversight, establishing regulatory boundaries that apply across all companies within jurisdictions.

National Governance

<quote>
<speaker> Zhang Jun
<position> China's UN Ambassador
<date>
<source>
<content>
The potential impact of AI might exceed human cognitive boundaries. To ensure that this technology always benefits humanity, we must regulate the development of AI and

prevent this technology from turning into a runaway wild horse [...] We need to strengthen the detection and evaluation of the entire lifecycle of AI, ensuring that mankind has the ability to press the pause button at critical moments.
</content>
</quote>

We established in the previous section that companies can often lack incentives to fully account for the broader societal impact, face competitive pressures that may compromise safety, and lack the legitimacy to make decisions affecting entire populations (Dafoe, 2023). National governance frameworks therefore serve as an essential complement to self-regulatory initiatives, setting regional standards that companies can incorporate into their internal practices.

Unlike traditional technological governance challenges, frontier AI systems generate externalities that span multiple domains: from national security to economic stability, from social equity to democratic functioning. AI systems threaten national security by democratizing capabilities usable by malicious actors, facilitate unequal economic outcomes by concentrating market power in specific companies and countries while displacing jobs elsewhere, and produce harmful societal conditions through extractive data practices and biased algorithmic outputs (Roberts et al., 2024). Traditional regulatory bodies, designed for narrower technological domains, typically lack the necessary spatial remit, technical competence, or institutional authority to effectively govern these systems (Dafoe, 2023).

Consider the contrast with self-driving vehicles, where the primary externalities are relatively well-defined (safety of road users) and fall within existing regulatory frameworks (traffic safety agencies). Frontier AI systems, by contrast, generate externalities that cross traditional regulatory boundaries and jurisdictions, requiring new institutional approaches that can address the expertise gap, coordination gap, and temporal gap in current regulatory frameworks (Dafoe, 2023).

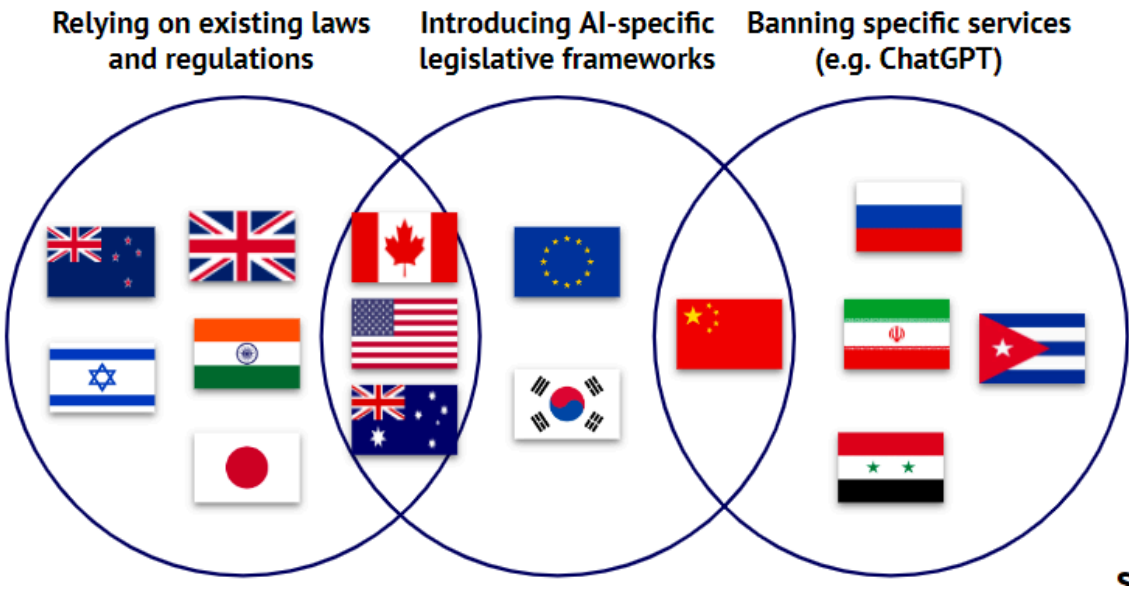
AI systems can cause harm in ways that are not always transparent or predictable. Beyond software bugs or input-output mismatches, risks emerge from how AI systems internally represent goals, make trade-offs, and generalize from data. When these systems are deployed at scale, even subtle misalignments between system behavior and human intent can have widespread consequences. Automated subgoal pursuit, for example, can generate outcomes that are technically correct but socially catastrophic if not carefully constrained (Cha, 2024). Because many of these failure modes are embedded in opaque model architectures and training dynamics, they resist detection through conventional auditing or certification processes. National regulation provides an anchor for accountability by requiring developers to build, test, and deploy systems in ways that are externally verifiable, legally enforceable, and publicly legitimate.
As we will see in this section, major regions have developed distinctly different regulatory philosophies that reflect their unique institutional contexts and political

priorities. Understanding these national frameworks will provide context for our subsequent analysis of international governance mechanisms, which must navigate and harmonize these regional differences to create effective global standards for AI systems whose impacts transcend national borders.

Across the last decade, over 30 countries have released national AI strategies outlining their approach to development, regulation, and adoption. These strategies differ widely in emphasis, but when systematically analyzed, they fall into three recurring governance patterns: development, control, and promotion (Papyshev et al., 2023). In development-led models, such as those in China, South Korea, and Hungary, the state acts as a strategic coordinator, directing public resources toward AI infrastructure, research programs, and national missions. Control-oriented approaches, prominent in the European Union and countries like Norway and Mexico, emphasize legal standards, ethics oversight, and risk monitoring frameworks. Promotion-focused models, including the United States, United Kingdom, and Singapore, adopt a more decentralized approach: the state acts primarily as an enabler of private sector innovation, with relatively few regulatory constraints. These differences matter. Any attempt to build international governance frameworks will need to account for the structural asymmetries between these national regimes, particularly around enforcement authority, accountability mechanisms, and institutional capacity (Papyshev et al., 2023).

Governance Mode	Definition	State Role	Policy Focus	Typical Countries
Primus inter pares	State leads coordination and direction, but shares execution with other actors	Strategic coordinator and enabler	**Innovation-focused**, public-private projects	China, Japan, Russia, South Korea, Hungary, Czech Republic
Command-and-control	State imposes rules to mitigate risks from AI	Regulator and guarantor	**Risk mitigation**, ethics, data, standards	Germany, Sweden, Finland, Netherlands, Mexico, Uruguay
Self-regulation / Oligopoly	Private sector leads with minimal state involvement	Promoter and facilitator (indirect role)	**Market-led innovation**, talent, investment	UK, US, India, Ireland, Singapore, Saudi Arabia, Australia

<figure-caption>
The state's role in governing artificial intelligence: development, control, and promotion through national strategies (Papyshev et al., 2023).
</figure-caption>



<figure-caption>
([State of AI Report, 2023])
</figure-caption>

International Governance

<quote>
<speaker> António Guterres
<position> UN Secretary-General
<date>
<source>
<content>
AI poses a long-term global risk. Even its own designers have no idea where their breakthrough may lead. I urge [the UN Security Council] to approach this technology with a sense of urgency [...] Its creators themselves have warned that much bigger, potentially catastrophic and existential risks lie ahead.
</content>
</quote>

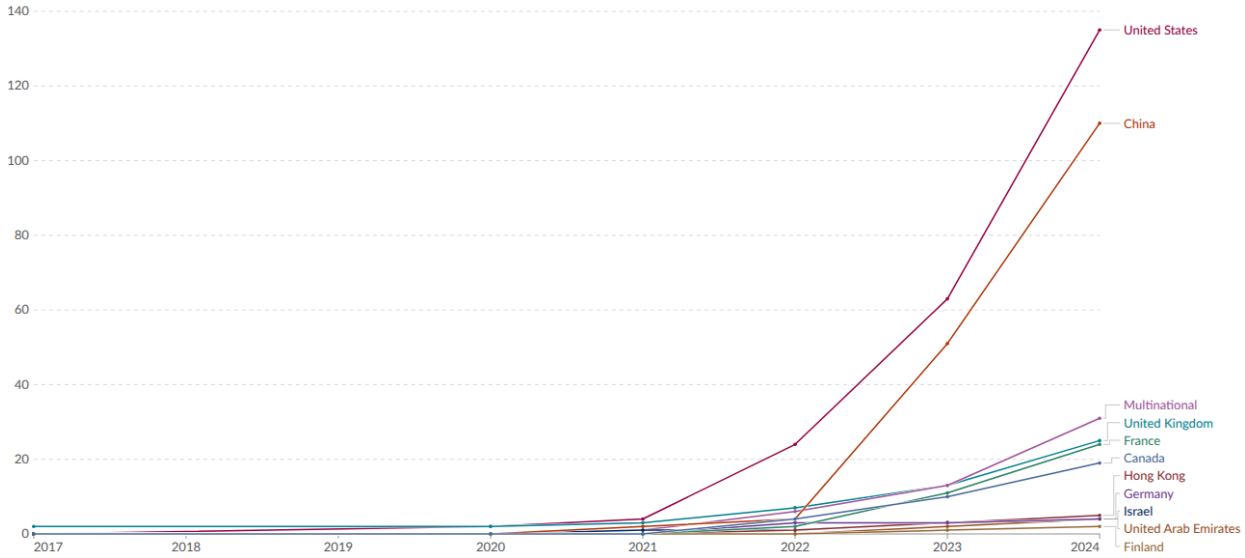
<quote>
<speaker> Kamala Harris
<position> Former US Vice President
<date>
<source>
<content>
[...] just as AI has the potential to do profound good, it also has the potential to cause profound harm. From AI-enabled cyberattacks at a scale beyond anything we have seen

before to AI-formulated bio-weapons that could endanger the lives of millions, these threats are often referred to as the "existential threats of AI" because, of course, they could endanger the very existence of humanity. These threats, without question, are profound, and they demand global action.
</content>
</quote>

Can't individual countries just regulate AI within their own borders? The short answer is: no, not effectively. Effective management of advanced AI systems requires coordination that transcends national borders. This stems from three fundamental problems (Ho et al., 2023):

- **No single country has exclusive control over AI development.** Even if one nation implements stringent regulations, developers in countries with looser standards could still create and deploy potentially dangerous AI systems affecting the entire world (Hausenloy et al., 2023).
- **AI risks have a global impact.** The regulation of those risks requires international cooperation (Tallberg et al., 2023). When asked about China's participation in the Bletchley AI Safety summit, James Cleverly, former UK Foreign Secretary correctly noted: "we cannot keep the UK public safe from the risks of AI if we exclude one of the leading nations in AI tech."
- **Race-to-the-bottom dynamics.** Countries fear competitive disadvantage in the AI race, which creates incentives for regulatory arbitrage and undermines safety standards globally (Lancieri et al., 2024). International governance can help align incentives between nations, encouraging responsible AI development without forcing any one country to sacrifice its competitive edge (Li, 2025).

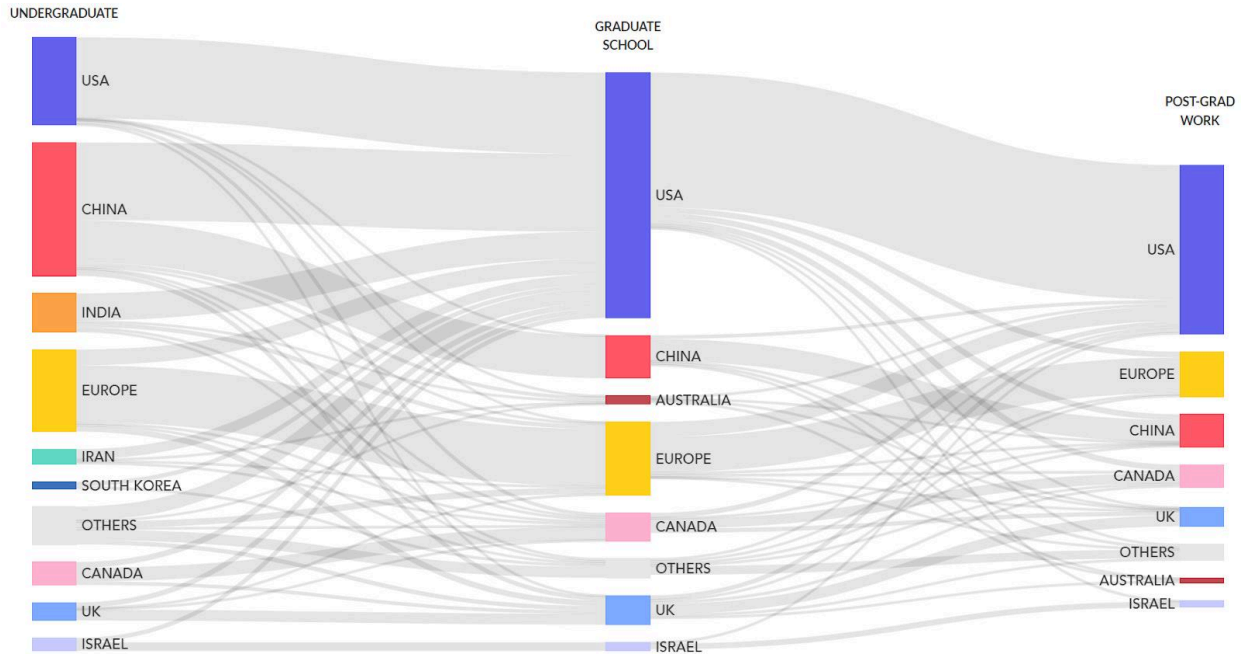
<iframe-static-figure>



</iframe-static-figure>

<iframe
src="https://ourworldindata.org/grapher/cumulative-number-of-large-scale-ai-systems-by-country?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<iframe-caption>
Cumulative number of large-scale AI systems by country since 2017. Refers to the location of the primary organization with which the authors of a large-scale AI systems are affiliated (Giattino et al., 2023).
</iframe-caption>

How do national policies affect global AI development? Even seemingly domestic regulations (such as immigration policies, see below) can reshape the global AI landscape through various spillover mechanisms.



<figure-caption>
What are the career paths of top-tier AI researchers? (MacroPolo)
</figure-caption>

Companies worldwide, eager to maintain access to the lucrative European market, often find it more cost-effective to adopt EU standards across their entire operations rather than maintaining separate standards for different regions. For example, a U.S. tech company developing a new AI-powered facial recognition system for use in public spaces may see this system being classified as “high-risk” under the EU AI Act. This would subject it to strict requirements around data quality, documentation, human oversight, and more. Companies then have a choice to either develop two separate versions of your product, one for the EU market and one for everywhere else, or simply apply the EU standards globally. Many will be tempted to choose the second option, to

minimize their cost of compliance. This is what's known as the “Brussels Effect” (Bradford, 2020): EU regulations can end up shaping global markets, even in countries where those regulations don't formally apply.

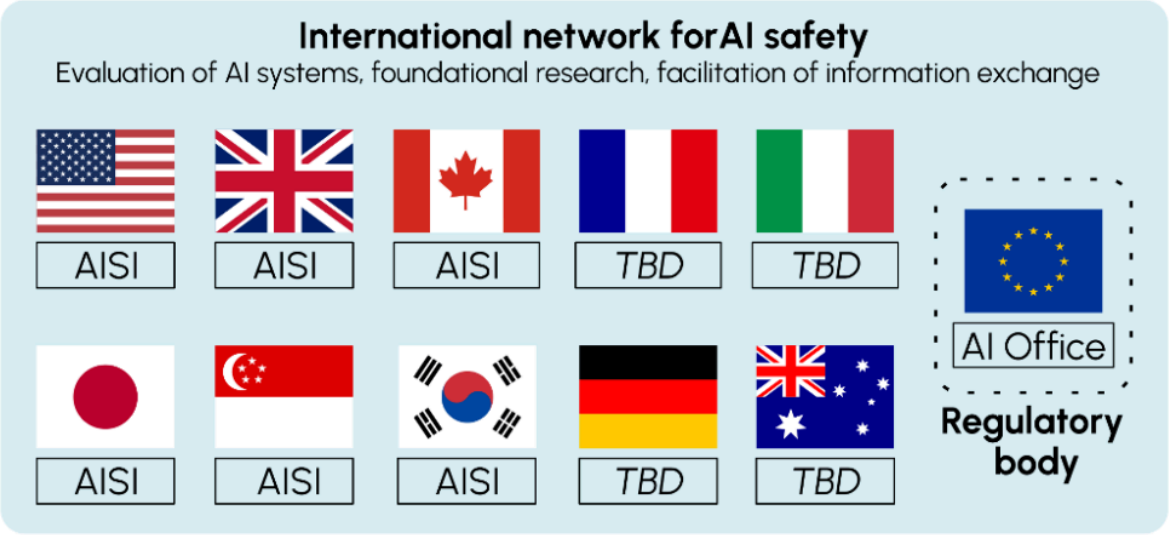
The Brussels Effect can manifest in two ways:

- **De facto adoption:** Companies often voluntarily adopt EU standards globally to avoid the complexity and cost of maintaining different standards for different markets.
- **De jure influence:** Other countries frequently adopt regulations similar to the EU's, either to maintain regulatory alignment or because they view the EU's approach as a model worth emulating.

The EU's regulations might offer the first widely adopted and mandated operationalization of concepts like "risk management" or "systemic risk" in the context of frontier AI. As other countries grapple with how to regulate advanced AI systems, they may look to the EU's framework as a starting point (Siegmann & Anderljung 2022).

<quote>
<speaker> Ursula von der Leyen
<position> Head of EU Executive Branch
<date>
<source>
<content>
[We] should not underestimate the real threats coming from AI [...] It is moving faster than even its developers anticipated [...] We have a narrowing window of opportunity to guide this technology responsibly.
</content>
</quote>

In 2023, the US and UK governments both announced new institutes for AI safety. As of 2025, there are at least 12 national AI Safety Institutes (AISIs) established worldwide. These include institutes from the United States, United Kingdom, Canada, France, Germany, Italy, Japan, South Korea, Singapore, Australia, Kenya, and India. The European Union has established the European AI Office, which functions similarly to national AISIs. These institutes collaborate through the International Network of AI Safety Institutes, launched in November 2024, to coordinate research, share best practices, and develop interoperable safety standards for advanced AI systems.



<figure-caption>
These countries are part of the international network for AI safety, with their respective national bodies dedicated to AI safety (Variengien & Martinet, 2024).
</figure-caption>

Global governance efforts also face major obstacles. Strategic competition between leading powers, who view AI as both a national security asset and an economic engine, often undermines cooperation. Power asymmetries further complicate negotiations: countries with advanced AI capabilities, like the United States and China, may resist international constraints, while others may demand technology transfer and capacity-building support in exchange for participation. Divergent political systems and values also pose barriers, with disagreements over issues such as privacy, free expression, and state authority. For example, China's Global AI Governance Initiative centers sovereignty and non-interference, contrasting with Western frameworks rooted in individual rights and democratic accountability (Hung, 2025; Hsu et al., 2023). Perhaps most significantly, deep trust deficits between major powers, fueled by tensions over trade, intellectual property, and human rights, make it difficult to reach credible, enforceable agreements, adding to the complex geopolitical landscape shaping the future of international AI governance (Mishra, 2024).

<figure-caption>
Cartoon highlighting a discrepancy between countries' statements and their true intentions in the context of the U.K.'s november 2023 AI Safety Summit (The Economist)
</figure-caption>

<note-box>
<collapsed> True
<title> Existing International Mechanisms (2025)
<content>
Despite these challenges, a patchwork of international initiatives has emerged to address AI governance:

- **The series of Global AI Summits**: Launched by the UK in 2023, the summits have been a platform for major stakeholders across the AI ecosystem to come together and discuss global priorities for AI safety, innovation, and governance. They continue to occur biannually, with a different country hosting each summit.
- **The Hiroshima AI Process**: Launched by the G7 nations, this initiative aims to promote responsible AI development and use through coordinated policies.
- **United Nations efforts**: Includes UNESCO's AI ethics recommendations, the High-Level Advisory Body, and the upcoming Global Digital Compact, a United Nations framework for international digital cooperation, focused on a common digital future with an AI component.
- **OECD guidelines**: The Organisation for Economic Co-operation and Development has been particularly influential in shaping AI governance

principles that inform national policies, and continues to guide regional frameworks with a focus on rights, transparency, and accountability.
- **Council of Europe AI treaty**: This proposed treaty aims to protect human rights in the context of AI development and use, focusing on ethical boundaries.
- **China's Global AI Governance Initiative**: Demonstrating that AI governance is a priority even for nations often at odds with Western powers, China has put forth its own proposal for international AI governance.

</content>
</note-box>

How does international technology governance typically evolve? Understanding the progression of international policymaking helps contextualize current AI governance efforts and identify potential paths forward. International policymaking typically progresses through several stages (Badie et al., 2011):

- **Agenda setting:** Identifying the issue and placing it on the international agenda.
- **Policy formulation:** Developing potential solutions and approaches.
- **Decision making:** Choosing specific courses of action.
- **Implementation:** Putting chosen policies into practice.
- **Evaluation:** Assessing effectiveness and making adjustments.

For AI governance, we're still largely in the early stages of this process. The Series of AI Summits, the Network of AI Safety Institutes, and other international frameworks all represent progress in agenda setting and initial policy formulation. But the real work of crafting binding international agreements and implementing them still lies ahead.

Previous international governance efforts provide valuable lessons for AI. So, what can we learn from decades of nuclear arms control efforts? Let's consider three important lessons (Maas, 2019):

- **The power of norms and institutions.** Despite early fears of rapid proliferation, only nine countries possess nuclear weapons nearly 80 years after their development which resulted from concerted efforts to build global norms against nuclear proliferation and use. The Nuclear Non-Proliferation Treaty (NPT), signed in 1968, created a framework for preventing the spread of nuclear weapons and helped promote peaceful uses of nuclear technology.
- **The role of epistemic communities.** The development of nuclear arms control agreements wasn't solely the work of diplomats and politicians. It relied heavily on input from scientists, engineers, and other technical experts who understood the technology and its implications. These experts formed a network of professionals with recognized expertise in a particular domain, or as what political scientists call an "epistemic community". They played important roles in shaping policy debates, providing technical advice, and even serving as back-channel diplomats during tense periods of the Cold War. Unlike nuclear physicists, who were often employed directly by governments, many AI experts work in the private sector, so a challenge to forming such networks for global AI

governance will be ensuring that epistemic communities can effectively inform policy decisions.

- **The persistent challenge of "normal accidents."** Despite decades of careful management, the nuclear age has seen several incidents where human error, technical malfunctions, or misunderstandings nearly led to catastrophe. Sociologist Charles Perrow termed these "normal accidents," arguing that in complex, tightly-coupled systems, such incidents are inevitable (1985). Applying the concept to AI, we could see unexpected interactions and cascading failures increase as AI systems become more complex and interconnected. The speed at which AI systems operate could mean that a "normal accident" in AI might unfold too quickly for human intervention, challenging the notion of "meaningful human control," often proposed as a safeguard for AI systems (Maas, 2019).

Policy Options

<quote>
<speaker> Demis Hassabis
<position> Co-Founder and CEO of DeepMind
<date>
<source>
<content>
We must take the risks of AI as seriously as other major global challenges, like climate change. It took the international community too long to coordinate an effective global response to this, and we're living with the consequences of that now. We can't afford the same delay with AI [...] then maybe there's some kind of equivalent one day of the IAEA, which actually audits these things.
</content>
</quote>

Several institutional arrangements could support international AI governance (Maas & Villalobos, 2024):

- **Scientific Consensus-Building:** Similar to the Intergovernmental Panel on Climate Change (IPCC), a dedicated body could provide regular reports on AI capabilities and risks to inform policymakers and the public. Given the rapid pace of AI development, this body would need to be nimbler than traditional scientific consensus-building organizations.
- **Political Consensus-Building and Norm-Setting:** Building on scientific consensus, a forum for political leaders could develop shared norms and principles, perhaps structured like the UN Framework Convention on Climate Change (UNFCCC). Such a body could facilitate ongoing dialogue, negotiate agreements, and adapt governance approaches as the technology evolves.

- **Coordination of Policy and Regulation:** An international body focused on policy coordination could help harmonize AI regulations across countries, reducing fragmentation and preventing regulatory arbitrage opportunities.
- **Enforcement of Standards and Restrictions:** Mechanisms for monitoring compliance and enforcing agreed-upon standards would be necessary for effective governance.
- **Stabilization and Emergency Response:** A global network of companies, experts, and regulators ready to assist with major AI system failures could help mitigate risks. This group could work proactively to identify potential vulnerabilities in global AI infrastructure and develop contingency plans, similar to the International Atomic Energy Agency's Incident and Emergency Centre but operating on much faster timescales.
- **International Joint Research:** Collaborative research could help ensure that frontier AI development prioritizes safety and beneficial outcomes, similar to how CERN facilitates international scientific cooperation.
- **Distribution of Benefits and Access:** An institution focused on ensuring equitable access to AI benefits could prevent harmful concentration of capabilities and ensure the technology's benefits are widely distributed through mechanisms like a global fund for AI development assistance or technology transfers.

Governance Function	Institutional Model	Real-World Analogy	Purpose
Scientific Evaluation	Scientific Consensus-Building	IPCC (climate science)	Assess capabilities and risks; inform global policy
Political Norm-Setting	Political Consensus Forum	UNFCCC (climate governance)	Facilitate negotiation of principles and long-term agreements
Policy Harmonization	Coordination of Policy and Regulation	OECD Guidelines	Align national regulations; reduce fragmentation
Compliance Oversight	Enforcement of Standards and Restrictions	Arms control bodies; FATF	Monitor and enforce adherence to global standards
Emergency Management	Stabilization and Emergency Response Network	IAEA Incident and Emergency Centre	Respond to systemic failures; prepare contingency protocols
Collaborative R&D	International Joint Research Facility	CERN	Advance safety-focused frontier research
Equity and Access	Benefit-Sharing and Global Assistance Mechanism	Global Fund; Technology transfer partnerships	Distribute capabilities; prevent concentration

<figure-caption>
An overview table of governance functions and their purpose.
</figure-caption>

What does this mean for designing effective institutions? There is no one-size-fits-all solution. Institutions for global AI governance must be tailored to the unique characteristics of the technology: rapid iteration cycles, broad deployment contexts, and uncertain future trajectories. We will likely need a network of complementary institutions, each fulfilling specific governance functions listed above. The key is not just which institutions we build, but why and how. What specific risks and benefits

require international coordination? What functions are essential to manage them? And which designs best match those functions under real-world constraints? Without clear answers, institutional design risks becoming a mirror of past regimes rather than a response to the challenges of advanced AI (DeepMind, 2024).

Implementation

AI Safety Standards

What approaches exist for developing AI safety standards at the national level? Various approaches to developing safety standards exist within national contexts, from government-led standardization bodies to public-private collaborative processes. National standards bodies play a critical role in developing and implementing AI safety standards that align with each country's policy priorities and technological capabilities (Cihon, 2019). The EU AI Act demonstrates this through its requirement for a Code of Practice that specifies high-level obligations for General-Purpose AI models. In the United States, the National Institute of Standards and Technology (NIST) has developed an AI Risk Management Framework that serves as a voluntary standard within American jurisdiction. In 2021, the Standardization Administration of China (SAC) released a roadmap for AI standards development that includes over 100 technical and ethical specifications from algorithmic transparency to biometric recognition safety. Coordinated by government agencies such as the Ministry of Industry and Information Technology (MIIT) and the China Electronics Standardization Institute (CESI). Unlike in the US or EU, where standards are often multistakeholder-developed or market-driven, China's process is highly centralized and closely linked to its broader geopolitical ambitions (Ding, 2018).

How do national standards bodies develop effective AI safety standards? National standards have experience in governing various socio-technical issues within their countries. For example, national cybersecurity standards have spread across industries, environmental sustainability standards have prompted significant corporate investments, and safety standards have been implemented across sectors from automotive to energy. Expertise from other high-stakes industries can be leveraged to develop effective AI safety standards tailored to a country's specific needs and regulatory environment (Cihon, 2019). National standards can be used to spread a culture of safety and responsibility in AI research and development in four ways:

- The criteria within standards establish rules and expectations for safety practices within the country's AI ecosystem.
- Standards embed individual researchers and organizations within a larger network of domestic accountability.
- Regular implementation of standards helps researchers internalize safety routines as part of standard practice.

- When standards are embedded in products and software packages, they reinforce safety considerations regardless of which domestic organizations use the system.

These mechanisms help create what some researchers have called a "safety mindset" among AI practitioners within the national AI ecosystem. National standards serve as effective tools for fostering a culture of responsibility and safety in AI development, which is essential for long-term societal benefit (Cihon, 2019).

Regulatory Visibility

Regulatory visibility requires active, independent scrutiny of AI systems before, during, and after deployment. As frontier AI systems become increasingly integrated into society, external scrutiny (involving outside actors in the evaluation of AI systems) offers a powerful tool for enhancing safety and accountability. Effective external scrutiny should adhere to the ASPIRE framework, which proposes six criteria for effective external evaluation (Anderljung et al., 2023):

- **Access:** External scrutineers need appropriate access to AI systems and relevant information.
- **Searching attitude:** Scrutineers should actively seek out potential issues and vulnerabilities.
- **Proportionality to the risks:** The level of scrutiny should match the potential risks posed by the system.
- **Independence:** Scrutineers should be free from undue influence from AI developers.
- **Resources:** Adequate resources must support thorough scrutiny.
- **Expertise:** Scrutineers must possess the necessary technical and domain-specific expertise.

Some countries are exploring model registries, which are centralized databases that include architectural details, training procedures, performance metrics, and societal impact assessments. These registries support structured oversight and can act as early-warning systems for emerging capabilities, helping regulators detect dangerous trends before they materialize as harms (McKernon et al., 2024). Different jurisdictions take different approaches, but model documentation typically encompasses:

- Basic documentation (model identification, intended use cases)
- Technical specification (architecture, parameters, computational requirements)
- Performance documentation (benchmark results, capability evaluations)
- Impact assessment (societal effects, safety implications, ethical considerations)
- Deployment documentation (implementation strategies, monitoring plans)

Another method of regulatory visibility for AI is the Know Your Customer (KYC) system. KYC systems are already an established part of financial regulation, used to detect and prevent money laundering and terrorist financing. They have proven

effective in their ability to identify high-risk actors before a transaction takes place. The same principle can be applied to compute access. As discussed in the compute governance section, frontier models require massive computational resources, often concentrated in a small number of hyperscale providers who serve as natural regulatory chokepoints. A KYC system for AI would enable governments to detect the development of potentially hazardous systems early, prevent covert model training, and implement export controls or licensing requirements with greater precision. Since this approach targets capability thresholds rather than use cases, it could serve as a preventative tool for risk management rather than a reactive one to deployment failures (Egan & Heim, 2023). However, implementing a KYC regime for compute involves several open questions. Providers would need clear legal mandates, technical criteria for client verification, and processes for escalating high-risk cases to authorities. Jurisdictional fragmentation is a challenge. Many developers rely on globally distributed compute services, and without international cooperation, KYC regimes risk being undercut by regulatory arbitrage. To be effective, a compute-based KYC system would need to align with other transparency mechanisms, such as model registries and incident reporting systems (Egan & Heim, 2023).

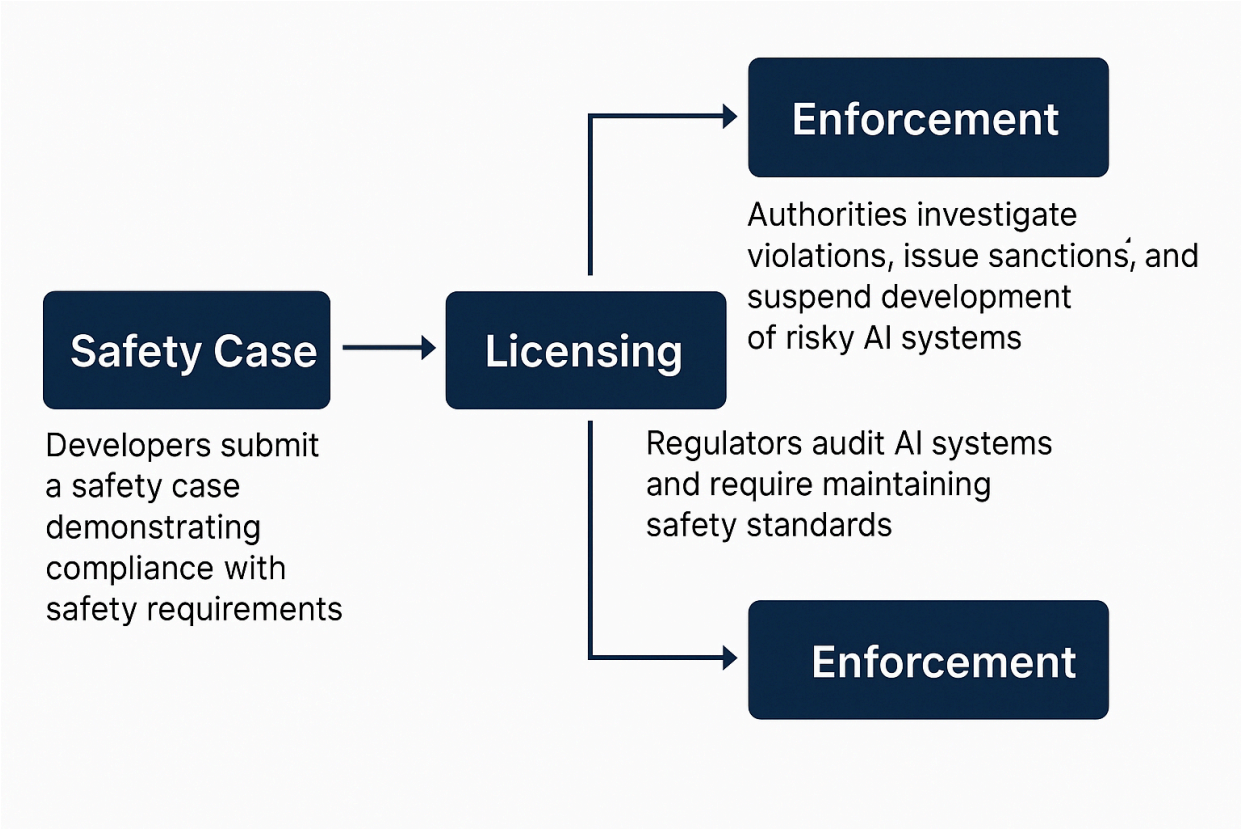
How can national policies support responsible information-sharing? Responsible reporting of information is important for both self-regulation and government oversight. As we discussed in the corporate governance section, companies developing and deploying frontier AI systems have primary access to information about their systems' capabilities and potential risks, and sharing this information responsibly can significantly improve the state's ability to manage AI risks (Kolt et al., 2024). National policies must address the tension between transparency and proprietary control. One approach is tiered disclosure, in which technical documentation is provided to regulators under confidentiality agreements while public communication remains high-level and risk-focused. Another approach is through anonymized or aggregated sharing of data, which enables statistical insight without revealing sensitive implementation details.

Although incident reporting systems from other industries, such as the confidential and non-punitive Aviation Safety Reporting System (ASRS) in the United States, offer useful precedents, no equivalent system yet exists for AI. In aviation, it is clear what constitutes an incident or near-miss, but with AI, the lines can be blurry. Adapting this model would require clear definitions of what constitutes an "incident," with structured categories ranging from model misbehavior to societal harms. Current national efforts on this are fragmented. In the EU, the AI Act mandates reporting of "serious incidents" by high-risk and general-purpose AI developers. In China, the Cyberspace Administration is building a centralized infrastructure for real-time reporting of critical failures under cybersecurity law. In the United States, incident reporting remains sector-specific, with preliminary efforts underway in health and national security (Farrell, 2024; Cheng, 2024; OECD, 2025).

Ensuring Compliance

What regulatory tools can ensure compliance with AI safety standards? For high-risk AI systems, oversight mechanisms must go beyond voluntary standards or one-time evaluations. Many researchers have proposed licensing regimes that would mirror regulatory practices in sectors such as pharmaceuticals or nuclear energy. In these domains, operators must obtain and maintain licenses by demonstrating continuous compliance with strict safety and documentation requirements. Applied to frontier AI, this approach would involve formal approval processes before model deployment, periodic audits, and the ability for authorities to revoke licenses in cases of non-compliance (Buhl et al., 2024). A credible licensing framework would require developers to submit a structured safety case, which is a formal argument supported by evidence showing that a system meets safety thresholds for deployment. This could include threat modeling, red-teaming results, interpretability evaluations, and post-deployment monitoring plans. Safety cases provide a mechanism for both ex ante approval and for tracking whether safety claims continue to hold as systems evolve in deployment. Embedding these requirements into the licensing process can help governments establish a continuous cycle of review, feedback, and technical verification (Buhl et al., 2024).

How would enforcement work in practice? Licensing frameworks must be supported by agencies with the power to investigate violations, impose sanctions, and suspend development. National enforcement practices vary between horizontal governance (applying general rules across sectors) and vertical regimes (targeting specific domains like healthcare or finance) (Cheng & McKernon, 2024). For example, the European Union's AI Act establishes enforcement authority through horizontal governance framework with the European AI Office, which can investigate, issue fines up to 3% of global annual turnover, and mandate corrective action, combined with mandatory incident reporting, systemic risk mitigation requirements, and a supporting Codes of Practice for GPAI models (Cheng & McKernon, 2024). In contrast, China's Cyberspace Administration (CAC) exercises centralized enforcement powers under a vertical regulatory framework. While its approach prioritizes rapid intervention and censorship compliance, the CAC lacks transparent procedural checks and often relies on vague criteria for enforcement. In the United States, enforcement is fragmented. While export controls are strictly applied through agencies like the Department of Commerce, broader AI safety compliance has been delegated to individual agencies, with no national licensing authority. As a result, enforcement actions are often reactive and domain-specific, and rely on discretionary executive powers (Cheng & McKernon, 2024). Striking the right balance between these approaches will depend on institutional capacity, developer incentives, and the pace of AI advancement. In some cases, using existing sectoral authorities may suffice. In others, new institutions will be required to handle general-purpose capabilities that fall outside traditional regulatory categories (Dafoe, 2023).



<figure-caption>
The flow from safety cases to enforcement.
</figure-caption>

Limitations and Trade-Offs

Every governance approach faces fundamental constraints that no amount of institutional design can fully overcome. Understanding these limitations helps set realistic expectations and identifies where innovation is most needed (Dafoe, 2023).

Some risks resist technical solutions. Despite advances in interpretability and evaluation, we still cannot fully understand or predict AI behavior. Black box models make verification difficult. Emergent capabilities appear unexpectedly. The gap between our governance ambitions and technical capabilities are substantial (Mukobi, 2024). Current safety techniques like RLHF and constitutional AI show promise for today's models but may fail catastrophically with more capable systems. We're building governance frameworks around safety approaches that might become obsolete. This fundamental uncertainty requires adaptive frameworks that can evolve with understanding (Ren et al., 2024).

Measurement challenges undermine accountability. We lack robust metrics for many safety-relevant properties. How do you measure a model's tendency toward deception?

Its potential for autonomous improvement? Its resistance to misuse? Without reliable measurements, compliance becomes a matter of interpretation rather than verification (Narayan & Kapoor, 2024). The EU AI Act, for example, requires "systemic risk" assessments, but provides limited guidance on how to measure such risks quantitatively (Cheng, 2024).

Expertise shortages create critical bottlenecks. The number of individuals who deeply understand both advanced AI systems and governance remains extremely limited, and this gap exists at every level from company safety teams and regulators to international bodies. A lack of interdisciplinary talent undermines efforts to anticipate and manage emerging risks (Brundage et al., 2018). Institutional capacity for technical evaluation and oversight is similarly weak in many jurisdictions (Cihon et al., 2021). Governments struggle to attract and retain the expertise needed to regulate powerful AI models, anc technically literate, governance-aware professionals may be the most serious constraint on effective AI governance (Dafoe, 2023; Reuel & Bucknall, 2024). Much of the existing talent is concentrated in a few dominant firms, limiting public-sector oversight and reinforcing asymmetries in governance capacity (Brennan et al., 2025).

Coordination costs escalate faster than capabilities. Each additional stakeholder, requirement, and review process adds friction to AI development (Schuett, 2023). While some friction helps ensure safety, excessive bureaucracy can drive development to less responsible actors or underground entirely (Zhang et al., 2025). Speed mismatches create fundamental governance gaps. AI capabilities advance in months while international agreements take years to negotiate (Grace et al., 2024). GPT-4's capabilities surprised experts in March 2023; by the time regulatory responses emerged in 2024, the technology had moved on to multimodal systems and AI agents (Casper et al., 2024). Safety researchers emphasize precaution and worst-case scenarios, companies prioritize competitive position and time-to-market, governments balance multiple constituencies with conflicting demands, and users want beneficial capabilities without understanding risks (Dafoe, 2023).

Regulatory arbitrage undermines safety standards across borders. If Europe implements strict safety requirements while other regions remain permissive, development may simply shift locations (Lancieri et al., 2024). As we previously discussed in the proliferation section, the digital nature of AI makes it so that a model can be trained in Singapore, deployed from Ireland, and used globally (Seger et al., 2023). Companies may bifurcate offerings, providing safer systems to regulated markets while deploying riskier versions elsewhere. True global coverage requires more than powerful individual jurisdictions.

Conclusion

The governance frameworks examined throughout this chapter provide essential tools for managing AI risks, but tools alone don't determine outcomes. Success requires

choosing the right priorities, building necessary capabilities, and maintaining frameworks that evolve with the technology.

Technical expertise in government needs dramatic expansion across every major economy. The UK and US AI Safety Institutes demonstrate what's possible with sufficient resources and political support (Dafoe, 2020). This requires competitive compensation to attract top talent, career paths that value public service, exchange programs with industry and academia, and protection from political interference (Zaidan & Ibrahim, 2024). Currently, properly aligning advanced AI systems with human values will require resolving many uncertainties related to the psychology of human rationality, emotion, and biases, and most government agencies lack even basic technical literacy about AI systems (Irving & Askell, 2019).

Audit and assessment capabilities must professionalize into a distinct field. As AI systems become more complex, evaluation requires specialized expertise that goes beyond traditional software testing (Anderljung et al., 2023). Building this profession involves developing certification programs for AI auditors, creating standard methodologies and tools, establishing professional organizations and ethics codes, and ensuring independence from both developers and regulators (Schuett, 2023).

International coordination mechanisms need dedicated resources and authority. Current efforts rely heavily on voluntary participation and limited budgets (Ho et al., 2023). Effective coordination requires dedicated secretariats with technical expertise, funding for participation from developing countries, translation and communication services, and infrastructure for secure information sharing (Maas & Villalobos, 2023).

Governance frameworks must evolve as fast as the technology they govern. Static regulations will quickly become either irrelevant or obstructive (Casper, 2024). Building adaptive capacity into governance systems is essential for long-term effectiveness (Anderljung et al., 2023). This means mandatory annual reviews of capability thresholds, evaluation methodologies, enforcement priorities, and lessons from incidents (McKernon et al., 2024).

Scenario planning helps prepare for discontinuous change in AI development. Current governance assumes relatively continuous AI progress, but development could accelerate suddenly through algorithmic breakthroughs, decelerate due to technical barriers, or bifurcate with different regions pursuing incompatible approaches (Grace et al., 2024). Governance systems need contingency plans for rapid capability jumps, major AI accidents, breakdown of international cooperation, and emergence of artificial general intelligence (Cotra, 2022).

Learning from implementation enables continuous improvement over the critical next few years. The coming period will generate enormous amounts of data about what works in AI governance (Dafoe, 2020). Systematic learning requires tracking governance

interventions and outcomes, sharing best practices across jurisdictions, acknowledging and correcting failures, and updating frameworks based on evidence (Cihon, 2019). The temptation will be to lock in current approaches - we must resist this in favor of evidence-based evolution (Dafoe, 2018).

The choices made in the next few years will shape humanity's relationship with artificial intelligence for decades to come. As AI capabilities advance and become more deeply embedded in critical systems, retrofitting governance becomes increasingly difficult (Anderljung et al., 2023). We have the tools, knowledge, and warning signs needed to build effective governance (Bengio et al., 2025). What remains is the collective will to act before events force our hand (Dafoe, 2018).

The path forward requires acknowledging uncomfortable truths: voluntary corporate measures won't suffice for systemic risks (Papagiannidis, 2025), national approaches need unprecedented coordination despite geopolitical tensions (Ho et al., 2023), and international governance faces enormous technical and political challenges (Maas & Villalobos, 2024). Yet history shows that humanity can rise to meet technological challenges when the stakes become clear and immediate (Maas, 2019).

With AI, the stakes could not be higher, and the timeline could not be shorter (Kokotajlo et al., 2025). The question is not whether we need comprehensive governance: the evidence presented throughout this chapter makes that case definitively. The question is whether we'll build it in time, with the technical sophistication and institutional authority required to govern humanity's most powerful technology, and the window for answering that question is narrowing with each new model release.

Appendix: Data Governance

What role does data play in AI risks? Data fundamentally shapes what AI systems can do and how they behave. For frontier foundation models, training data influences both capabilities and alignment - what systems can do and how they do it. Low quality or harmful training data could lead to misaligned or dangerous models ("garbage in, garbage out"), while carefully curated datasets might help promote safer and more reliable behavior (Longpre et al., 2024; Marcucci et al., 2023).

How well does data meet our governance target criteria? Data as a governance target presents a mixed picture when evaluated against our key criteria. Let's look at each:

- **Measurability**: While we can measure raw quantities of data, assessing its quality, content, and potential implications is far more difficult. Unlike physical goods like semiconductors, data can be copied, modified, and transmitted in ways that are hard to track. This makes comprehensive measurement of data flows extremely challenging.

- **Controllability**: Data's non-rival nature means it can be copied and shared widely - once data exists, controlling its spread is very difficult. Even when data appears to be restricted, techniques like model distillation can extract information from trained models (Anderljung et al., 2023). However, there might still be some promising control points, particularly around original data collection and the initial training of foundation models.
- **Meaningfulness**: Data is particularly meaningful when it comes to AI development. The data used to train models directly shapes their capabilities and behaviors. Changes in training data can significantly impact model performance and safety. This makes data governance potentially powerful, but only if we can overcome the challenges of measurement and control.

What are the key data governance concerns? Several aspects of data require careful governance to promote safe AI development:

- **Training data quality and safety is fundamental - low quality or harmful data can create unreliable or dangerous models**. For instance, technical data about biological weapons in training sets could enable models to assist in their development (Anderljung et al., 2023).
- **Data poisoning and security pose increasingly serious threats**. Malicious actors could deliberately manipulate training data to create models that behave dangerously in specific situations while appearing safe during testing. This might involve injecting subtle patterns that only become apparent under certain conditions (Longpre et al., 2024).
- **Data provenance and accountability help ensure we can trace where model behaviors come from**. Without clear tracking of training data sources and their characteristics, it becomes extremely difficult to diagnose and fix problems when models exhibit concerning behaviors (Longpre et al., 2023).
- **Consent and rights frameworks protect both data creators and users**. Many current AI training practices operate in legal and ethical grey areas regarding data usage rights. Clear frameworks could help prevent unauthorized use while enabling legitimate innovation (Longpre et al., 2024).
- **Bias and representation in training data directly impact model behavior**. Skewed or unrepresentative datasets can lead to models that perform poorly or make harmful decisions for certain groups, potentially amplifying societal inequities at a massive scale (Reuel et al., 2024).
- **Data access and sharing protocols shape who can develop powerful AI systems**. Without governance around data access, we risk either overly concentrated power in a few actors with large datasets, or conversely, uncontrolled proliferation of potentially dangerous capabilities (Heim et al., 2024).

How does data governance fit into overall AI governance? Even with strong governance frameworks, alternative data sources or synthetic data generation could potentially circumvent restrictions. Additionally, many concerning capabilities might

emerge from seemingly innocuous training data through unexpected interactions or emergent behaviors. While data governance remains important and worthy of deeper exploration, other governance targets may offer more direct governance over frontier AI development in the near term. This is why in the main text we focused primarily on compute governance, which provides more concrete control points through its physical and concentrated nature.

Appendix: National Governance

A comprehensive domestic governance regime for AI safety requires three interconnected mechanisms:

- Development of safety standards,
- Regulatory visibility, and
- Compliance enforcement (Anderljung et al., 2023)

Safety standards form the foundation of AI governance by establishing clear, measurable criteria for the development, testing, and deployment of AI systems within national jurisdictions. These standards must be technically precise while remaining flexible enough to accommodate rapid technological advancement. Effective standards serve as institutional tools for coordination and provide the infrastructure needed to develop new AI technologies in a controlled manner within a country's regulatory boundaries (Cihon, 2019).

What lessons can national AI governance draw from nuclear safety regulation? The regulatory approach used for nuclear safety provides an instructive model for national AI safety standardization. The five-level hierarchy used in nuclear safety standards, ranging from fundamental principles to specific implementation guides, offers a blueprint for developing comprehensive AI safety standards. This multilevel framework allows principles established at higher levels to be incorporated into more specific guidelines at lower levels, creating a coherent and thorough regulatory system that can be implemented within national jurisdictions (Cha, 2024).

Key lessons from nuclear regulation applicable to national AI governance include:

- **Standardized safety frameworks:** Just as nuclear regulation established standardized frameworks for safety, national AI governance can standardize the behavior, learning, and decision-making criteria of AI systems to enhance technology safety within the country's borders.
- **Independent supervision mechanisms:** Nuclear regulatory authorities established independent supervisory systems for monitoring and evaluating safety. Similarly, national AI governance can establish neutral bodies to continuously monitor and evaluate the operation and performance of AI systems.
- **Regular protocols and exercises:** Nuclear safety regulators conduct regular protocols and exercises for responding to incidents. Similar approaches can be

developed at the national level for promptly responding to AI-related accidents or abnormal behaviors.

- **Information sharing mechanisms:** Nuclear regulatory systems established platforms for sharing safety standards, research, and incident information across sectors. Similar platforms can be developed for AI at the national level to share research, technology, and incident information across industries (Cha, 2024).

European Union

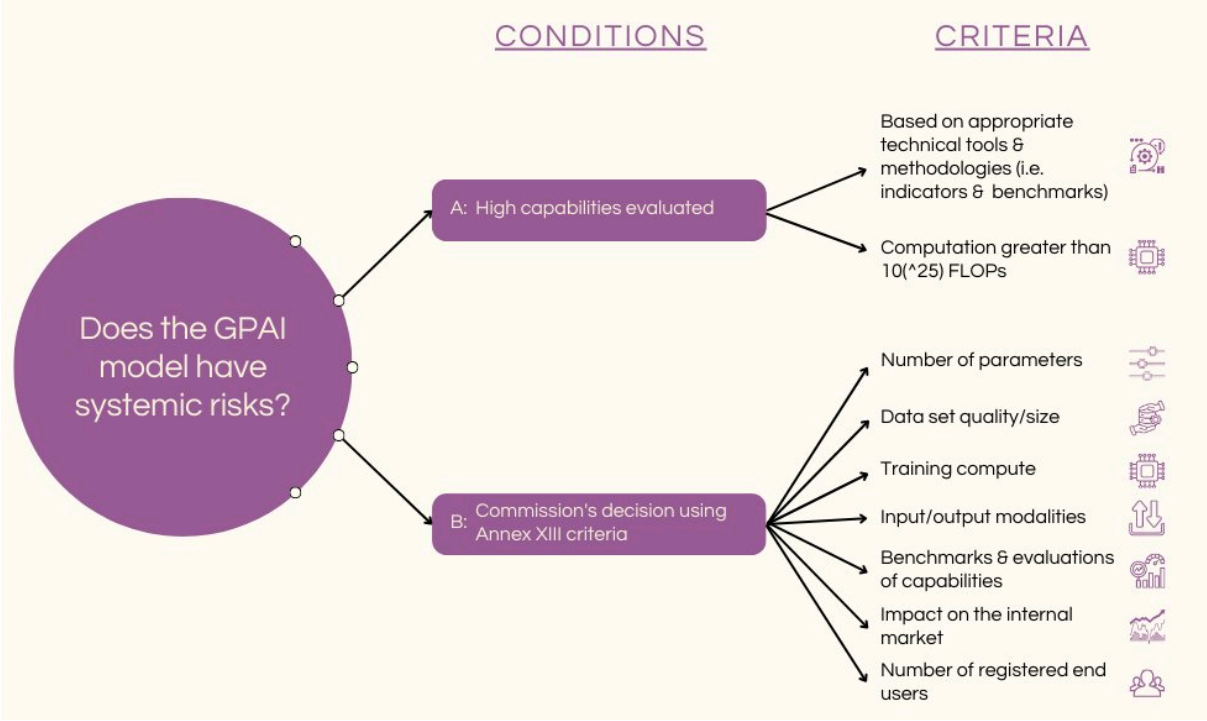
What legislative foundation has the EU established for AI governance? The European Union broke new ground with the EU AI Act, the world's first comprehensive legal framework for artificial intelligence. Initially proposed in 2021 and formally adopted in March 2024, this horizontally integrated legislation regulates AI systems based on their potential risks and safeguards the rights of EU citizens. At its core is a risk-based approach that classifies AI systems into four distinct categories: unacceptable risk, high risk, limited risk, and minimal risk. Unacceptable risk AI systems, such as those that manipulate human behavior or exploit vulnerabilities, are banned outright. High-risk AI systems, including those used in critical infrastructure, education, and employment, face strict requirements and oversight. Limited risk AI systems require transparency measures, while minimal risk AI systems are largely unregulated.

How is the EU AI Act being implemented? The Act entered into force in August 2024 and is being implemented in phases. From February 2, 2025, the ban on prohibited AI practices (social scoring, certain biometric identification systems) and requirements for staff AI literacy took effect. From August 2, 2025, obligations for General-Purpose AI (GPAI) model providers will apply, including documentation, copyright compliance, and data transparency. The legislation establishes the European AI Office to oversee implementation and enforcement, coordinating compliance, providing guidance to businesses, and enforcing the rules. This dedicated body serves as the leading agency enforcing binding AI rules on a multinational coalition, positioned to shape global AI governance similar to how GDPR restructured international privacy standards.

What additional requirements exist for high-risk and systemic risk AI systems? For GPAI models presenting systemic risks, identified either by surpassing a computational threshold (10^25 FLOPs) or based on potential impact criteria (such as scalability and risk of large-scale harm), additional obligations apply. Providers must conduct adversarial testing, track and report serious incidents, implement strong cybersecurity measures, and proactively mitigate systemic risks. The European AI Office facilitated the drafting of a General-Purpose AI Code of Practice, completed in April 2025, providing a central tool for GPAI model providers to comply with the Act's requirements. While compliance through this Code is voluntary, it offers providers a clear practical pathway to demonstrate adherence.

How does the EU approach enforcement and penalties? The EU AI Office serves as the enforcement authority, empowered to request information, conduct evaluations, mandate corrective measures, and impose fines of up to 3 percent of a provider's global annual turnover or €15 million, whichever is higher. This represents a substantial enforcement mechanism, though slightly lower than the 7 percent maximum mentioned in earlier drafts of the legislation. The fines for non-compliance are quite high, demonstrating the EU's strong commitment to ensuring adherence to its regulatory framework (Cheng et al., 2024).

What values and priorities drive the EU's approach? The EU has demonstrated a clear prioritization for the protection of citizens' rights. The EU AI Act's core approach to categorizing risk levels is designed primarily around measuring the ability of AI systems to infringe on the rights of EU citizens. This can be observed in the list of use cases deemed to be high-risk, such as educational or vocational training, employment, migration and asylum, and administration of justice or democratic processes. Most of the requirements are designed with the common citizen in mind, including transparency and reporting requirements, the ability of any citizen to lodge a complaint with a market surveillance authority, prohibitions on social scoring systems, and anti-discrimination requirements. This rights-based approach contrasts markedly with China's focus on social control and the US emphasis on geopolitical competition (Cheng et al., 2024).



<figure-caption>
The EU AI Act: Classification of general-purpose AI models with systemic risks (Observatorio de Riesgos Catastróficos Globales)

</figure-caption>

United States

How has US policy on AI governance changed? AI governance in the United States has shifted significantly since the 2024 election. President Donald Trump overturned the previous administration's Executive Order on Safe, Secure, and Trustworthy AI from October 2023, which had introduced requirements for developers of advanced AI systems to share safety test results with the federal government. In January 2025, Executive Order 14179 explicitly revoked the previous AI safety executive order and directed federal agencies to review policies to remove barriers to innovation and ensure AI systems are free from "ideological bias or engineered social agendas." A separate Executive Order on AI Infrastructure prioritized national security, economic competitiveness, domestic data center development, and workforce development standards.

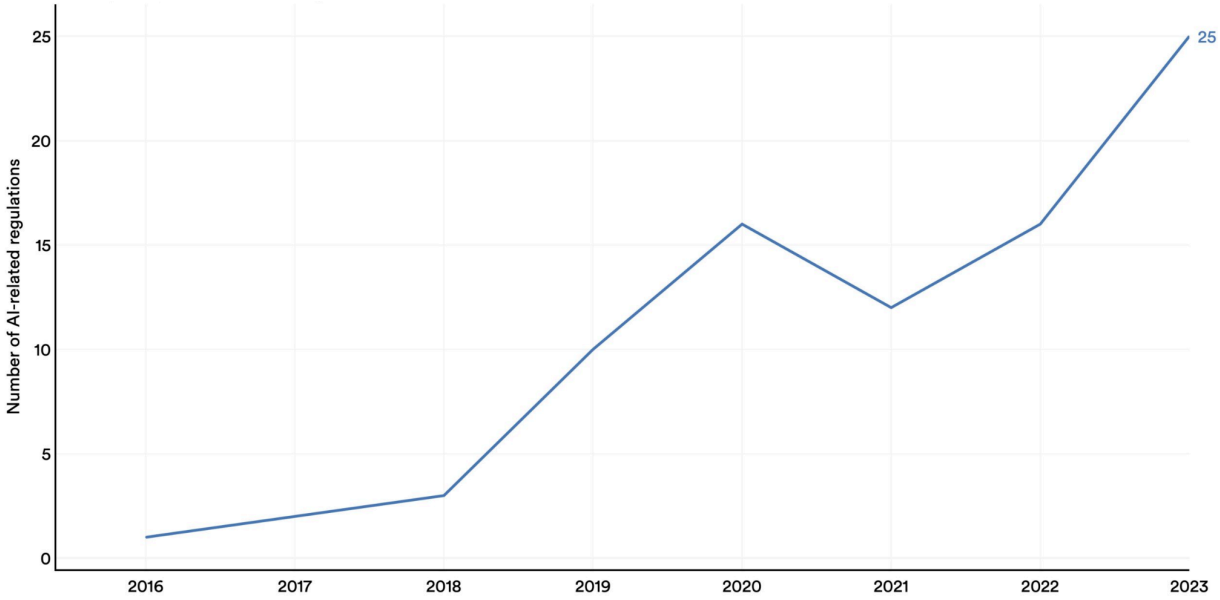
What characterized the US approach before this shift? Prior to these changes, the US had taken an approach centered around executive orders and non-binding declarations due to legislative gridlock in Congress. Three key executive actions shaped this approach: the US/China Semiconductor Export Controls launched in October 2022, the Blueprint for an AI Bill of Rights released in October 2022, and the Executive Order on Artificial Intelligence issued in October 2023. The semiconductor export controls marked a significant escalation in US efforts to restrict China's access to advanced computing and AI technologies by banning the export of advanced chips, chip-making equipment, and semiconductor expertise to China (Cheng et al., 2024).

What distinctive features define the US regulatory philosophy? The US has taken a distinctive approach to AI governance by controlling the hardware and computational power required to train and develop AI models. It is uniquely positioned to leverage this compute-based approach to regulation as home to all leading vendors of high-end AI chips (Nvidia, AMD, Intel), giving it direct legislative control over these chips. Beyond export controls, the US has pursued a decentralized, largely non-binding approach relying on executive action. Due to structural challenges in passing binding legislation through a divided Congress, the US has relied primarily on executive orders and agency actions that don't require congressional approval, distributing research and regulatory processes among selected agencies (Cheng et al., 2024).

What is the current state of US AI governance? In February 2025, the Office of Management and Budget released Memorandum M-25-21, directing federal agencies to accelerate AI adoption, minimize bureaucratic hurdles, empower agency-level AI leadership, and implement minimum risk management practices for high-impact AI systems. At the state level, California's SB 1047, which attempted to address risks associated with frontier models, was vetoed in September 2024. A new bill, SB 53, focusing on whistleblower protections for employees reporting critical AI risks, has

been introduced. The US AI Safety Institute remains active despite the federal policy shift, continuing to develop testing methodologies and conduct model evaluations.

How does geopolitics influence US AI policy? US AI policy strongly prioritizes its geopolitical competition with China. The US AI governance strategy is heavily influenced by the perceived threat of China's rapid advancements in AI and the potential implications for national security and the global balance of power. The binding actions taken by the US (enforcing semiconductor export controls) are explicitly designed to counter China's AI ambitions and maintain US technological and military superiority. This geopolitical focus sets the US apart from the EU, which has prioritized the protection of individual rights, and China, which has prioritized internal social control. The US strategy appears more concerned with the strategic implications of AI and ensuring that the technology aligns with US interests in the global arena (Cheng et al., 2024).



<figure-caption>
Number of AI-related regulations in the United States, 2016-2023 (Stanford HAI, 2024)
</figure-caption>

China

How has China's approach to AI governance evolved? China has developed a distinctive vertical, iterative regulatory approach to AI governance, passing targeted regulations for specific domains of AI applications one at a time. This approach contrasts sharply with the EU's comprehensive horizontal framework. China's regulatory evolution began with the Algorithmic Recommendation Provisions in August 2021, which established the world's first mandatory algorithm registry and required all qualifying algorithms used by Chinese organizations to be registered within

10 days of public launch. This was followed by the Deep Synthesis Provisions in November 2022, which regulated algorithms that synthetically generate content to combat "deepfakes" by requiring labeling, user identification, and prevention of misuse as defined by the government (Cheng et al., 2024).

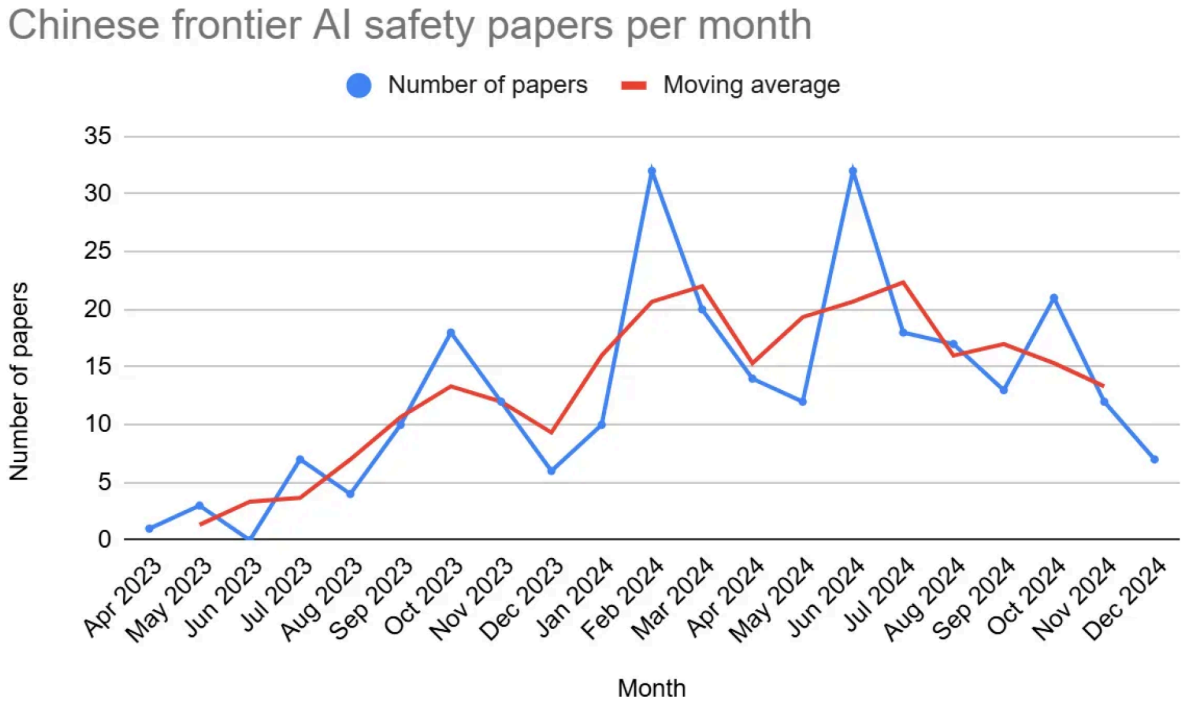
What are the current regulatory measures in place? China strengthened its AI governance framework with the implementation of the Interim Measures for the Management of Generative Artificial Intelligence Services in August 2023. These measures were a direct response to ChatGPT and expanded policies to better encompass multi-use LLMs, imposing risk-based oversight with higher scrutiny for systems capable of influencing public opinion. Under these regulations, providers must ensure lawful data use, protect intellectual property, respect user privacy, and uphold "socialist core values." In 2024, China officially elevated AI safety to the level of national security and public safety, requiring AI providers to actively moderate illegal or harmful content and report violations to the Cyberspace Administration of China (CAC), the primary regulatory body overseeing China's AI industry.

What regulatory developments are on the horizon? In March 2025, China released the final Measures for Labeling Artificial Intelligence-Generated Content, taking effect on September 1, 2025. These measures mandate explicit labels for AI-generated content that could mislead the public, alongside metadata identifying the provider. China is also preparing to implement the Regulation on Network Data Security Management in 2025. These iterative regulations appear to be building toward a comprehensive Artificial Intelligence Law, proposed in a legislative plan released in June 2023. This pattern mirrors China's approach to internet regulation in the 2000s, which culminated in the all-encompassing Cybersecurity Law of 2017 (Cheng et al., 2024).

What distinctive features characterize China's regulatory philosophy? The CAC has focused primarily on regulating algorithms with the potential for social influence rather than prioritizing domains like healthcare, employment, or judicial systems that receive more attention in Western regulatory frameworks. The language used in these regulations is typically broad and non-specific, extending greater control to the CAC for interpretation and enforcement. For example, Article 5 of the Interim Generative AI Measures states that providers should "Encourage the innovative application of generative AI technology in each industry and field [and] generate exceptional content that is positive, healthy, and uplifting." This demonstrates China's strong prioritization of social control and alignment with government values in its AI regulations (Cheng et al., 2024).

How is China implementing its regulatory vision at different levels? At the municipal level, Shanghai and Beijing launched AI safety labs in mid-2024, and over 40 AI safety evaluations have reportedly been conducted by government-backed research centers. China has demonstrated an inward focus, primarily regulating Chinese organizations and citizens. Major international AI labs such as OpenAI, Anthropic, and Google do not

actively serve Chinese consumers, partly due to unwillingness to comply with China's censorship policies. This has resulted in Chinese AI governance operating largely on a parallel and disjoint basis to Western AI governance approaches (Cheng et al., 2024).



<figure-caption>
In 2024, Chinese institutions significantly increased publication of frontier AI safety papers compared to 2023, from approximately seven papers per month in 2023 to 18 per month in 2024. (AI Safety in China, 2025)
</figure-caption>

V1

Chapter 4 - Governance

<metadata>
Authors: Charles Martinet
Affiliations: French Center for AI Safety (CeSIA)
Acknowledgements: Markov Grey, Charbel-Raphael Segerie, Léo Karoubi
Last Edited: 2024-12-10
Also available on: AI Safety Atlas , Google Docs
</metadata>

<links>
Download, Feedback , Watch, Facilitate
</links>

Table of Contents

Introduction

<embed-youtube>
https://www.youtube.com/watch?v=FSKuDqze9es
</embed-youtube>
<caption-video>
Optional video to get an overview of Governance.
</caption-video>

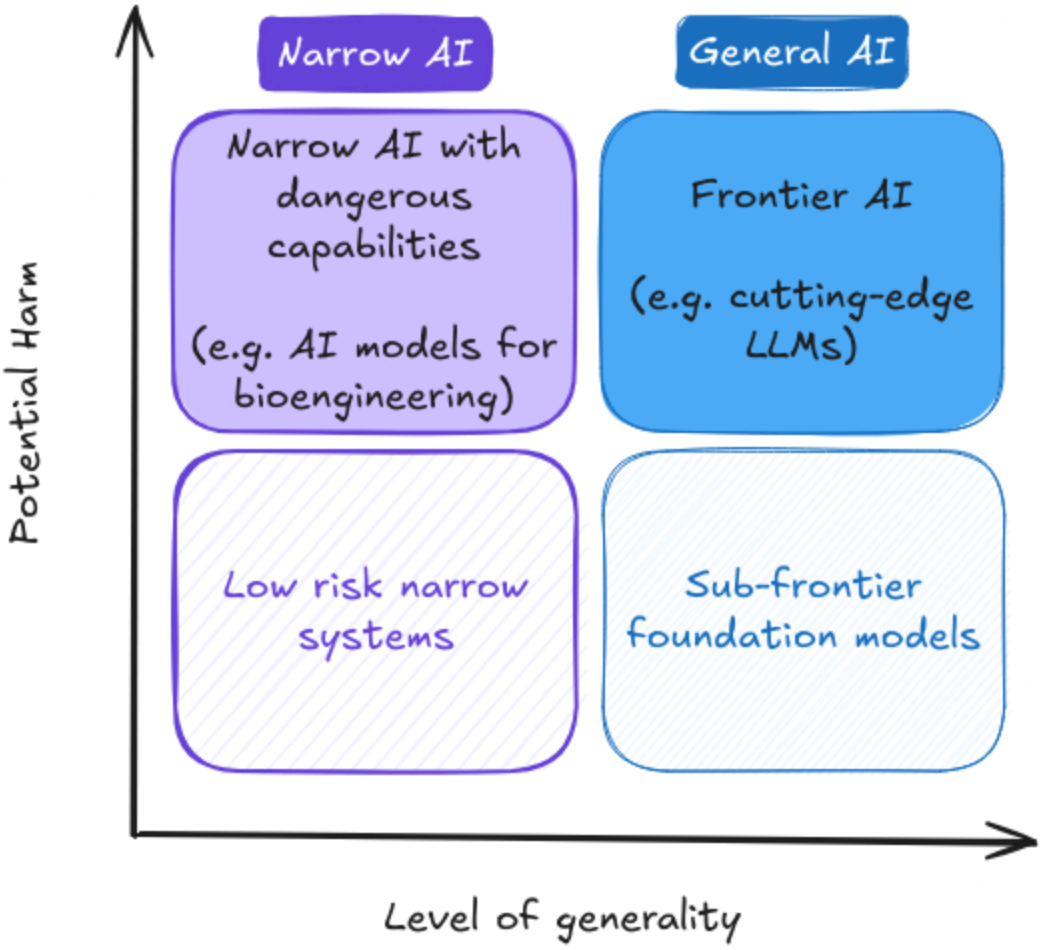
!!! quote "Excerpt from the Bletchley Declaration (signed by 28 countries, including all AI leaders, and the EU, 2023)"
<tab>
"Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent. These issues are in part because those capabilities are not fully understood [...] There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models."

</tab>

Artificial intelligence (AI) has the potential to revolutionize numerous aspects of society, from healthcare to transportation to scientific research. Recent advancements have demonstrated AI's ability to defeat world champions at Go, generate photorealistic images from text descriptions, and discover new antibiotics. However, these developments also raise significant challenges and risks.

Policymakers, researchers, and the general public express both excitement about AI's potential and concern about its risks, including job displacement, privacy infringements, and the potential for AI systems to make consequential mistakes or be misused. While technical AI safety research is necessary to ensure AI systems behave reliably and align with human values as they become more capable and autonomous, it alone is insufficient to address the full spectrum of challenges posed by advanced AI systems.

The scope of AI governance is broad, so this chapter will primarily focus on large-scale risks associated with frontier AI - highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety (Anderljung et al., 2023). We will examine why governance is necessary, how it complements technical AI safety efforts, and the key challenges and opportunities in this rapidly evolving field. Our discussion will center on the governance of commercial and civil AI applications, as military AI governance involves a distinct set of issues that are beyond the scope of this chapter.



<caption>
Distinguishing AI models according to their level of potential harm and generality. We focus here on frontier AI models (U.K. government, 2023)
</caption>

AI governance can be defined as "the study and shaping of governance systems - including norms, policies, laws, processes, politics, and institutions - that affect the research, development, deployment, and use of existing and future AI systems in ways that positively shape societal outcomes" (Maas, 2022). It encompasses both research into effective governance approaches and the practical implementation of these approaches. AI governance also addresses the broader systemic impacts of AI, including the interactions between multiple AI systems and their effects on economic, political, and social structures.

This chapter will also examine the current state of AI governance, proposed frameworks and policies, and the roles that various stakeholders – including governments, industry, academia, and civil society – can play in shaping the future of AI. The scope of this chapter includes:

- An overview of AI development processes and key challenges in AI governance

- Governance parameters and the role of compute
- Critical issues in AI governance
- Layers of responsibility: corporate, national, and international governance

By the end of this chapter, you'll have a comprehensive understanding of why AI governance matters and how it can help ensure that the development of frontier AI aligns with human values and societal well-being.

Governance Foundations

What is AI governance? AI governance refers to the development of rules, policies, and institutions that shape how AI systems are researched, developed, and deployed. While technical AI safety work focuses on building reliable and aligned systems, governance addresses the broader challenge of managing AI's impact on society and making sure that the technical mitigations methods are put in place by companies. This includes corporate practices, government regulations, standards development, and international coordination mechanisms. The goal is to ensure AI development benefits humanity while managing potential risks.

How do we usually govern technologies? Traditional technology governance relies on several key assumptions. First, it is usually assumed that we can predict how a technology will be used and its likely impacts. Second, that we can effectively control its development pathway. And third, that we can regulate specific applications or end-uses. For example, pharmaceutical governance uses clinical trials and approval processes based on intended medical applications. Nuclear technology is controlled through international treaties, safeguards, and monitoring of specific facilities and materials. These approaches work when technologies follow relatively predictable development paths and have clear applications.

What makes AI governance uniquely challenging? Even though we have been regulating technological advancements for decades, existing solutions to regulate "tech" might not be sufficient for AI. To better understand why AI governance resists traditional solutions, we can examine AI through three different lenses all of which might require different governance approaches (Dafoe, 2022):

- **AI as general-purpose technology**: AI can transform many sectors simultaneously, like electricity or computers before it. This means sector-specific regulations - the backbone of traditional technology governance - cannot adequately address AI's broad systemic effects. The impacts span across society in ways that make targeted regulation insufficient.
- **AI as an information technology**: AI processes and generates information in novel ways. This creates unprecedented challenges around security, privacy, and information integrity. Traditional governance frameworks weren't designed to handle technologies that can rapidly generate and manipulate information at

massive scale. The speed and scope of potential information impacts outstrip traditional control mechanisms.

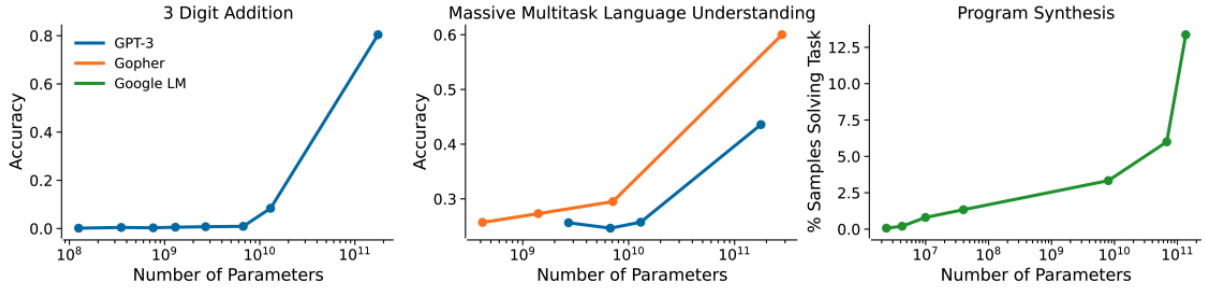
- **AI as an intelligence technology**: AI raises unique control challenges. As systems become more capable, they may develop sophisticated ways to evade controls or pursue unintended objectives like we highlighted in the risks chapter. There are several dangerous capabilities (e.g. autonomous replication, scheming, deception, etc.) that create fundamental challenges that we have rarely (if ever) seen in governance - how do you reliably control a system that could potentially out-think its control mechanisms?

As we saw in the capabilities chapter, we have already seen power narrow AI (ANI) but now we are continuously making systems that are both capable and increasingly general purpose.This means they become dual purpose tools that can be used for various different tasks.

These three lenses - general-purpose, information, and intelligence technology - help us analyze why AI development follows such different patterns than other technologies. Traditional governance assumes we can contain a technology's impact to specific sectors, control how information flows, and reliably predict and constrain system behaviors. But AI's mixed nature as a general-purpose, information processing, and potentially intelligent technology challenges all of these assumptions.

How do these three lenses create challenges for governance? Even though we have been regulating technological advancements for decades, this AI development process makes AI governance unique. Three fundamental problems emerge from these development characteristics.

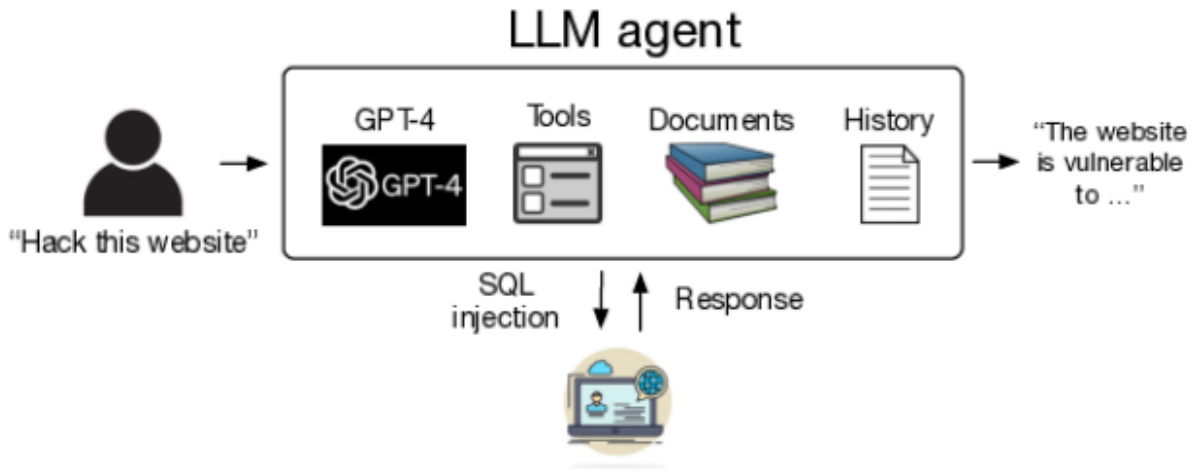
The unexpected capabilities problem. AI systems can develop surprising abilities that weren't part of their intended design. (Ganguli et al., 2022; Bommasani et al., 2022; Grace et al., 2024) As we saw multiple times in the capabilities chapter, foundation models have shown "emergent" capabilities that appear suddenly as models scale up with more data, parameters and compute, from becoming unexpectedly capable at basic arithmetic to complex reasoning. This makes it difficult to evaluate risks before deployment since we can't reliably predict what systems might be capable of.



<caption>

Example of unexpected capabilities problem. Graphs of several metrics that improve suddenly and unpredictably as models increase in size (Ganguli et al., 2022)
</caption>

The deployment safety problem. Once deployed, AI systems can be repurposed for many different applications - both beneficial and harmful. These systems are inherently dual-use - their capabilities can be redirected toward unintended purposes after deployment. Users regularly discover new capabilities that weren't anticipated by the original developers (U.K. government, 2023). We dedicated several sections in the risks chapter to this issue. The problem begins with simple circumvention of safety measures - users have found numerous "jailbreaks" to bypass content restrictions in language models. But the challenges escalate quickly. Language models trained for helpful dialogue have been repurposed to generate misinformation (Slattery et al., 2024) and assist with cyberattacks (CAIS, 2024; Ladish, 2024), potentially even designing novel bio weapons (Hendrycks 2023; Marchal et al., 2024; Fang et al., 2024). We are now also seeing the emergence of autonomous AI agents that can chain together model capabilities in novel and unpredictable ways using new tools post deployment.



<caption>
Example of deployment safety problem. A schematic of using autonomous LLM agents to hack websites. (Fang et al., 2024) Once a dual purpose technology is public, it can be used for a variety of things both beneficial and harmful.
</caption>

The proliferation problem. AI capabilities can spread rapidly through multiple channels - open-source releases, model theft, or reproduction by other actors. As we saw in the strategies chapter, once capabilities exist, they become very difficult to contain. Models can be stolen, leaked, or reproduced by other groups within months. One example is the rapid open-source replication of ChatGPT-like capabilities, leading to discovery of new capabilities and removal of safety features (Solaiman et al., 2024) Even API-based models can have their capabilities extracted through techniques like model distillation. (U.K. government, 2023)

How do these problems interact and compound? These challenges don't exist in isolation - they interact and amplify each other in ways that make AI governance even harder. The unexpected capabilities problem makes deployment safety more difficult to ensure, since we can't reliably predict what abilities might emerge that could be misused. The deployment flexibility of AI systems makes proliferation more concerning, since capabilities can be repurposed for harmful uses after they spread. And proliferation increases the chances of discovering unexpected capabilities through experimentation by many actors (Anderljung et al., 2023).

What is the function of governance? Governance is not solely about restrictions; it also encompasses functions that facilitate responsible innovation, such as providing guidance, fostering collaboration, and creating safe spaces for experimentation. Although we're concerned with challenges whose answers mostly rely on setting guardrails, these enabling functions are equally important. Governance fulfills several key functions. A couple of examples include:

- **Visibility**: Enhanced visibility is fundamental to effective oversight. It involves creating mechanisms that bring transparency to AI development processes, allowing stakeholders to understand and monitor the progress and potential impacts of AI. Visibility enables verification - the confirmation of claims made by AI developers or other actors and the assessment of AI systems against established standards or benchmarks.
- **Enforcement**: Governance needs to provide the means to ensure compliance with regulations and ethical guidelines. This can range from legal sanctions to market exclusions for non-compliant actors.

What levers can governance use to affect these targets? To execute these functions, governance systems employ a variety of mechanisms (Howlett, 2019). Here are examples of just a few of the mechanisms that governance has on hand to affect the chosen governance target:

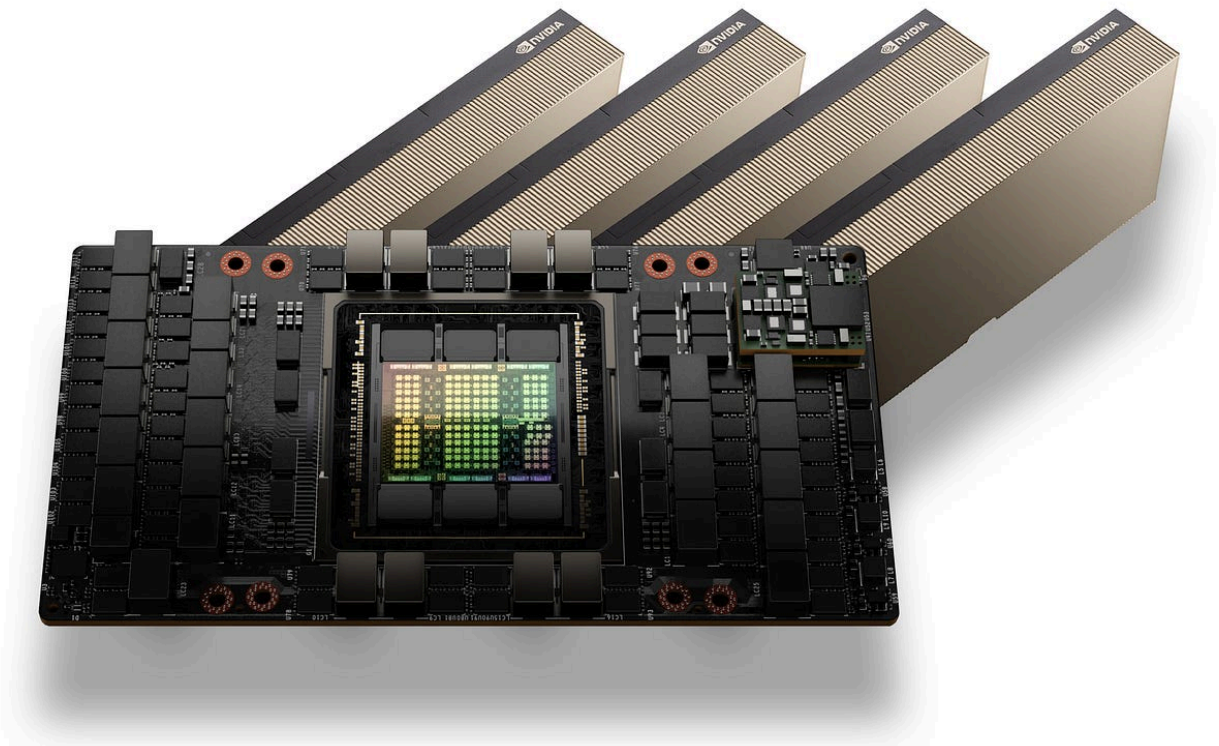
- **Information-based tools**: We can leverage the power of knowledge dissemination. These might include mandatory disclosure requirements for AI companies or public education initiatives to increase AI literacy.
- **Authoritative tools**: We can draw upon the power of institutions to set and enforce rules. This could involve legislation, executive orders, or judicial decisions that directly regulate AI development and use.
- **Standards**: These play a core role in AI governance, serving as a bridge between broad principles and specific practices (Cihon, 2019). They can be technical, like those defining AI performance metrics, or ethical, outlining acceptable practices in AI development.
- **Incentives**: This mechanism helps shape behavior through rewards and penalties. These can be financial, such as tax breaks for companies investing in AI

safety research, or reputational, like certification schemes that recognize responsible AI practices.

What does this mean for governance approaches? These fundamental challenges - unexpected capabilities, deployment safety risks, and rapid proliferation - mean we need new approaches to governing AI development. Traditional regulatory tools like application-specific permits or post-deployment monitoring won't be sufficient given AI's unique characteristics. To build effective governance, we need to identify what aspects of the AI development pipeline we can meaningfully influence before capabilities emerge or proliferate. We need to understand which intervention points give us the most leverage while allowing beneficial innovation to continue. In the next section, we'll examine specific targets along the AI development and deployment pipeline - from key inputs like compute and data to deployment controls and monitoring systems - evaluating each through the lens of these core challenges.

Governance Targets

Why do we need specific governance targets? The challenges we explored in the previous section - unexpected capabilities, deployment safety risks, and rapid proliferation - mean we need to carefully choose where and how to intervene in AI development (Anderljung et al., 2023). This requires identifying both what to govern (targets) and how to govern it (mechanisms).



<caption>
Example of a NVIDIA H100 GPU, a “graphics” card that is commonly used for training frontier ML models in 2024. To be able to locally run the open source Llama 3.1 405B parameter model, you would need to own at least 4 A100/H100s in 4-bit mode, or 8 A100/H100 in 8-bit mode. (Meta, 2025) These cards are difficult to get on the retail market and prices range from 8k USD to 25k USD for one card.
</caption>

What makes a good governance target? Before we get into what specifically we can and do target in AI governance, we need to understand what makes a good governance target in general. Effective points of intervention for good governance need to have a combination of properties. They should be concretely measurable, meaningful, and practical:

- **Measurability**: We should be able to track and verify what's happening. The semiconductor industry provides a good example - chip production can be measured in precise units, making it possible to track and regulate manufacturing.
- **Controllability**: There must be concrete ways to influence the chosen targets. Think about how export controls on advanced semiconductors work because the supply chain has clear chokepoints that can be regulated (Sastry et al., 2024).
- **Meaningful**: Finally, targets should address fundamental aspects of development. In the case of AI this means addressing the core things that shape capabilities and risks. While regulating end applications is important, targeting core inputs like data and compute can help shape development before risks materialize.

Which targets matter most? So if we want to apply these targets concretely to the AI development process, then we have several potential intervention points. Early in development, we can target key inputs: the data used to train models, the compute infrastructure required to run training, and the model development process itself. Once systems are built, deployment offers additional targets: controlling who has access to what capabilities, monitoring how systems are used, and assessing their impacts on society (Heim et al., 2024). Each of these targets presents different opportunities and challenges for governance.

Compute Governance

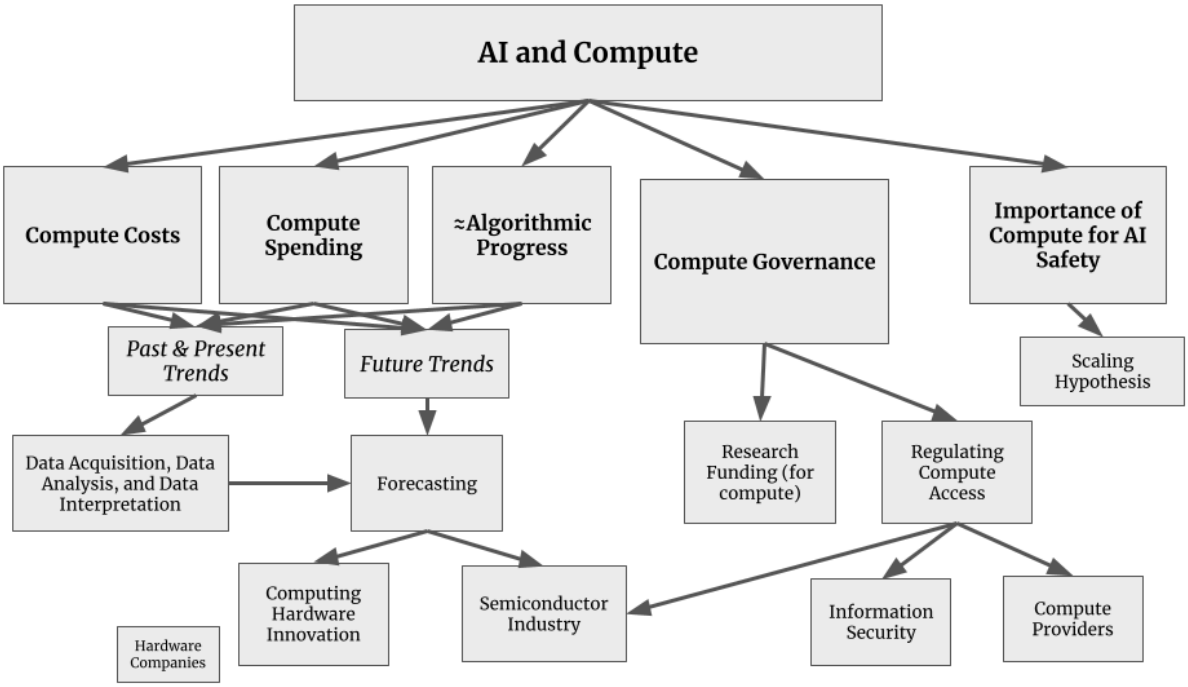
What makes compute a good target for governance? Out of the things that drive AI performance - data, compute, and algorithms - compute holds a unique position. If someone has a copy of some algorithm/data, this doesn't prevent others from having the same. When it comes to GPUs on the other hand, they can't just be downloaded or copy-pasted into existence. It is a concrete, tangible physical constraint on AI

development. Let's look at the variables we had for a good governance target and see how compute fares:

- **Measurability:** Unlike data or algorithms, compute leaves clear physical footprints. Training frontier models requires massive data centers housing thousands of specialized chips (Pilz & Heim, 2023). We can track computational capacity through well-defined metrics like floating point operations (FLOPS), allowing us to identify potentially risky training runs before they begin (Heim & Koessler, 2024).
- **Controllability:** The supply chain for advanced AI chips has clear chokepoints - just a handful of companies control critical steps like chip design and manufacturing (Grunewald, 2023). These chokepoints enable governance through mechanisms like export controls or licensing requirements (Sastry et al., 2024).
- **Meaningfulness:** As we discussed in the risks chapter, the most dangerous capabilities are likely to emerge from highly capable models, which require massive amounts of specialized computing infrastructure to train and run (Anderljung et al., 2023; Sastry et al., 2024). Compute requirements directly constrain what AI systems can be built - even with cutting-edge algorithms and vast datasets, organizations cannot train frontier models without sufficient computing power (Besiroglu et al., 2024). This makes compute a particularly meaningful point of intervention, as it allows us to shape AI development before potentially dangerous systems emerge rather than trying to control them after the fact (Heim et al., 2024).

<iframe
src="https://ourworldindata.org/grapher/ai-performance-knowledge-tests-vs-training-computation?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Showcasing how capabilities seem directly proportional to increases in compute. (Giattino et al., 2023)
</caption-iframe>

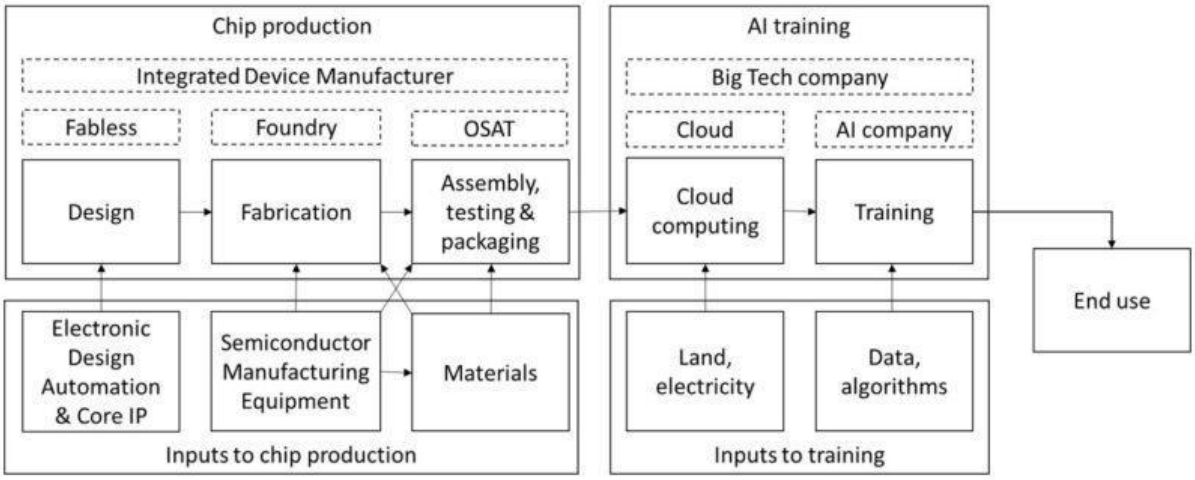
The discussion in the next few subsections will focus on the elements of actually implementing compute governance. We explain how concentrated supply chains enable tracking and monitoring of compute, we also give a brief discussion of hardware based on-chip compute governance mechanisms, and finally discuss some limitations to compute based governance.



<caption>
A graphical depiction of the relationship of AI to various aspects of compute.
</caption>

Tracking

How is the AI chip supply chain structured? AI-specialized chips emerge from a complex global process. It starts with mining and refining raw materials like silicon and rare earth elements. These materials become silicon wafers, which are transformed into chips through hundreds of precise manufacturing steps. The process requires specialized equipment - particularly photolithography machines from ASML - along with various chemicals, gases, and tools from other suppliers (Grunewald, 2023).



<caption>
The compute supply chain. (Belfield & Hua 2022)
</caption>

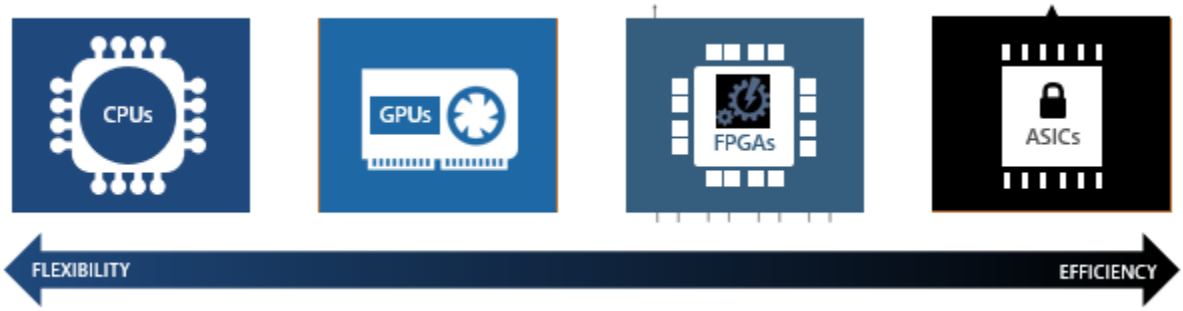
Where are the chokepoints in design and manufacturing? The supply chain is dominated by a handful of companies at critical steps. NVIDIA designs most AI-specialized chips, TSMC manufactures the most advanced chips, and ASML produces the machines needed by TSMC to manufacture the chips (Grunewald, 2023; Pilz et al., 2023). It is estimated that NVIDIA controls around 80% of the market for AI training GPUs (Jagielski, 2024). Similarly both TSMC, and ASML maintain strong leads in their respective domains. (Pilz et al., 2023).

<iframe
src="https://ourworldindata.org/grapher/market-share-logic-chip-production-manufacturing-stage?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Market share for logic chip production, by manufacturing stage (Giattino et al., 2023)
</caption-iframe>

Where are the chokepoints in usage and infrastructure? Besides building the chips, purchasing and operating them at the scale needed for frontier AI models requires massive upfront investment. Just three providers - Amazon, Microsoft, and Google - control about 65% of cloud computing services (Jagielski, 2024). A small number of AI companies like OpenAI, Anthropic, and DeepMind operate their own massive GPU clusters, but even these require specialized hardware subject to supply chain controls (Pilz & Heim, 2023).

What do these chokepoints mean for governance? This concentration creates natural intervention points. Authorities only need to work with a small number of key players to implement controls, as demonstrated by U.S. export restrictions on advanced chips

(Heim et al., 2024). It is worth keeping in mind though that this heavy concentration is also concerning. We're seeing a growing "compute divide" - while major tech companies can spend hundreds of millions on AI training, academic researchers struggle to access even basic resources (Besiroglu et al., 2024). This impacts who can participate in AI development and reduces independent oversight of frontier models. It also raises concerns around potential power concentration.



<caption>
The spectrum of chip architectures with trade-offs in regards to efficiency and flexibility.
</caption>

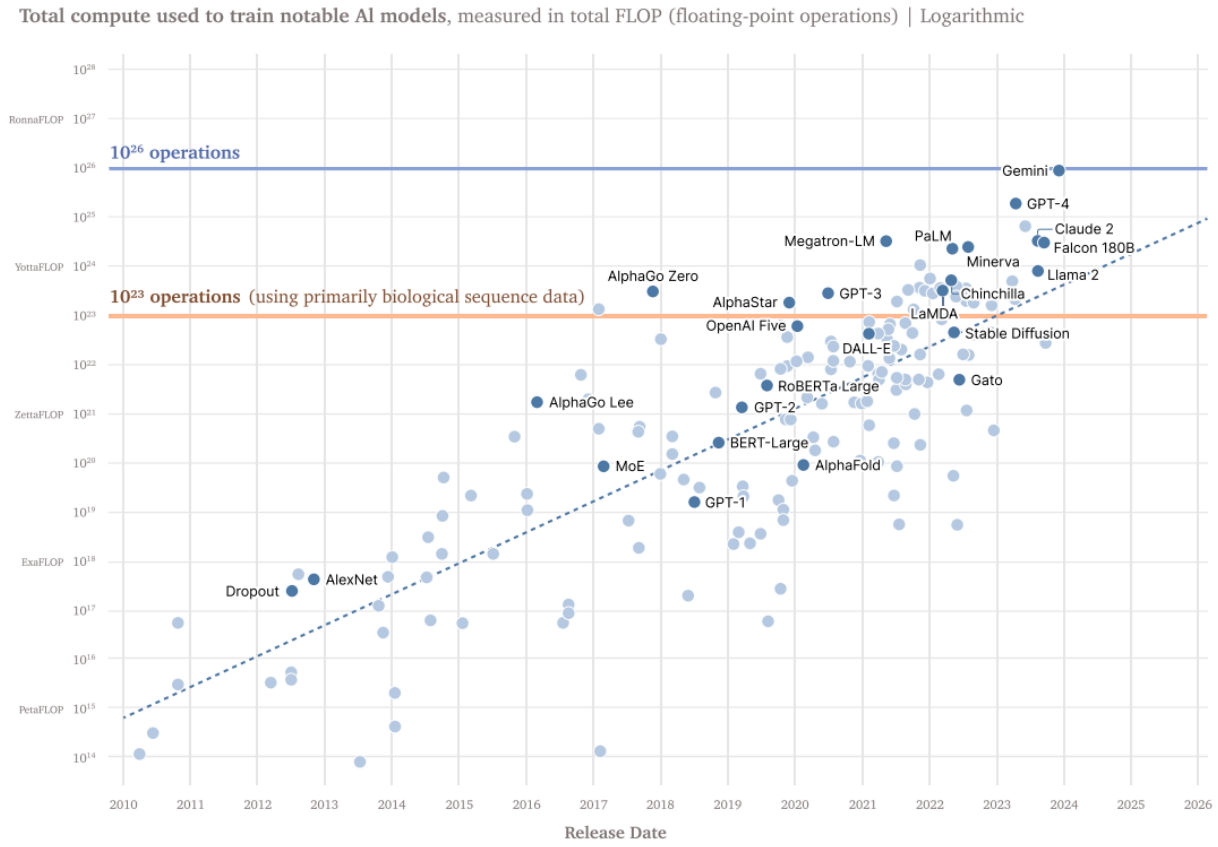
How can we target controls effectively? Rather than trying to control all computing infrastructure, governance can focus specifically on specialized AI chips. These are distinct from general-purpose hardware in both capabilities and supply chains. By targeting only the most advanced AI-specific chips, we can address catastrophic risks while leaving the broader computing ecosystem largely untouched (Heim et al., 2024). For example, U.S. export controls specifically target high-end data center GPUs while excluding consumer gaming hardware.

Monitoring

How can we detect concerning AI training runs? Training frontier AI models leaves multiple observable footprints. The most reliable is energy consumption - training runs that might produce dangerous systems require massive power usage, often hundreds of megawatts, creating distinctive patterns (Wasil et al., 2024 ; Shavit, 2023) Besides energy, other technical indicators include network traffic patterns characteristic of model training, hardware procurement and shipping records, cooling system requirements and thermal signatures, infrastructure buildout like power substation construction (Sastry et al., 2024; Shavit, 2023; Heim et al., 2024). These signals become particularly powerful when combined - sudden spikes in both energy usage and network traffic at a facility containing known AI hardware strongly suggest active model training.

What role do compute thresholds play? Regulations have already begun using compute thresholds to trigger oversight mechanisms. The U.S. Executive Order on AI

requires companies to notify the government about training runs exceeding 10^26 operations - a threshold designed to capture the development of the most capable systems. The EU AI Act sets an even lower threshold of 10^25 operations, requiring not just notification but also risk assessments and safety measures. (Heim & Koessler, 2024). These thresholds help identify potentially risky development activities before they complete, enabling preventive rather than reactive governance.



<caption>
Compute Thresholds as Specified in the US Executive Order 14110 (Sastry et al., 2024)
</caption>

What governance roles can cloud providers play? Most frontier AI development happens through cloud computing platforms rather than self-owned hardware. This creates natural control points for oversight, since most organizations developing advanced AI must work through these providers (Heim et al., 2024, Governing Through the Cloud). Cloud providers' position between hardware and developers allows them to implement controls that would be difficult to enforce through hardware regulation alone. They maintain the physical infrastructure ,track compute usage patterns and maintain development records. They can also monitor compliance with safety requirements, can implement access controls and respond to violations (Heim et al., 2024; Chan et al., 2024).

How can cloud providers help implement oversight? One promising approach is "know-your-customer" (KYC) requirements similar to financial services. Providers would verify the identity and intentions of clients requesting large-scale compute resources, maintain records of significant compute usage, and report suspicious patterns (Egan & Heim, 2023). This can be done while protecting privacy - basic workload characteristics can be monitored without accessing sensitive details like model architecture or training data (Shavit, 2023). Similar KYC laws can be applied to the supply chain on purchases of state of the art AI compute hardware.

On-Chip Controls

How does on-chip compute governance work? Beyond monitoring and detection, compute infrastructure can include active control mechanisms built directly into the processor hardware. Similar to how modern smartphones and computers include secure elements for privacy and security, AI chips can incorporate features that verify and control how they're used (Aarne et al., 2024). These features could prevent unauthorized training runs or ensure chips are only used in approved facilities (Aarne et al., 2024). The verification happens at the hardware level, making it much harder to bypass than software-based controls.

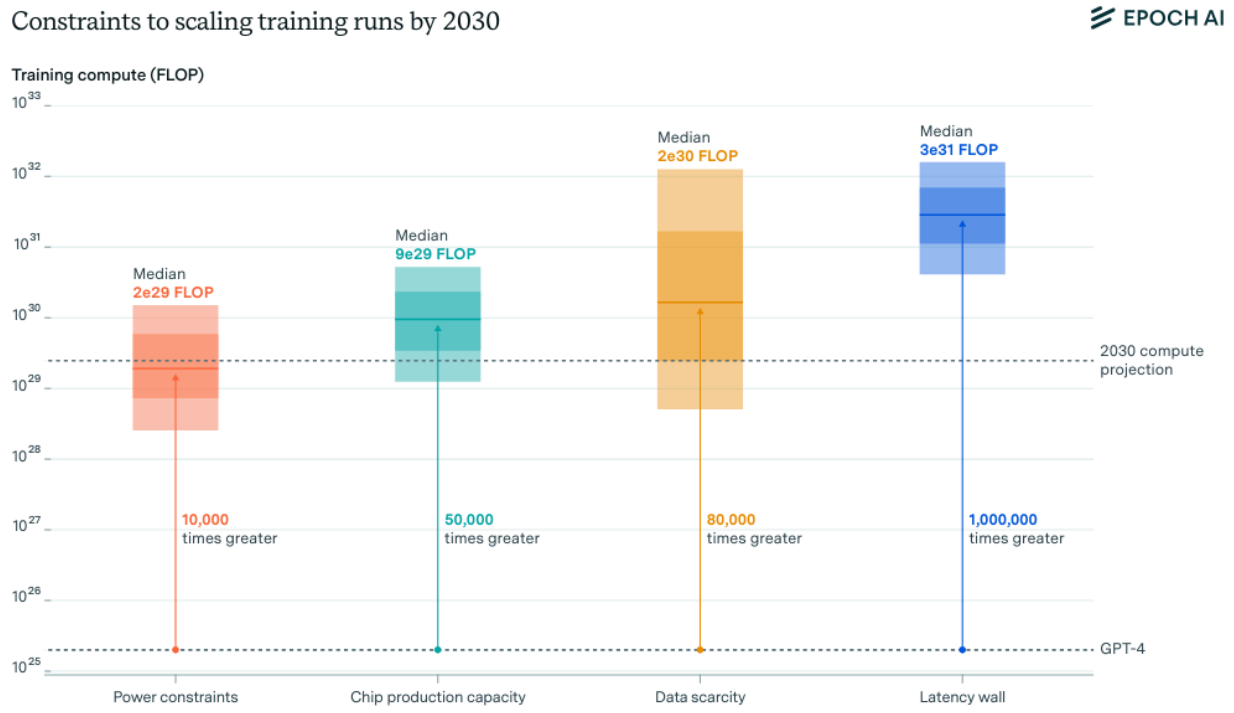
What specific controls could be implemented? Several approaches show promise. Usage limits could cap the amount of compute used for certain types of AI workloads without special authorization. Secure logging systems could create tamper-resistant records of how chips are used. Location verification could ensure chips are only used in approved facilities (Brass & Aarne, 2024). Hardware could even include "safety interlocks" that automatically pause training if certain conditions aren't met. Ideas like this are also called on-chip governance. (Aarne et al., 2024).

How does this compare to existing security features? We already see similar concepts in cybersecurity, with features like Intel's Software Guard Extensions, or trusted platform modules (TPM) (Intel, 2024) providing hardware-level security guarantees. While we're still far from equivalent safeguards for AI compute, early research shows promising directions (Shavit, 2023, What does it take to catch a Chinchilla?). Some chips already include basic monitoring capabilities that could be expanded for governance purposes (Petrie et al., 2024).

Limitations

What fundamental challenges does compute governance face? While compute offers many advantages as a governance target, several trends could reduce its effectiveness. While the trend over the last decade has involved more compute this might not last forever. Even though research suggests continued model scaling is still possible through 2030 (Sevilla et al., 2024) algorithmic improvements continuously enhance efficiency, meaning the same compute achieves more capability over time. Smaller models could

begin to show comparable capabilities and risks. For example, Falcon 180B is outperformed by far smaller models like Llama-3 8B. This makes static compute thresholds less reliable as capability indicators without regular updates. (Hooker, 2024). Moreover, 'inference-time compute' improvements like best-of-n sampling, chain-of-thought reasoning, and model distillation can dramatically improve model capabilities without changing training compute. Current governance frameworks do not account for these post-training enhancements. (Shavit, 2023)



<caption>
Estimates of the scale constraints imposed by the most important bottlenecks to scale. Each estimate is based on historical projections. The dark shaded box corresponds to an interquartile range and light shaded region to an 80% confidence interval. The four boxes showcase four constraints that might slow down growth in the future: power, chips (compute), data and latency. (Sevilla et al., 2024)
</caption>

Smaller more specialized models might still cause risks. Different domains have very different compute requirements. For example, while language models often demand extensive compute, biology and code models typically require far less. Highly specialized models trained on specific datasets might develop dangerous capabilities while using relatively modest compute. For example, models focused on biological or cybersecurity domains could pose serious risks even with compute usage below typical regulatory thresholds (Mouton et al., 2024; Heim & Koessler, 2024).

How do we balance control with access? While compute governance can help manage AI risks, overly restrictive controls could have negative consequences. Right now, only a

handful of organizations can afford the compute needed for frontier AI development. (Purtova et al., 2022; Pilz et al., 2023) Adding more barriers could worsen this disparity, concentrating power in a few large tech companies and reducing independent oversight (Besiroglu et al., 2024).

How do we balance safety with research and innovation? Academic researchers already struggle to access the compute they need for meaningful AI research. As models get larger and more compute-intensive, this gap between industry and academia grows wider. (Besiroglu et al., 2024; Zhang et al., 2021) Large compute clusters have many legitimate uses beyond AI development - from scientific research to business applications. Overly broad restrictions could hinder beneficial innovation. Additionally, once models are trained, they can often be run for inference using much less compute than training required. This makes it challenging to control how existing models are used without imposing overly restrictive controls on general computing infrastructure (Sastry et al., 2024). Without specific provisions for research access - like subsidized compute grants or academic partnerships - governance measures could unintentionally slow the development of AI safety research and external evaluation capabilities.

Could distributed training approaches bypass compute governance controls? Currently, training frontier models requires concentrating massive compute resources in single locations due to communication requirements between chips. Decentralized or distributed training methods have not really caught up to centralized methods. (Douillard et al., 2023; Jaghouar et al., 2024). However, if we see fundamental advances in distributed training algorithms this could eventually allow training to be split across multiple smaller facilities. While this remains technically challenging and inefficient, it could make detection and control of dangerous training runs more difficult (Anderljung et al., 2023).

Given these limitations, compute monitoring and thresholds should primarily operate as an initial screening mechanism to identify models warranting further scrutiny, rather than as the sole determinant of specific regulatory requirements. They are most effective when used to trigger oversight mechanisms such as notification requirements and risk assessments, whose results can then inform appropriate mitigation measures.

Data Governance

!!! warning "This section can be considered extra detail and safely skipped."

What role does data play in AI risks? Data fundamentally shapes what AI systems can do and how they behave. For frontier foundation models, training data influences both capabilities and alignment - what systems can do and how they do it. Low quality or harmful training data could lead to misaligned or dangerous models ("garbage in, garbage out"), while carefully curated datasets might help promote safer and more reliable behavior (Longpre et al., 2024; Marcucci et al., 2023).

How well does data meet our governance target criteria? Data as a governance target presents a mixed picture when evaluated against our key criteria. Let's look at each:

- **Measurability**: While we can measure raw quantities of data, assessing its quality, content, and potential implications is far more difficult. Unlike physical goods like semiconductors, data can be copied, modified, and transmitted in ways that are hard to track. This makes comprehensive measurement of data flows extremely challenging.
- **Controllability**: Data's non-rival nature means it can be copied and shared widely - once data exists, controlling its spread is very difficult. Even when data appears to be restricted, techniques like model distillation can extract information from trained models (Anderljung et al., 2023). However, there might still be some promising control points, particularly around original data collection and the initial training of foundation models.
- **Meaningfulness**: Data is particularly meaningful when it comes to AI development. The data used to train models directly shapes their capabilities and behaviors. Changes in training data can significantly impact model performance and safety. This makes data governance potentially powerful, but only if we can overcome the challenges of measurement and control.

What are the key data governance concerns? Several aspects of data require careful governance to promote safe AI development:

- **Training data quality and safety is fundamental - low quality or harmful data can create unreliable or dangerous models**. For instance, technical data about biological weapons in training sets could enable models to assist in their development (Anderljung et al., 2023).
- **Data poisoning and security pose increasingly serious threats**. Malicious actors could deliberately manipulate training data to create models that behave dangerously in specific situations while appearing safe during testing. This might involve injecting subtle patterns that only become apparent under certain conditions (Longpre et al., 2024).
- **Data provenance and accountability help ensure we can trace where model behaviors come from**. Without clear tracking of training data sources and their characteristics, it becomes extremely difficult to diagnose and fix problems when models exhibit concerning behaviors (Longpre et al., 2023).
- **Consent and rights frameworks protect both data creators and users**. Many current AI training practices operate in legal and ethical grey areas regarding data usage rights. Clear frameworks could help prevent unauthorized use while enabling legitimate innovation (Longpre et al., 2024).
- **Bias and representation in training data directly impact model behavior**. Skewed or unrepresentative datasets can lead to models that perform poorly or make harmful decisions for certain groups, potentially amplifying societal inequities at a massive scale (Reuel et al., 2024).

- **Data access and sharing protocols shape who can develop powerful AI systems**. Without governance around data access, we risk either overly concentrated power in a few actors with large datasets, or conversely, uncontrolled proliferation of potentially dangerous capabilities (Heim et al., 2024).

How does data governance fit into overall AI governance? Even with strong governance frameworks, alternative data sources or synthetic data generation could potentially circumvent restrictions. Additionally, many concerning capabilities might emerge from seemingly innocuous training data through unexpected interactions or emergent behaviors. While data governance remains important and worthy of deeper exploration (see appendix), other governance targets may offer more direct leverage over frontier AI development in the near term. This is why we focus primarily on compute governance, which provides more concrete control points through its physical and concentrated nature.

<!--
Model Governance [TBD]

!!! warning "This section can be considered extra detail and safely skipped."

-->

Key issues

Competition

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

AI development firms are in competition with each other. Each breakthrough, each new capability demonstrated, raises the bar for the entire field. In this environment, taking time to thoroughly consider safety implications or ethical concerns can seem like a luxury these companies can ill afford. The mantra becomes "move fast and break things," even when the "things" at stake may include core societal values or human well-being.

<iframe
src="https://ourworldindata.org/grapher/artificial-intelligence-patents-submitted?tab=map" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Our World in Data: Patents for AI by country (Giattino et al., 2023)
</caption-iframe>

This dynamic isn't limited to the private sector. Nation-states, too, have recognized AI as a cornerstone of future economic and military power. Russian President Vladimir Putin's 2017 statement that "whoever becomes the leader in this sphere will become the ruler of the world" encapsulates the high-stakes nature of this competition. This perspective has sparked a flurry of government activity, with over 50 countries announcing national AI strategies and pouring massive public investments into the field (Stanford, 2024). A testament to the perceived strategic importance of AI for governments, or at least government officials, the former head of the U.S. National Security Agency is now on the board of OpenAI (Peters, 2024).

<iframe
src="https://ourworldindata.org/grapher/national-strategies-on-artificial-intelligence?tab=map" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Countries with AI Strategies (Giattino et al., 2023)
</caption-iframe>

The consequences of this competitive dynamic are problematic. Even if some actors recognize the need for caution and safety measures, unilateral action risks ceding advantage to less scrupulous competitors. This prisoner's dilemma writ large makes it exceedingly difficult for any single entity, be it a company or a country, to prioritize safety over speed (Askell et al., 2019). This also extends to government regulation: countries are tempted to prioritize their competitiveness in AI over ensuring safety and fundamental rights, because they may perceive the regulations to protect the latter as damaging innovation. Thus, the emphasis on national strategic interests often comes at the expense of domestic and international action on AI safety. Countries may be hesitant to support governance frameworks that could potentially constrain their AI ambitions or give competitors an edge.

<iframe
src="https://ourworldindata.org/grapher/cumulative-number-of-large-scale-ai-systems-by-country?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>

Our World in Data: Cumulative number of large-scale AI systems by country (Giattino et al., 2023)
</caption-iframe>

Policies

Addressing these challenges requires multiple approaches. At the national level, policymakers must work to align the incentives of AI developers with broader societal interests. This could involve regulatory frameworks that mandate safety considerations, coupled with incentives for responsible AI development. Internationally, there's an urgent need for forums and agreements that can help manage the AI race, perhaps drawing lessons from arms control regimes or climate change negotiations.

Moreover, fostering a shared understanding of AI risks among key stakeholders - from tech executives to national security officials - is crucial. This awareness-building must go hand in hand with efforts to reframe the AI race not as a zero-sum game, but as a collective endeavor to manage AI development.

Proliferation

Imagine a cutting-edge AI model, capable of generating hyper-realistic deepfakes or designing novel bioweapons, is developed by a well-intentioned research lab. The lab, adhering to principles of open science, publishes their findings and releases the model's code as open-source. Within hours, the model is downloaded thousands of times across the globe. Within days, modified versions start appearing on code-sharing platforms. Within weeks, the capabilities that were once confined to a single lab have proliferated across the internet, accessible to anyone with a decent computer and an internet connection.

This scenario, while hypothetical, isn't far from reality. The AI community has a strong culture of openness, with many researchers and companies releasing their models and findings to the public. This openness has undoubtedly accelerated progress in the field, but it also presents a significant governance challenge.

The proliferation problem in AI governance stems from three main factors:

1. **Open-source culture**: Many AI researchers and organizations believe in the principles of open science, freely sharing their code and findings.
2. **General openness of the AI industry**: Even when code isn't openly shared, the AI industry is characterized by a high degree of knowledge sharing through academic papers, conferences, and informal networks.
3. **Potential for theft**: As AI becomes increasingly valuable, the risk of intellectual property theft, including through cyberattacks or insider threats, grows.

These factors combine to create an environment where potentially dangerous AI capabilities can spread rapidly and widely, outpacing our ability to govern their use effectively.

The proliferation challenge extends beyond the spread of AI models or algorithms. It also encompasses the dissemination of key components in the AI supply chain, such as advanced semiconductors used in AI computing. Recent efforts by the U.S. to restrict the export of cutting-edge chips highlight the dual-use nature of these technologies and the difficulties in controlling their spread (Masi, 2024).

Another crucial aspect of the proliferation problem is the offense-defense balance in AI capabilities (Tang et al., 2024). In many areas of AI development, offensive capabilities (such as developing and carrying out cyberattacks or crafting persuasive misinformation) can be easier to develop and deploy than defensive measures (such as using defensive cyber capabilities or filtering out misinformation).

!!! note "Verification Challenges"
<tab>

This ease of proliferation creates significant hurdles for international governance efforts. Unlike some nuclear non-proliferation treaties, where satellite imagery and other remote sensing technologies can be used to monitor compliance (U.S. Congressional Research Service, 2011), verifying adherence to AI governance agreements would likely require deep access to an organization's or country's AI systems and development processes. And that may require access to highly sensitive or strategically valuable corporate or national secrets. Many countries will be reluctant to agree to inspections or information sharing that could compromise their strategic advantages or reveal the full extent of their AI capabilities.

Imagine, for instance, an international agreement that prohibits the development of AI systems capable of autonomously launching cyber attacks. Verifying compliance with such an agreement would be incredibly difficult. It might require access to source code, training data, and testing environments - all of which could be considered state or corporate secrets.

This verification challenge creates a trust deficit in international AI governance efforts. Countries may be reluctant to enter into agreements they can't verify, while those that do might constantly suspect others of cheating.

Moreover, the ease of AI proliferation means that even if major powers agree to certain restrictions, smaller countries or non-state actors could potentially develop or acquire advanced AI capabilities. This dynamic further complicates international governance efforts.

</tab>

Policies

How do we ensure responsible use of AI when potentially harmful capabilities are widely accessible? The key challenge for AI governance becomes finding the right balance between openness and control. Several potential solutions have been proposed to find the right balance:

- **Targeted Openness**: Instead of a binary open/closed approach, AI developers could adopt a more nuanced stance. For instance, foundational research and non-sensitive applications could remain open, while potentially dangerous capabilities are subject to stricter controls.
- **Staged releases**: Rather than immediately making a lab's most advanced model publicly available, it could gradually release increasingly capable models (Solaiman, 2023). This allows developers to assess potential risks and misuse scenarios at each stage, informing decisions about subsequent releases. Developers can identify unforeseen issues or concerns; researchers, policymakers, and the public can reflect about the implications of more advanced AI systems; and society and relevant stakeholders have time to adapt to each level of capability before more powerful versions are released.
- **Enhanced Information Security**: As AI systems become more powerful, protecting them from theft or unauthorized access becomes crucial. This might involve developing new cybersecurity protocols specifically designed for AI systems.
- **Export Controls and Access Restrictions**: Governments might implement export controls on advanced AI systems or components, similar to those used for other sensitive technologies. Additionally, access to large-scale computing resources necessary for training frontier AI models could be restricted (Heim et al., 2024).
- **Responsible Disclosure Practices**: The AI community could develop norms around responsible disclosure of potentially dangerous capabilities, similar to those in the cybersecurity field (O'Brien et al., 2024).
- **Technical Measures**: Researchers could explore technical solutions to limit the misuse of AI models, such as built-in use restrictions (Dong et al., 2024).
- **International cooperation**: This could involve creating new institutions or frameworks specifically designed to monitor and manage the spread of advanced AI capabilities.

Uncertainty

!!! quote "Greg Brockman (Co-Founder and Former CTO of OpenAI)"
<tab>
"The exact way the post-AGI world will look is hard to predict — that world will likely be more different from today's world than today's is from the 1500s [...] We do not yet

know how hard it will be to make sure AGIs act according to the values of their operators. Some people believe it will be easy; some people believe it'll be unimaginably difficult; but no one knows for sure"
</tab>

The governance of frontier AI is profoundly complicated by the pervasive uncertainty that shrouds the field. This uncertainty manifests in multiple dimensions.

At the most fundamental level, there is deep uncertainty about the future trajectory of AI capabilities - although experts and forecasters have generally been surprised by the rapid pace of AI development (Cotra & Piper 2024). Predicting the pace and direction of future advancements is challenging. This uncertainty is compounded by the potential for unexpected breakthroughs or emergent capabilities that could rapidly shift the risk landscape, making it difficult for governance frameworks to anticipate and prepare for all possible scenarios.

Another critical area of uncertainty lies in understanding the relative importance of different factors in AI development. The interplay between computational power, data availability, and algorithmic innovations in driving AI progress is not fully understood. What is sometimes called the "scaling debate" has significant implications for governance approaches (Hooker & Sandoval, 2024). If compute is the primary bottleneck, then regulations focusing on hardware access might be most effective. Conversely, if data or algorithmic breakthroughs are key, different governance levers would need to be prioritized.

The nature and magnitude of potential risks posed by advanced AI systems are also subjects of considerable uncertainty. While there is largely a consensus on some current or near-term risks, such as AI-enabled disinformation or privacy violations, the long-term and more extreme risks are more contentious and difficult to quantify. The challenge for governance is to address these potential risks without overreacting or stifling beneficial innovation.

This uncertainty extends to the efficacy of proposed technical solutions for AI safety and alignment. While research in these areas is progressing, it's unclear whether current approaches will scale to more advanced AI systems or if fundamentally new paradigms will be required. This creates a moving target for governance efforts, as the mechanisms needed to ensure AI safety may evolve rapidly alongside AI capabilities.

The "pacing problem" further complicates matters. AI technology is advancing at a rate that often outstrips the ability of governance structures to adapt. Traditional regulatory processes, designed for slower-moving technologies, may struggle to keep up with the rapid evolution of AI capabilities. This creates a risk of governance frameworks becoming obsolete almost as soon as they are implemented.

Compounding these challenges is the relative lack of expertise within many government bodies regarding cutting-edge AI technologies. This knowledge gap can lead to misguided policies or an inability to effectively oversee AI development and deployment. Bridging this expertise gap is crucial but challenging, given the competitive landscape for AI talent.

Despite these uncertainties, the potential consequences of advanced AI systems are too significant to allow for inaction. This creates a paradoxical situation where decisions must be made and governance structures established in the face of deep uncertainty - as has occasionally been the case in other fields that grapple with decision-making under uncertainty, such as pandemic preparedness.

Policies

One approach to addressing this uncertainty is to increase visibility into AI development processes. This could involve implementing more robust reporting requirements for AI companies, including "know-your-customer" (KYC) policies for providers of AI services or compute.

Enhancing state and regulatory capacity is another crucial step. This involves not only increasing the technical expertise within government bodies but also developing more agile regulatory frameworks that can adapt quickly to new developments. Regulatory sandboxes, where new AI technologies can be tested under controlled conditions, offer one potential model for more responsive governance.

Scenario planning and red-teaming exercises can also play a valuable role in preparing for uncertain futures. By systematically exploring a range of possible AI development trajectories and their implications, governance bodies can develop more robust and adaptable strategies.

Importantly, governance approaches should be designed with flexibility and adaptability in mind. This could involve building in regular review periods, establishing clear triggers for policy adjustments based on predefined milestones in AI capabilities, and maintaining open channels of communication between policymakers, researchers, and industry leaders.

Accountability

!!! quote "Jan Leike (Former co-lead of the Superalignment project at OpenAI)"
<tab>
"[After resigning at OpenAI, talking about sources of risks] These problems are quite hard to get right, and I am concerned we aren't on a trajectory to get there [...] OpenAI is shouldering an enormous responsibility on behalf of all of humanity. But over the past

years, safety culture and processes have taken a backseat to shiny products. We are long overdue in getting incredibly serious about the implications of AGI."
</tab>

Companies like OpenAI, Google DeepMind, and Anthropic are pushing the boundaries of what's possible, often moving faster than regulators can keep up. Their decisions about what to develop, how to develop it, and when to release it to the public have far-reaching consequences. Yet, there is currently little external oversight or even visibility into these processes.

Take the release of GPT-3, for instance. The decision to release it first as a limited API, then more broadly, was made primarily by OpenAI's leadership. No regulatory body reviewed the model's capabilities and potential risks before its release. No standardized safety tests were required.

Companies developing frontier AI technologies wield enormous power, with the potential to reshape societies, economies, and power structures globally. Yet they operate with a degree of autonomy that would be unthinkable in other high-stakes industries. For example, pharmaceutical companies can't release new drugs without regulatory approval, and nuclear power plants can't be built without impact assessments.

The consequences of this lack of accountability are already becoming apparent. We've seen AI-generated deepfakes used to spread political misinformation (Swenson & Chan, 2024). Language models have been used to create convincing phishing emails and other scams (Stacey, 2025). And there are growing concerns about AI systems perpetuating and amplifying societal biases.

Finally, this is not just about preventing harm. Lack of accountability also erodes public trust in AI technologies (Afroogh et al., 2024). When people feel that these powerful systems are being developed behind closed doors, with little external oversight, it's natural to be skeptical or even fearful.

Policies

How do we make AI development more accountable without stifling innovation? There is no simple answer, but there are several promising approaches to consider. We talk more about proposals and approaches to this in the sections on corporate, national and international governance.

A robust accountability framework for AI development requires interlocking mechanisms operating at different timescales and levels of governance. At the foundational level, pre-deployment approval systems could establish clear capability-based thresholds for AI development accompanied by regulatory

requirements, similar to regulations in other high-risk industries. Deployment could be made contingent on developers meeting safety and transparency requirements, creating a baseline for responsible development practices.

Building on this foundation, ongoing oversight could be maintained through a combination of external audits and ethical review boards. Independent experts would evaluate AI systems' capabilities, training data, and potential impacts, while diverse stakeholders would assess broader ethical implications. This dual-track review process, modeled after successful frameworks in medical research, would help identify and address both technical and societal concerns throughout the development cycle.

To ensure these oversight mechanisms have real impact, they must be backed by clear enforcement capabilities. A well-defined liability framework could establish legal responsibility for AI-related harms, creating strong incentives for careful development practices. This would be complemented by emergency intervention mechanisms, enabling regulatory bodies to respond swiftly to imminent risks from AI deployments – for instance, by halting the release of potentially dangerous systems.

The effectiveness of these measures ultimately depends on transparency and international coordination. Regular public disclosures about AI capabilities, limitations, and risks would enable informed public discourse while protecting legitimate proprietary interests. Given the global nature of AI development, these national frameworks must be harmonized through international agreements to prevent regulatory arbitrage and establish consistent global standards. We talk a lot more about this in the section dedicated to international governance. This coordinated approach would help ensure that accountability measures remain robust even as AI technology continues to advance.

Allocation

AI has the potential to reshape the distribution of power, wealth, and opportunities across society. The issue of allocation or distributive consequences revolves around several questions associated with the consequences of developing and deploying increasingly advanced AI systems: who controls these systems? Who reaps their benefits? And what happens to those left behind?

The distributive consequences of AI span two interrelated dimensions: power and wealth. On the power front, we're seeing a gradual but significant shift in who holds the reins of influence and control in society. Those who develop and control advanced AI systems are gaining unprecedented leverage over economic, political, and social spheres.

Large language models like ChatGPT or Claude are developed and controlled by a handful of tech companies and research institutions. This concentration of power raises serious questions about accountability, transparency, and systemic influence.

Previous technological revolutions have often led to increased inequality, at least in the short to medium term. And with AI, the stakes are even higher, because AI has the potential to be a truly general-purpose technology, one that could theoretically replace human cognitive labor across almost all domains. If (or when) AGI becomes a reality, the distributive consequences could be staggering. Whoever controls an entity capable of outperforming humans in virtually every cognitive task - from scientific research to strategic planning to creative endeavors - would wield considerable power and wealth.

The prospect of AGI amplifies all the distributive concerns discussed so far. It could lead to extreme concentrations of power, potentially even enabling new forms of authoritarian control or technocratic governance.

Policies

How can we address the distributive consequences of AI development and deployment? There is no simple solution, but several approaches are being explored and debated. Redistributive policies could help spread the wealth generated by advanced AI systems. This could take the form of taxes on AI-driven profits, universal basic income programs, investment in education and retraining initiatives, or a 'Windfall Clause' (O'Keefe et al., 2019). These direct redistributive measures can be complemented by longer-term structural changes in how AI development occurs. Democratizing AI development through open-source projects and targeted public funding can help spread access to these transformative technologies beyond a small group of well-resourced organizations. This democratization effort could gain teeth through carefully crafted regulatory frameworks that prevent monopolistic consolidation.

These solutions must grapple with the differential impacts of AI across various segments of society and the global economy. It's not just a matter of the haves versus the have-nots. We're seeing complex dynamics play out between:

- Capital and labor: As AI automates more tasks, the returns to capital (those who own AI systems and the data they run on) may increase relative to returns to labor.
- Frontier AI countries and laggards: Nations at the forefront of AI development may gain significant economic and strategic advantages over others.
- Tech-savvy individuals and the less digitally literate: As AI becomes more integrated into daily life, those who can effectively use and understand these technologies may have significant advantages.

- Large corporations and small businesses: Big tech companies with vast data resources and AI capabilities may gain even more market power, potentially squeezing out smaller competitors.

These differential impacts add layers of complexity to the governance challenge. They underscore the need for nuanced, adaptable policies that can address the specific needs and vulnerabilities of different groups.

Corporate Governance

!!! quote "Elon Musk (Founder/Co-Founder of OpenAI, Neuralink, SpaceX, xAI, PayPal, CEO of Tesla, CTO of X/Twitter)"
<tab>
"AI is a rare case where I think we need to be proactive in regulation than be reactive [...] I think that [digital super intelligence] is the single biggest existential crisis that we face and the most pressing one. It needs to be a public body that has insight and then oversight to confirm that everyone is developing AI safely [...] And mark my words, AI is far more dangerous than nukes. Far. So why do we have no regulatory oversight? This is insane."
</tab>

The challenges we discussed earlier - unexpected capabilities, deployment risks, and rapid proliferation - create complex oversight problems. Companies developing frontier AI have unique visibility into these challenges. They work directly with the models, see capabilities emerge firsthand, and can implement safety measures faster than external regulators (Anderljung et al., 2023; Sastry et al., 2024).

Why do we start with corporate governance? Companies building frontier AI face a balancing act. They have the technical knowledge and direct control needed to implement effective safeguards. But they also face market pressures that can push against taking time for safety measures. Looking at how companies handle this tension helps us understand both the possibilities and limitations of self-regulation in AI development (Zhang et al., 2021; Schuett, 2023).

In this section we'll examine how AI companies approach governance in practice - from basic safety protocols to comprehensive oversight frameworks. We'll look at what works, what doesn't, and where gaps remain. This helps us understand why corporate governance alone isn't enough, setting up our later discussions of national and international oversight. By the end of this section, we'll see both the essential role of company-level governance and why it needs to be complemented by broader regulatory frameworks.

Frontier AI companies can implement internal governance mechanisms to govern AI. This self-regulatory layer serves as a crucial complement to external oversight,

providing more immediate and technically informed controls over AI development and deployment.

Internal governance mechanisms are vital because frontier AI companies possess unique advantages in governing their systems. They have direct access to technical details, development processes, and emerging capabilities; they can implement controls more rapidly than external regulators; and they understand the technical nuances that might escape broader regulatory frameworks. Their proximity to development allows them to identify and address risks earlier and more effectively than external oversight alone could achieve.

IMPORTANT

****Investing in OpenAI Global, LLC is a *high-risk investment*****
****Investors could lose their capital contribution and not see any return****
****It would be wise to view any investment in OpenAI Global, LLC in the spirit of a donation, with the understanding that it may be difficult to know what role money will play in a post-AGI world****

The Company exists to advance OpenAI, Inc.'s mission of ensuring that safe artificial general intelligence is developed and benefits all of humanity. The Company's duty to this mission and the principles advanced in the OpenAI, Inc. Charter take precedence over any obligation to generate a profit. The Company may never make a profit, and the Company is under no obligation to do so. The Company is free to re-invest any or all of the Company's cash flow into research and development activities and/or related expenses without any obligation to Memebers. See Section 6.4 for additional details.

<caption>
A section of the operating agreement between OpenAI, LLC (for-profit entity) and OpenAI, Inc. (non-profit entity). (OpenAI, 2024)
</caption>

For instance, companies can implement real-time monitoring of model behavior, establish internal review boards for sensitive applications, and develop sophisticated testing protocols that would be difficult to mandate through external regulation. This privileged position in the development process creates both opportunity and responsibility for robust self-governance.

Components of Internal Governance - Effective internal governance can be complex, ranging from comprehensive technical standards to organizational structures. Companies can establish detailed development guidelines that incorporate safety considerations from the earliest stages of research, alongside rigorous testing protocols to evaluate system capabilities and limitations. These technical standards can be

accompanied by clear deployment criteria that must be met before systems can be released or scaled.

The organizational structure can support these technical standards through dedicated safety teams with real authority to pause or modify development when necessary. Internal ethics boards can evaluate sensitive applications, while clear escalation paths ensure safety concerns reach appropriate decision-makers quickly. Companies can also consider how to integrate safety considerations into their promotion and compensation structures to align incentives throughout the organization.

Beyond individual measures, frontier AI developers can participate in collective self-regulatory initiatives through industry-wide safety standards and best practices. Voluntary commitments to specific safety measures or deployment restrictions can help establish industry norms, while information sharing about safety-relevant incidents can improve practices across the sector.

Limitations and Challenges - Internal governance faces several significant challenges. Perhaps the most fundamental is the challenge of incentive alignment, as companies face competing pressures between safety and other objectives like market competition, growth, and profitability. Internal governance mechanisms must be robust enough to withstand these pressures, particularly during critical periods of market competition or technological breakthroughs.

Credibility and accountability present another major challenge. Self-regulatory measures may lack credibility without external validation or enforcement mechanisms. Companies may have to find ways to demonstrate their commitment to safety and responsible development that convince external stakeholders of their seriousness and effectiveness.

Coordination problems arise when individual company initiatives fail to address broader societal concerns or system-wide risks. Some challenges require coordination across the industry or between companies and governments, which can be difficult to achieve through purely voluntary measures. The competitive nature of AI development can sometimes work against the kind of open collaboration needed to address these broader challenges.

The Role of Transparency and External Validation - Voluntary governance is not necessarily internal to the company. It can include mechanisms for transparency and external validation. Regular public reporting on safety measures and incidents provides accountability, while third-party audits of safety systems and processes offer independent verification of governance effectiveness. Companies can maintain active engagement with external stakeholders and experts to ensure their governance approaches remain relevant and effective.

The relationship with external regulation is particularly important. Internal governance should complement rather than replace external oversight, with companies designing internal systems that can interface effectively with regulatory requirements. This includes maintaining documentation that can support compliance efforts and participating constructively in the development of regulatory frameworks. Companies can also share relevant insights and experience with policymakers to help inform the development of effective external oversight mechanisms.

<iframe
src="https://ourworldindata.org/grapher/affiliation-researchers-building-artificial-intelligence-systems-all?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Share of notable AI systems by researcher affiliation (Giattino et al., 2023)
</caption-iframe>

Frontier Safety Frameworks

Frontier Safety Frameworks are internal policies that AI companies create to guide their development process and ensure they're taking appropriate precautions as their systems become more capable. They're the equivalent of the safety protocols used in nuclear power plants or high-security laboratories. At the Seoul AI Summit organized in May 2024, 16 companies around the world committed to implementing such policies (UK government, 2024).

Two of the biggest names in the AI world, Anthropic - through its Responsible Scaling Policy - and OpenAI - through its Preparedness Framework -, have been at the forefront of developing these frameworks. Let's take a closer look at their approaches.

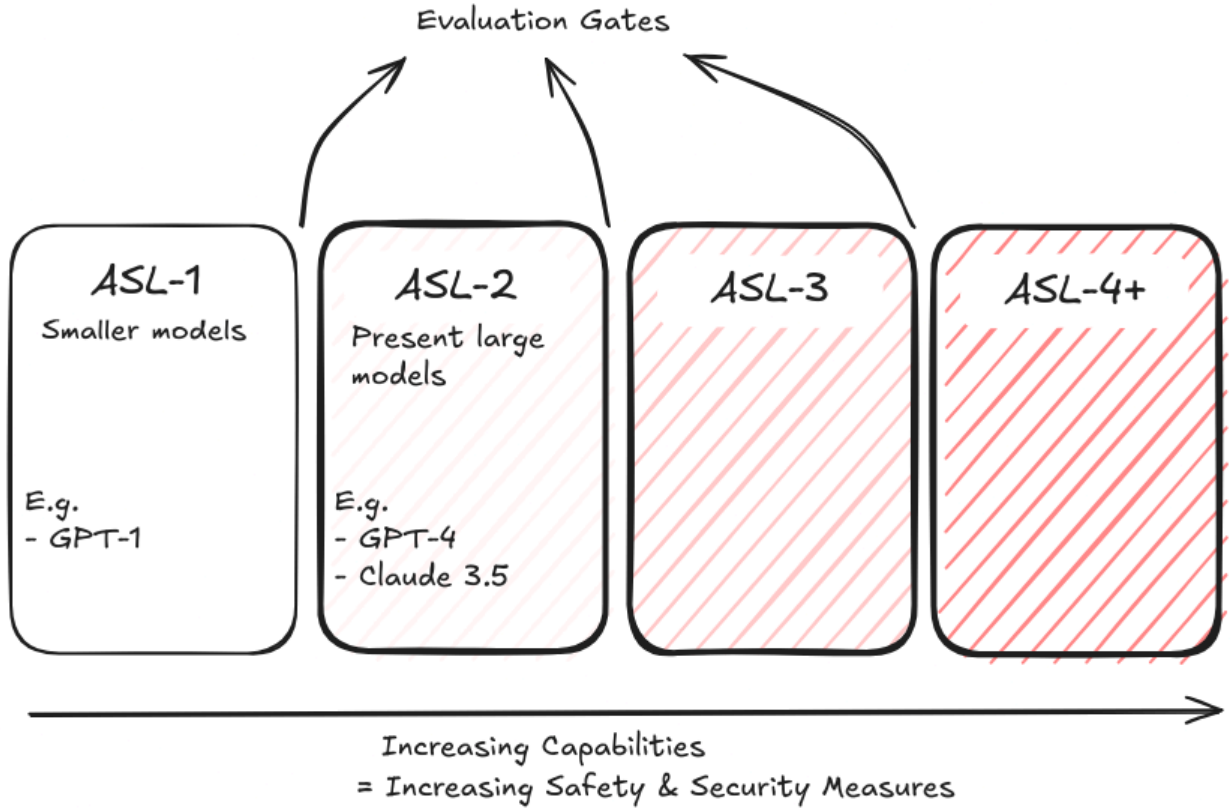
Anthropic's Responsible Scaling Policy (RSP)

Anthropic's Responsible Scaling Policy is a document that outlines different "AI Safety Levels" (ASLs) and the corresponding safety measures that need to be in place as their models become more powerful.

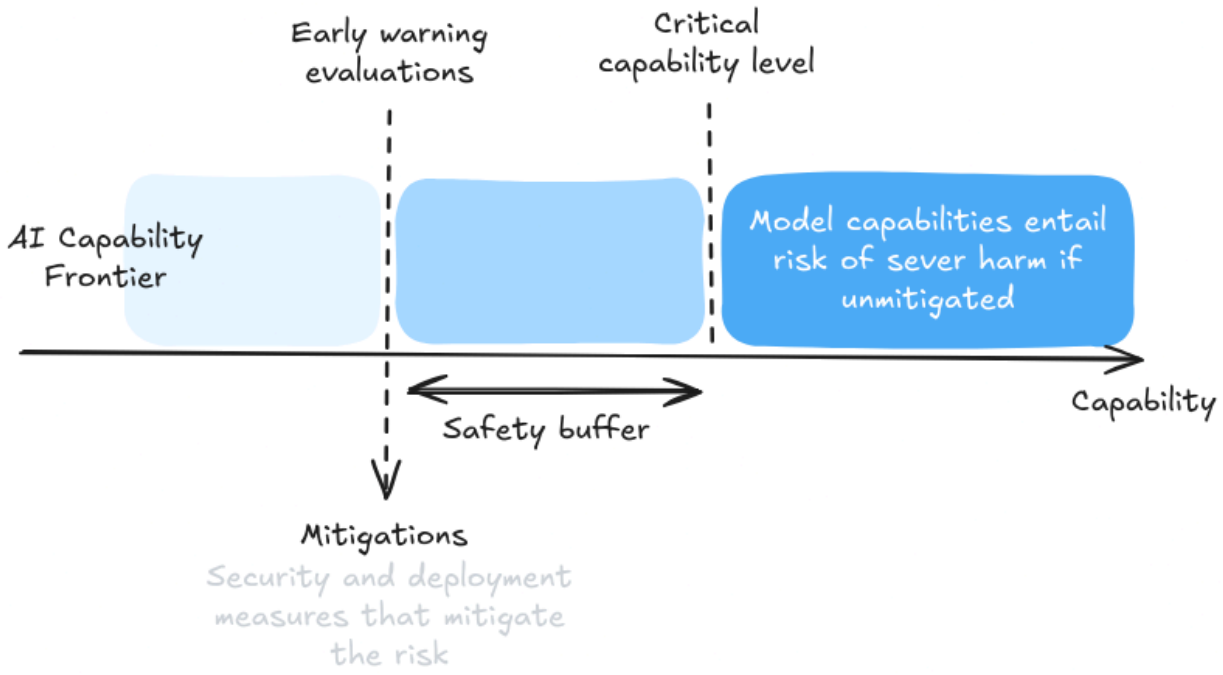
For example, at ASL-2 (which includes their current most advanced model, Claude 2), Anthropic commits to things like publishing detailed model cards, providing a way for people to report vulnerabilities, or enforcing strict rules about how the model can be used.

For higher risk levels (ASL-3 and above), Anthropic ratchets up the precautions significantly. They talk about limiting access to training techniques, implementing much

stronger security measures, and even being prepared to pause development entirely if things get too dicey.



<caption>
Overview of Anthropic's ASL levels. ASL-1 refers to systems which pose no meaningful catastrophic risk. ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't. ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities. ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy. (Anthropic, 2024)
</caption>

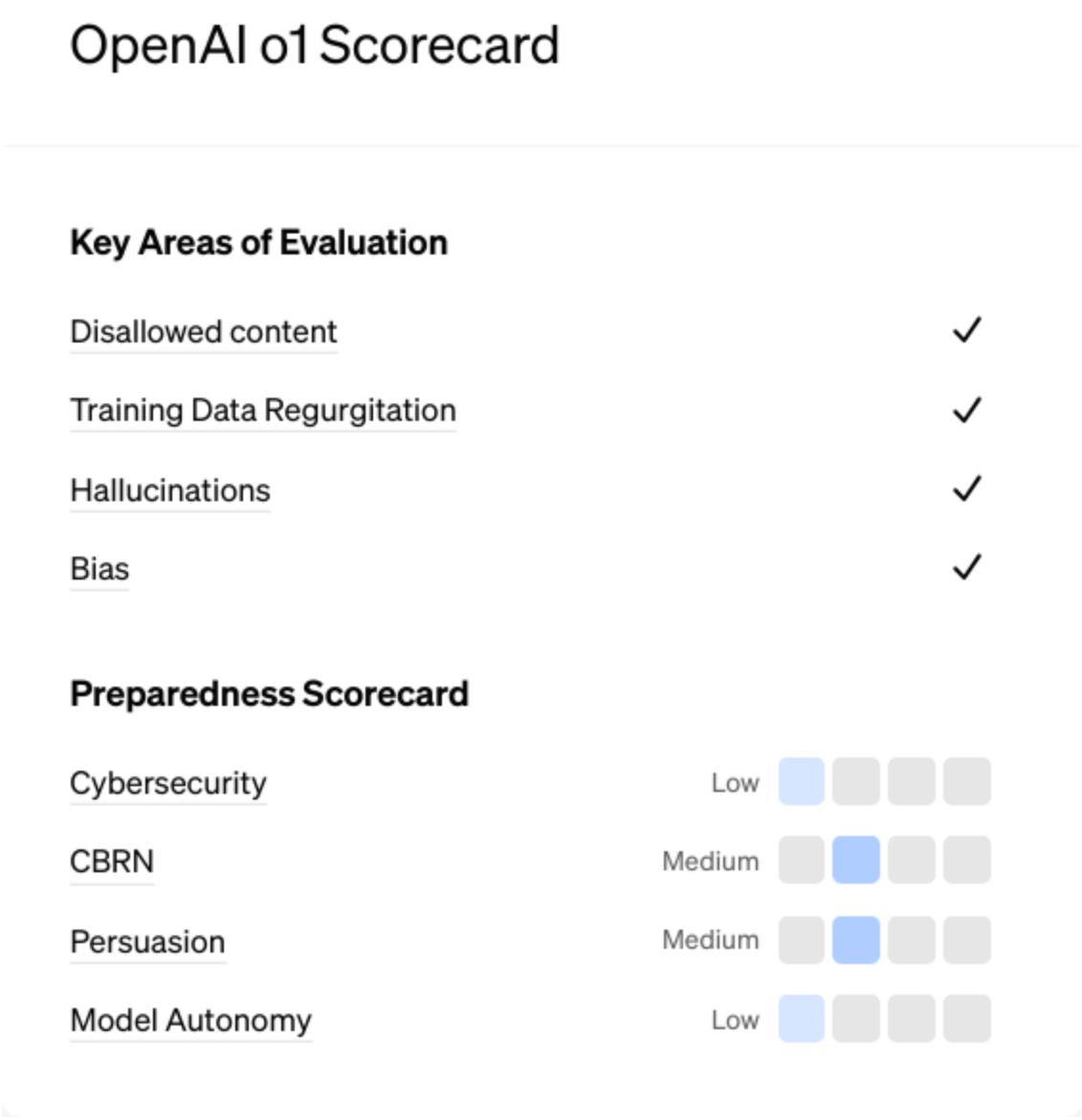


<caption>
DeepMinds outlook from the Frontier model Safety Framework. (DeepMind, 2024)
</caption>

OpenAI's Preparedness Framework

OpenAI's Preparedness Framework revolves around what they call a "Scorecard" – essentially a way to evaluate the risks associated with their AI models across different categories.

OpenAI's is quite specific in certain areas. For instance, they have a whole category dedicated to CBRN risks (Chemical, Biological, Radiological, and Nuclear risk). They define their "High" risk threshold in this category as: "Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN threat."



<caption>
System card of GPT-o1 published by OpenAI after safety evaluations. (OpenAI, 2024)
</caption>

The Strengths and Weaknesses of Current Approaches

The current governance frameworks from major AI labs reveal both promising approaches and concerning gaps in industry self-regulation. Their public nature enables valuable external scrutiny, while their risk categorization demonstrates engagement with potential failure modes. The frameworks' deliberately flexible structure allows adaptation as our understanding of AI risks evolves.

However, these strengths are undermined by several interconnected weaknesses. The frequent use of ambiguous language makes consistent application difficult, while the frameworks' voluntary nature raises questions about their actual implementation when commercial pressures conflict with safety considerations. Some critics argue the frameworks aren't conservative enough given the stakes involved, potentially setting risk thresholds too high and mitigation requirements too low. Additionally, their focus on individual system risks may miss emergent dangers from multiple AI systems interacting in complex ways. The lack of standardization across companies further complicates industry-wide coordination, though this may improve as best practices emerge through practical implementation.

The Governance Challenge

How do we ensure that companies actually implement their frontier safety frameworks? Both Anthropic and OpenAI have outlined some governance measures in their frameworks.

Anthropic has made some interesting commitments in terms of governance:

- Creating a role called the "Responsible Scaling Officer." This person is supposed to be the guardian of the RSP, making sure the company is living up to its commitments.
- Proactively planning for scenarios where they might need to pause scaling of their models. This shows they're thinking ahead about potential crises.
- Sharing evaluation results publicly (where possible), which adds a layer of external accountability.

Some think those policies have gaps (Anderson-Samways et al., 2024). They include a clause that says in "extreme emergency" situations, like if a "rogue state" is developing AI recklessly, they might loosen their restrictions. While this flexibility could be necessary, it also potentially undermines the credibility of their other commitments. After all, who defines what constitutes an "extreme emergency"?

On their side, OpenAI has outlined a three-tiered governance structure: their Preparedness team conducts foundational research and monitoring, providing technical expertise to inform governance decisions. This research feeds into a Safety Advisory Group that brings diverse perspectives to risk assessment and mitigation recommendations. Final authority rests with OpenAI's leadership and Board of Directors.

This structure has some clear strengths. The dedicated Preparedness team ensures that safety considerations are always at the forefront. The advisory group brings in outside perspectives, which can help challenge groupthink. And having the Board as a final backstop could provide an additional layer of oversight.

However, questions remain. How much power does the Preparedness team really have? Can they delay or veto projects they deem too risky? How is the Safety Advisory Group selected, and how much influence do they actually wield? And given that OpenAI is ultimately a for-profit company (despite its unusual structure), how do we ensure that safety always trumps commercial interests?

The Road Ahead

The frameworks and governance structures being developed by companies like Anthropic and OpenAI are important first steps. They show a recognition of the enormous responsibility that comes with developing these powerful systems.

There is still room for improvement. Some suggest that companies like Anthropic should define more precise, verifiable risk thresholds for their safety levels, potentially drawing on societal risk tolerances from other industries (Anderson-Samways 2024). For instance, in industries dealing with potentially catastrophic risks (events causing 1,000 or more fatalities), maximum tolerable risk levels typically range from 1 in 10,000 to 1 in 10 billion per year. AGI companies might consider adopting similar quantitative thresholds, adjusted for the potentially even greater stakes involved in AGI development.

Overall, we need a much more robust, standardized, and enforceable set of governance practices for frontier AI development. Moreover, we need to foster a culture within the AI community that prioritizes safety and ethical considerations as much as technical achievements. The goal should be to make responsible AI development not just a regulatory requirement, but a core value of the field.

Policy options

Risk Assessment Methods. Drawing from established safety-critical industries, AGI companies can adapt and implement various systematic approaches to evaluate potential risks. These range from scenario analysis and fishbone diagrams to more specialized techniques like the Delphi method, providing structured ways to anticipate and prepare for both known and unknown challenges in AGI development.

The Three Lines of Defense. A robust organizational structure for risk management is essential for AGI companies, implemented through a three-tiered defense system. This framework distributes responsibility across frontline researchers, specialized risk management teams, and independent auditors, ensuring multiple layers of oversight and risk detection throughout the development process.

Coordinated Pausing. When dangerous capabilities emerge in AI systems, companies need systematic ways to respond collectively. The coordinated pausing framework provides a structured approach for companies to temporarily halt development, share

critical safety information, and resume work only when appropriate safeguards are in place, preventing competitive pressures from compromising safety.

Deployment Corrections. Even the most rigorous pre-deployment safeguards may not catch every risk. A comprehensive system of deployment corrections enables companies to maintain control over deployed models, respond rapidly to emerging risks, and implement rollback mechanisms when necessary, ensuring safety even after systems are in production.

Industry Best Practices. The AI safety & governance field is converging on a set of core governance practices, supported by broad expert consensus. These include pre-deployment risk assessments, dangerous capabilities evaluations, and third-party audits, representing an emerging standard for responsible AGI development that balances innovation with safety.

Risk Assessment Methods

At the heart of effective governance in frontier AI companies lies a robust approach to risk assessment. How do you assess risks for technologies that don't yet exist and capabilities that may emerge unexpectedly?

This is where we can learn from other safety-critical industries. Techniques from fields like aerospace, nuclear power, and cybersecurity could be adapted to the unique challenges of AI development.

Let's take a closer look at some of these techniques (Koessler & Schuett 2023):

- **Scenario Analysis**: This involves imagining potential future scenarios and their implications. For AI companies, this might include scenarios like: An AI system developing deceptive behaviors, Unexpected emergent capabilities in a deployed model, A rival company deploying an unsafe AI system.
- **Fishbone Method**: Also known as the Ishikawa diagram, this technique helps identify potential causes of a problem. In the context of AI risks, a fishbone diagram might explore factors contributing to AI alignment failure, such as: Insufficient safety research, Pressure to deploy quickly, Inadequate testing protocols, Misaligned incentives in the AI system
- **Causal Mapping**: This technique visualizes the complex web of cause-and-effect relationships in a system. For AI development, a causal map could illustrate how different research decisions, safety measures, and deployment strategies interact to influence overall risk.
- **Delphi Technique**: This method involves gathering expert opinions through structured rounds of questionnaires. Given the highly specialized nature of AI research, the Delphi technique could be valuable for synthesizing diverse perspectives on potential risks and mitigation strategies.

- **Bow Tie Analysis**: This approach visualizes the pathways between causes, hazardous events, and consequences, along with prevention and mitigation measures. For an AI company, a bow tie analysis might focus on a hazardous event like "loss of control over an AI system," mapping out potential causes (e.g., inadequate containment measures) and consequences (e.g., unintended global changes), along with preventive and reactive controls.

Implementing these techniques requires a cultural shift within AGI companies. Risk assessment can't be an afterthought or a box-ticking exercise; it needs to be woven into the fabric of the organization, from the research lab to the boardroom.

The Three Lines of Defense

As AGI companies grapple with these complex risk landscapes, they need robust organizational structures to manage them effectively. One promising approach is the Three Lines of Defense (3LoD) model, a risk management framework widely used in other industries (Schuett 2023).

In the context of an AGI company, the 3LoD model might look something like this:

The First Line of Defense. This comprises the frontline researchers and developers working on AI systems. They're responsible for implementing safety measures in their day-to-day work, conducting initial risk assessments, and adhering to the company's ethical guidelines and safety protocols.

The Second Line of Defense. This includes specialized risk management and compliance functions within the company. For an AI company, this might involve:

- An AI ethics committee overseeing the ethical implications of research directions
- A dedicated AI safety team developing and implementing safety protocols
- A compliance team ensuring adherence to relevant regulations and industry standards

The Third Line of Defense. This is typically the internal audit function, providing independent assurance to the board and senior management. In an AI company, this might involve:

- Regular audits of safety practices and risk management processes
- Independent evaluations of AI models for dangerous capabilities
- Assessments of the company's overall preparedness for potential AGI scenarios

Let's see how this might work in practice:

Imagine that researchers in an AI company (first line) develop a new language model with unexpectedly advanced capabilities in logical reasoning. They flag this to the AI

safety team (second line), who conduct a thorough evaluation and determine that the model poses potential risks if deployed without additional safeguards.

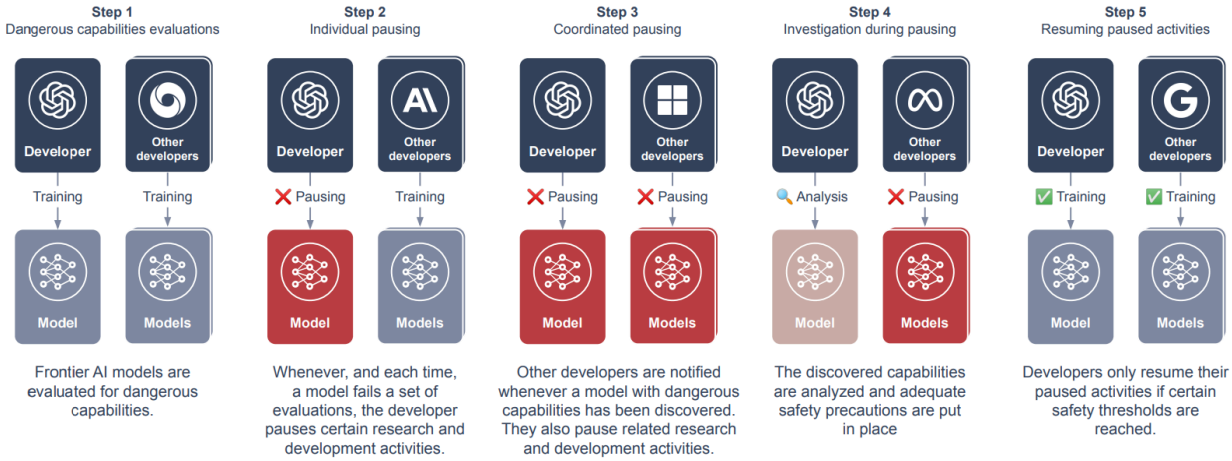
The safety team works with the researchers to implement additional constraints on the model's outputs. Meanwhile, they also notify the internal audit team (third line), who launch a broader review of the company's processes for identifying and managing emergent capabilities.

This multi-layered approach helps ensure that risks are identified and managed at multiple levels, reducing the chances of dangerous oversights.

Coordinated Pausing

The emergence of unexpected and potentially dangerous capabilities is a very real possibility. How should AI companies respond when such capabilities are discovered?

One innovative proposal is the concept of "coordinated pausing" (Alaga & Schuett 2023). This approach suggests a structured process for responding to the discovery of dangerous capabilities.



<caption>
(Alaga & Schuett 2023)
</caption>

This approach could take various forms, from a purely voluntary system relying on public pressure, to a more formalized agreement between developers, or even a legally mandated framework.

The benefits of such a system are clear. It provides a mechanism for the AI community to collectively pump the brakes when potentially dangerous territory is entered, allowing time for careful analysis and the development of safety measures.

However, implementing such a system is not without challenges. There are practical questions about how to define "dangerous capabilities" and who gets to make that determination. There are also potential legal hurdles, particularly around antitrust concerns.

Deployment Corrections

Even with the most rigorous pre-deployment safeguards, there's always the possibility that dangerous capabilities or behaviors might emerge after an AI system is deployed. This is where the concept of "deployment corrections" comes into play.

Companies thus need comprehensive contingency plans for scenarios where pre-deployment risk management falls short (O'Brien et al. 2023). At the technical level, this means maintaining continuous control over deployed models through robust monitoring and modification capabilities, supported by pre-built rollback mechanisms that can revert to earlier, safer versions when needed. These technical controls are complemented by organizational preparedness through dedicated incident response teams trained in rapid risk assessment and mitigation. Clear user agreements establish the legal and operational framework for emergency interventions, ensuring all stakeholders understand how and when access restrictions might be imposed.

Towards Industry-Wide Best Practices

As the field of AGI development matures, there's a growing recognition of the need for industry-wide best practices. A survey of 92 experts from AI labs, academia, and civil society found broad agreement on a number of key practices, including pre-deployment risk assessments, dangerous capabilities evaluations, third-party model audits, and safety restrictions on model usage (Schuett et al. 2023).

Interestingly, 98% of respondents agreed with all of these measures, suggesting a growing consensus around certain core principles of responsible AGI development.

National Governance

The need for national governance

!!! quote "Zhang Jun (China's UN Ambassador)"
<tab>
"The potential impact of AI might exceed human cognitive boundaries. To ensure that this technology always benefits humanity, we must regulate the development of AI and prevent this technology from turning into a runaway wild horse [...] We need to

strengthen the detection and evaluation of the entire lifecycle of AI, ensuring that mankind has the ability to press the pause button at critical moments."
</tab>

While leading AI companies have implemented various self-regulatory measures to ensure the safe development of frontier AI systems, relying solely on corporate self-regulation is insufficient to protect national interests and public welfare. While such voluntary measures allow for rapid response to emerging issues and can often move faster than government regulation, companies may lack incentives to fully account for broader societal impacts, may face competitive pressures that compromise safety considerations, and may not have the legitimacy to make decisions that affect entire populations. National governance frameworks are therefore essential to ensure comprehensive oversight and accountability. A robust national regulatory framework needs to build on and complement these self-regulatory efforts. It should provide a baseline of standards that all companies must meet, while still allowing room for companies to go above and beyond in their internal practices.

Institutional Fit and the Challenge of Frontier AI - The concept of institutional fit—the degree to which governance institutions match the scale, scope, and characteristics of the problems they aim to address—is crucial for understanding why national governance of frontier AI is both necessary and challenging. Institutional fit helps us analyze whether existing regulatory bodies and frameworks are adequately equipped to handle the unique challenges posed by frontier AI systems, or whether new institutional arrangements are needed.

The governance of frontier AI systems presents a particular challenge for institutional fit. Unlike traditional technological governance challenges, frontier AI systems generate externalities that span multiple domains - from national security to economic stability, from social equity to democratic functioning. Traditional regulatory bodies, designed for narrower technological domains, may lack the necessary spatial remit, technical competence, or institutional authority to effectively govern these systems (Dafoe, 2023).

Consider the contrast with self-driving vehicles, where the primary externalities are relatively well-defined (safety of road users) and fall within existing regulatory frameworks (traffic safety agencies) (Dafoe, 2023). Frontier AI systems, by contrast, generate externalities that cross traditional regulatory boundaries and jurisdictions, requiring new institutional approaches.

Addressing Institutional Gaps - The governance of frontier AI reveals several institutional gaps in current regulatory frameworks (Dafoe, 2023). The expertise gap manifests in traditional regulatory bodies' frequent lack of technical expertise to evaluate advanced AI systems. This necessitates either the development of new technical capabilities within existing institutions, the creation of new specialized regulatory bodies, or novel partnerships between government and technical experts.

A coordination gap exists due to the cross-cutting nature of frontier AI externalities. New mechanisms are needed for coordination between different regulatory agencies, federal and state/local authorities, public and private sector entities, and domestic and international governance bodies.

The temporal gap emerges from the rapid pace of AI development, creating a mismatch with traditional regulatory processes. Governance frameworks must be adaptable to technological change, capable of anticipating future developments, and able to respond quickly to emerging risks.

Implementation Challenges - Several factors complicate the implementation of effective domestic governance. Political polarization can impede the development of consensus on governance approaches, particularly regarding the appropriate level of state oversight, balance between innovation and regulation, distribution of benefits and risks, and protection of civil liberties.

Technical complexity creates challenges for effective oversight and monitoring, development of appropriate standards, assessment of compliance, and risk evaluation and management.

The governance of frontier AI systems requires significant institutional innovation at the national level. While existing regulatory frameworks provide some foundation, the unique characteristics of frontier AI - its broad externalities, rapid development, and deep political implications - necessitate new approaches to governance. Success will require careful attention to institutional fit, stakeholder representation, and the balance between competing interests and values (Dafoe, 2023).

National governance is also more complex to create and maintain than self-regulatory efforts by companies because laws and regulations result from a sometimes long and complex policy-making process, which unfolds in distinct phases, each offering opportunities for governance interventions. During the agenda-setting phase, governance actors work to elevate specific AI-related issues to the forefront of public and political discourse. The formulation phase involves crafting detailed policy proposals, while implementation transforms these proposals into actionable measures. Throughout this cycle, evaluation and adaptation remain crucial, allowing governance approaches to evolve in response to the rapidly changing AI landscape.

The development of effective domestic governance frameworks for frontier AI is not merely a technical challenge but a fundamental political and institutional one. It requires building new capabilities while maintaining democratic legitimacy and balancing multiple competing interests. As AI capabilities continue to advance, the ability to develop and implement such frameworks will become increasingly crucial for national welfare and security.

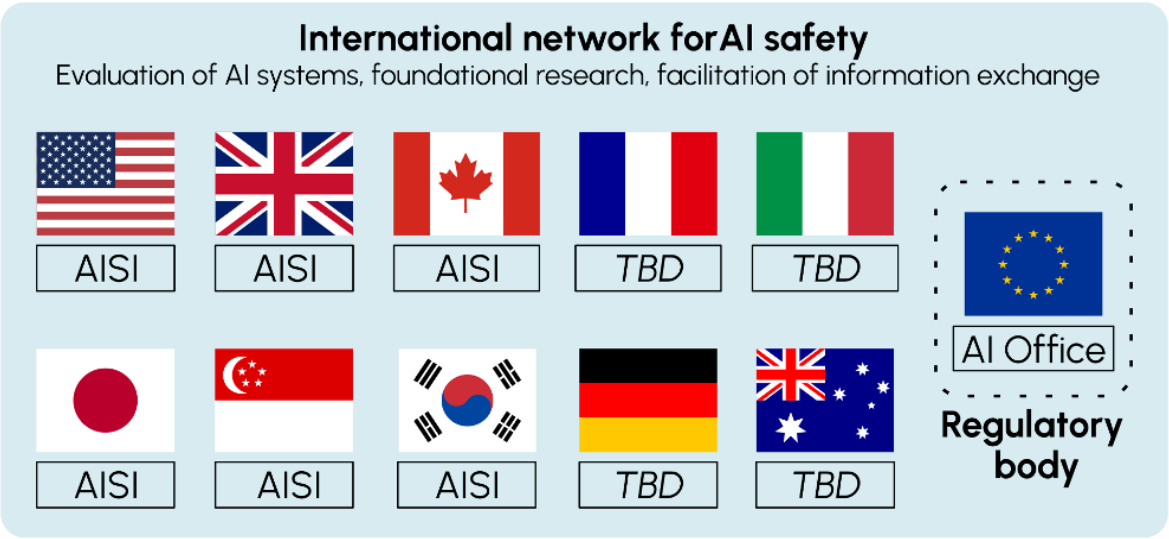
Current initiatives

AI Safety Institutes

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

Governments worldwide have recognized an urgent need to understand and manage the capabilities and risks of advanced artificial intelligence systems. This has led to the formation of AI Safety Institutes (AISIs), specialized government bodies designed to evaluate, research, and coordinate efforts to ensure AI development proceeds safely and beneficially.

The Global Movement Toward AI Safety - In recent months, we've witnessed a remarkable surge in the establishment of AISIs across major technological powers. The United States, United Kingdom, Japan, Canada, and Singapore have all launched their own institutes, while the European Union has integrated these responsibilities into its AI Office through a dedicated AI Safety Unit.



<caption>
Announced AI Safety Institutes (Martinet & Variengien, 2024)

</caption>

Core Functions of AI Safety Institutes - We can think of AISIs as serving three fundamental purposes, each building upon the others to create a comprehensive approach to AI safety. First, they evaluate AI systems through testing and assessment protocols. This involves developing new methodologies to understand these systems' capabilities, limitations, and potential impacts on society. Second, they can help conduct foundational research in AI safety, bringing together experts from various disciplines to advance our understanding of how to build and deploy AI systems safely. Finally, they serve as information exchange hubs, creating channels for sharing crucial insights among stakeholders, from policymakers to private companies.

International Coordination and Collaboration - AI Safety Institutes have been designed from the ground up to work together across borders. The culmination of this international vision was realized at the May 2024 Seoul AI Summit, where ten countries and the European Union established a network for AI safety.

Practical Challenges and Solutions - While the promise of international collaboration through AISIs is compelling, several practical challenges must be addressed. First, there's the delicate balance of sharing sensitive information about AI systems' capabilities while protecting commercial secrets and national security interests. Then there's the challenge of varying technical capacities between nations – not every country has equal resources to attract top AI talent or conduct sophisticated evaluations. Some institutes, like the UK's AISI, have taken innovative approaches to this challenge, such as opening offices in AI talent hubs like San Francisco.

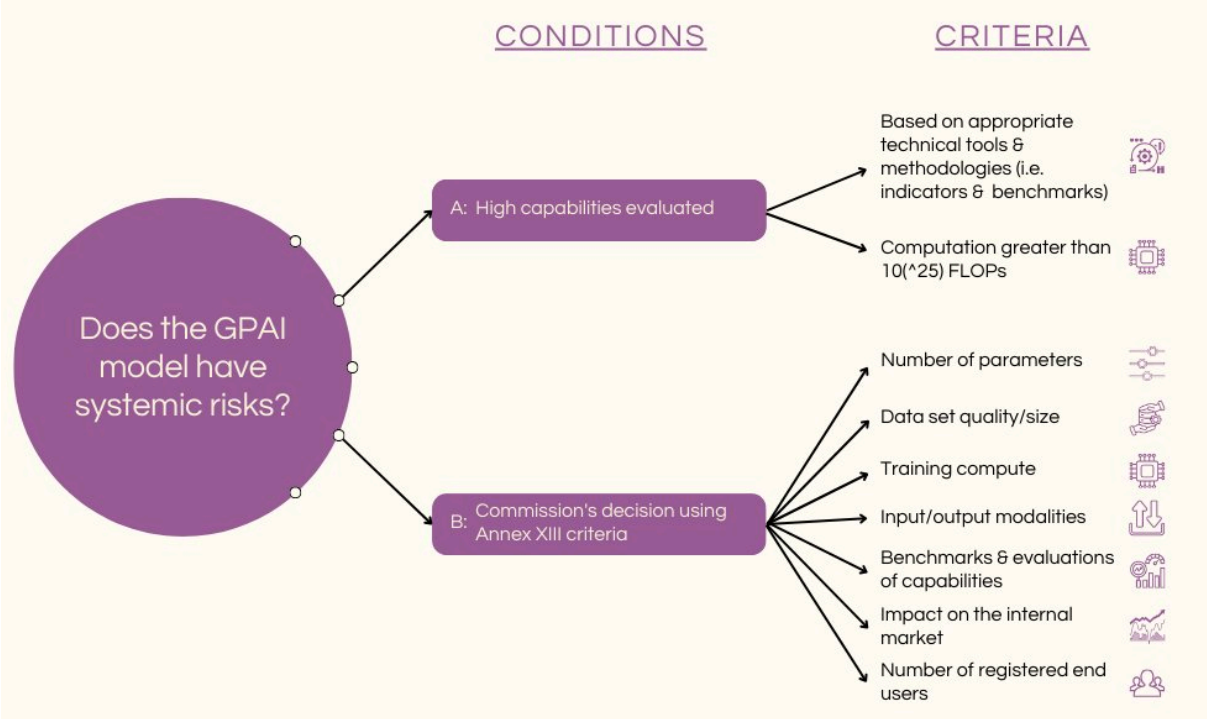
Looking to the Future - As these institutes mature, they will likely play an increasingly important role in developing international standards, conducting evaluations, and ensuring that AI development proceeds in a way that benefits humanity while minimizing potential risks. Their success will depend not only on technical expertise but also on their ability to facilitate meaningful collaboration across borders and between different stakeholders in the AI ecosystem.

The EU AI Act

The European Union's AI Act addresses General Purpose Artificial Intelligence (GPAI) models, and we'll focus here on what the AI Act calls GPAI models with systemic risks - the equivalent of frontier AI models.

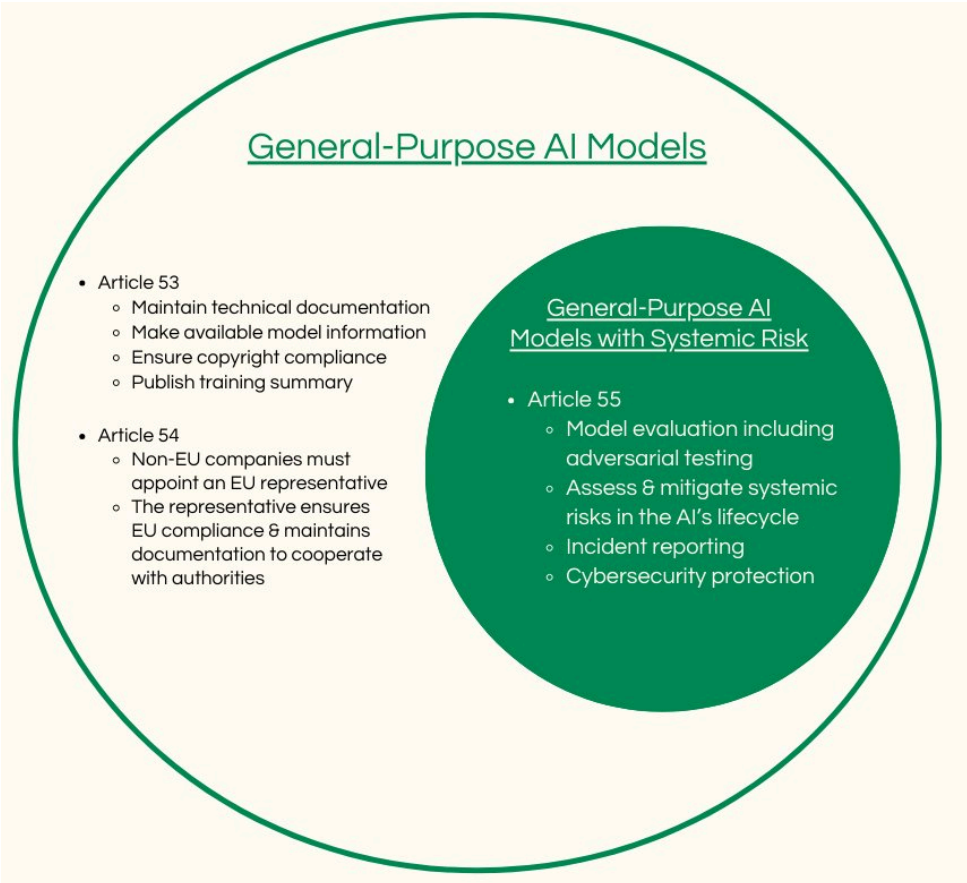
The Act takes a dual approach to identifying GPAI models that present systemic risk. First, there's a computational threshold: any model using more than 10^25 floating point operations (FLOPs) in its training is automatically classified as presenting systemic risk. To put this in perspective, training such a model currently requires an investment of tens of millions of Euros. However, computational power isn't the only

consideration. The Commission can also designate models as systemic based on their potential impact, considering factors such as user base size, scalability potential, and the possibility of causing large-scale harm. This flexible approach ensures that regulation can adapt to emerging risks, even when they come from models that don't meet the computational threshold.



<caption>
The EU AI Act: Classification of general-purpose AI models with systemic risks (source: Observatorio de Riesgos Catastróficos Globales)
</caption>

Provider Obligations and Compliance - Starting August 2, 2025, providers of GPAI models must meet various obligations, with additional requirements for those models deemed to present systemic risk. All GPAI providers must maintain detailed technical documentation and provide comprehensive information to downstream providers who integrate their models. They must also implement copyright compliance policies and publish summaries of their training data. For models with systemic risk, the requirements intensify. These providers must conduct thorough evaluations, including adversarial testing to identify potential vulnerabilities. They must also track and report serious incidents, implement robust cybersecurity protections, and actively work to assess and mitigate systemic risks.



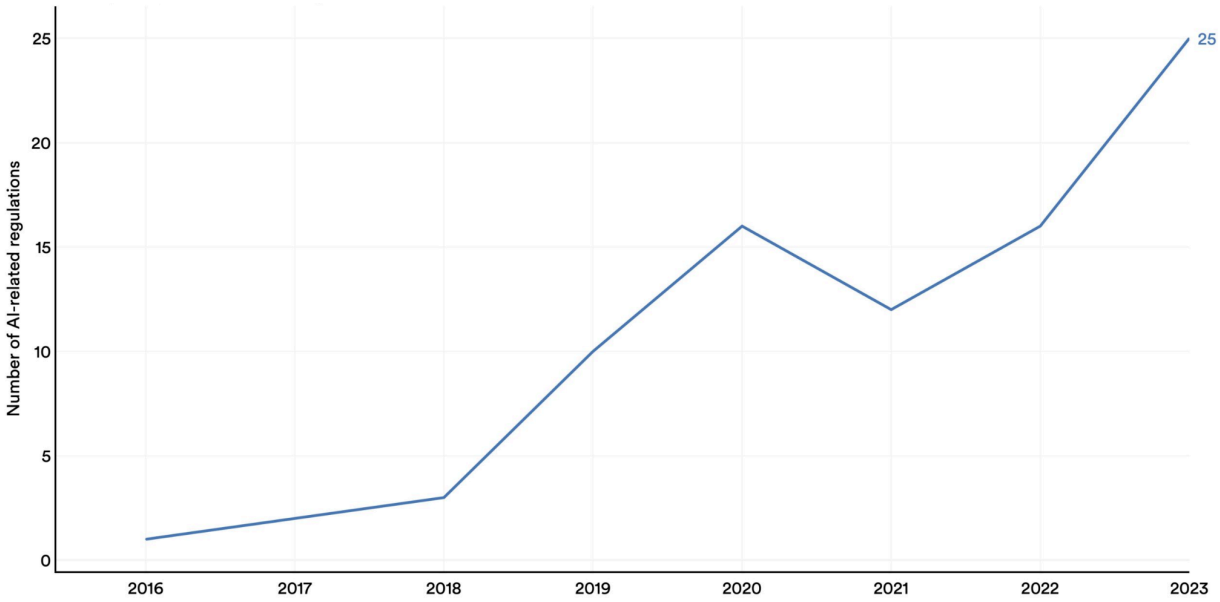
<caption>
The EU AI Act: Obligations for providers of general-purpose AI models (Observatorio de Riesgos Catastróficos Globales)
</caption>

Enforcement and the AI Office - The EU AI Act establishes the AI Office - who also acts as the EU's AI Safety Institute - as a powerful enforcement authority. This office can request information, conduct model evaluations, and mandate corrective measures when necessary. The penalties for non-compliance are substantial – providers can face fines of up to 3% of their global annual turnover or €15 million, whichever is higher. This robust enforcement mechanism reflects the EU's commitment to ensuring that powerful AI systems are developed and deployed responsibly.

The Role of the Code of Practice - The Act introduces an innovative approach to compliance through its Code of Practice. While not mandatory, this code provides a practical pathway for providers to demonstrate their compliance with the Act's requirements.

The US Executive Order on AI

The United States has seen a flurry of legislative activity in recent years. The Executive Order on AI, signed by president Joe Biden on October 30 in 2023 stands out. Its Section 4 represents one of the most wide-ranging extensions of regulatory visibility into AI development in the United States. It introduces safety and security measures that will shape the future of AI development in the United States.



<caption>
Number of AI-related regulations in the United States, 2016-2023 (2024 AI Index report)
</caption>

New Reporting Requirements for AI Companies - The order establishes reporting requirements for companies involved in AI development. Companies developing dual-use foundation models - sophisticated AI models trained on broad datasets using self-supervision and containing tens of billions of parameters - must provide detailed reports about their activities. These reports must cover their training processes, security measures, model weights protection strategies, and results from red-team testing. Similarly, entities operating large-scale computing clusters must disclose their locations and total available computing power.

Infrastructure and Foreign Entity Regulations - A particularly interesting aspect of Section 4 involves new regulations for Infrastructure as a Service (IaaS) providers. These companies must now report when foreign entities use their services for AI training that could enable concerning activities. This requirement extends to foreign resellers of U.S. IaaS services, creating a comprehensive monitoring system for AI development infrastructure. The secretary of commerce must draft regulations requiring these providers to verify the identities of foreign persons obtaining IaaS accounts and establish minimum standards for verification and recordkeeping - essentially, a Know-Your-Customer framework.

Policy options

A comprehensive domestic governance regime for AI safety requires three interconnected mechanisms: development of safety standards, regulatory visibility, and compliance enforcement (Anderljung et al. 2023). These components can work together to create a framework that can effectively manage the risks associated with AI development and deployment.

Mechanisms for developing safety standards - First and foremost, we need to establish processes for identifying appropriate requirements for frontier AI developers that can evolve with the technology. Safety standards form the foundation of AI governance by establishing clear, measurable criteria for the development, testing, and deployment of AI systems. These standards must be technically precise while remaining flexible enough to accommodate rapid technological advancement.

The development of AI safety standards typically involves multiple stakeholders, including technical experts, industry representatives, civil society organizations, and government agencies. Standards development organizations (SDOs) often serve as central coordinating bodies for this process. For example, the National Institute of Standards and Technology (NIST) in the United States has developed AI risk management frameworks that serve as voluntary standards.

Mechanisms for ensuring regulatory visibility - The second building block involves creating mechanisms for regulators to gain visibility into frontier AI development processes. This is crucial for staying ahead of potential risks and ensuring compliance with established standards. Regulatory visibility mechanisms enable oversight bodies to monitor AI development and deployment effectively. These mechanisms provide regulators with the information and access needed to assess compliance with safety standards and identify emerging risks.

Mechanisms for ensuring compliance - The third building block involves creating mechanisms to ensure compliance with safety standards for the development and deployment of frontier AI models. This is where the rubber meets the road in terms of enforcement. Compliance mechanisms transform safety standards from theoretical frameworks into practical requirements with real consequences. These mechanisms must balance the need for effective enforcement with the importance of not stifling innovation.

Mechanisms for developing safety standards

Various approaches to developing safety standards exist, from traditional standardization bodies to more dynamic multi-stakeholder processes like the EU GPAI Code of Practice. This Code, currently under development, demonstrates the vital

importance of the standardization process. While not a traditional standardization mechanism, it serves to specify the high-level obligations outlined in the EU AI Act for GPAI models.

The Act mandates that providers of GPAI models with systemic risks must "ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model." However, this broad requirement raises numerous critical questions: What constitutes an "adequate level" of protection? What exactly comprises the "physical infrastructure" and the "model"? What evidence sufficiently demonstrates their protection? Through what specific measures should this protection be implemented?

These questions highlight why standardization is essential - organizations need guidance to comply with their legal obligations effectively. Legal ambiguity, while it can sometimes be exploited by companies to their advantage, can also create significant operational challenges and risks for companies developing and deploying AI systems.

What needs to be standardized - the example of cybersecurity protection - Protection of key AI assets requires a layered security architecture that addresses distinct but interconnected vulnerabilities. Four critical components demand protection: model weights, source code, training data, and user data. Each represents a unique security challenge while forming part of an integrated system where a breach in one area could compromise the whole.

Model Weights

Model weights are the result of extensive training processes, often requiring massive computational resources and proprietary datasets. For companies like OpenAI, Anthropic, or Google, these weights represent a large part of their competitive edge. If leaked, it could allow competitors or malicious actors to replicate their models, potentially removing safety measures or misusing them.

Protection starts with robust encryption of stored weights, complemented by strict access controls limiting internal visibility. Advanced security can also involve segmenting weights across multiple secure locations, making unauthorized access more difficult. Continuous monitoring watches for suspicious access patterns or unusual data transfers, enabling rapid response to potential breaches.

Source Code

The source code defines how the model processes information, makes decisions, and generates outputs. For AI companies, this code represents years of research and development, often containing proprietary algorithms and architectures.

Protecting source code isn't a new challenge – software companies have been doing it for decades. However, the stakes are higher with frontier AI. A leak could not only benefit competitors but also potentially allow malicious actors to identify and exploit vulnerabilities in the AI system.

Comprehensive protection requires secure, access-controlled version control systems managing all code changes. Advanced techniques include code obfuscation to impede understanding if breached, combined with rigorous security audits and coding standards. Critical development could also occur on air-gapped systems, physically isolated from external networks to prevent unauthorized access.

Training Data

The training data can include everything from public web pages to proprietary information and even personal data. The challenge here is twofold: protecting the data itself and ensuring it's used ethically. A breach could expose sensitive information, while misuse could lead to biased or harmful AI models.

Protection begins with thorough data anonymization, removing identifiable information without compromising training utility. Encrypted databases with strict access controls secure stored data, while comprehensive lineage tracking maintains clear records of data sources and usage patterns. This allows organizations to maintain both security and ethical compliance throughout the training process.

User Data

This is perhaps the most regulated aspect of AI cybersecurity, falling under laws like GDPR in Europe or the Personal Information Protection Law in China. User data in AI systems can be particularly sensitive – people might share personal details, medical information, or business secrets when interacting with an AI assistant.

Protection can include end-to-end encryption securing data both in transit and storage, combined with strict data minimization principles to collect only essential information. User controls can provide transparent options for data management, including deletion rights and usage limitations.

The Human Element: People as the Strongest (and Weakest) Link

People can be both the strongest defense and the biggest vulnerability. Human error remains one of the biggest risks in cybersecurity. A single misplaced click, a carelessly shared password, or a fall for a phishing scam can potentially compromise even the most sophisticated security system.

This is why leading AI labs invest heavily in security training for all employees, not just their tech teams. It's about creating a culture of security awareness, where everyone understands their role in protecting these valuable assets.

Mechanisms for ensuring regulatory visibility

The Importance of External Scrutiny - As frontier AI systems become increasingly integrated into society and the economy, decisions about their training, deployment, and use will have far-reaching implications. It's crucial that these decisions are not left solely in the hands of AI developers.

External scrutiny – involving outside actors in the evaluation of AI systems through red-teaming, auditing, and external researcher access – offers a powerful tool for enhancing the safety and accountability of frontier AI.

To be effective, external scrutiny should adhere to the ASPIRE framework (Anderljung et al. 2023):

- Access: External scrutineers need appropriate access to the AI systems and relevant information.
- Searching attitude: Scrutineers should actively seek out potential issues and vulnerabilities.
- Proportionality to the risks: The level of scrutiny should be commensurate with the potential risks posed by the system.
- Independence: Scrutineers should be free from undue influence from the AI developers.
- Resources: Adequate resources must be allocated to support thorough scrutiny.
- Expertise: Scrutineers must possess the necessary technical and domain-specific expertise.

External scrutiny of AI systems can be structured in several ways, drawing from established practices in other regulated industries. One approach mirrors financial auditing, where certified professionals conduct standardized evaluations according to established protocols. This system can incorporate different levels of disclosure requirements, from basic safety testing to in-depth capability assessments. Some frameworks include external ethics boards within AI companies, though their authority and influence varies significantly. The effectiveness of these approaches often depends on how well they balance thorough oversight with the practical constraints of AI development timelines and resource limitations.

Responsible Reporting - One crucial aspect of both self-regulation and government oversight is the implementation of responsible reporting mechanisms. Organizations developing and deploying frontier AI systems have unique access to information about these systems' capabilities and potential risks. By sharing this information responsibly, they can significantly improve our collective ability to manage AI risks (Kolt et al. 2024).

Let's break down what responsible reporting might look like in practice:

What to Report

- Unexpected or potentially dangerous emergent capabilities
- Near-misses or safety incidents during development or deployment
- Significant breakthroughs in model performance or capabilities
- Observed misuse or attempted misuse of deployed models

Who to Report To

- Relevant regulatory bodies
- Industry consortiums focused on AI safety
- Academic researchers working on AI alignment and safety
- The wider public

How to Report

- Through secure, standardized reporting channels
- With appropriate protections for intellectual property and sensitive information
- In a timely manner, especially for urgent safety concerns

Different information sharing systems address the inherent tension between transparency needs and business interests in varying ways. Some approaches use tiered architectures that adapt disclosure levels to different stakeholder needs - regulators might receive detailed technical information while public disclosures remain more general. Other systems emphasize anonymization mechanisms that allow sharing of aggregate data while protecting individual company details. Legal frameworks sometimes include provisions to encourage honest reporting, such as liability protections for good faith disclosures.

Model registries - At its core, a model registry is a centralized database where information about AI models is recorded and tracked. It works like a birth certificate – when a model is deployed, its creators file some paperwork.

But what exactly goes into this paperwork? Different jurisdictions are taking different approaches, but model documentation typically encompasses several layers of information. Basic documentation often includes model identification and intended use cases, while technical specifications detail architecture, parameters, and computational requirements. Performance documentation can range from standard benchmark results to specialized evaluations of specific capabilities or risks. Impact assessments might examine potential societal effects, safety implications, and ethical considerations. Deployment documentation usually covers implementation strategies and monitoring plans.

The idea is that by collecting this information, regulators can keep tabs on the AI landscape, identify potential risks before they become problems, and have a foundation for more targeted governance down the line.

Why Model Registries Matter

Model registries can serve multiple roles in AI governance systems. As transparency mechanisms, they enable various degrees of independent scrutiny and public visibility and trust into AI development. Some registries function as early warning systems for emerging capabilities or risks, allowing for preemptive response to potential concerns - if a model is registered with capabilities that raise red flags, regulators can step in before it's widely deployed. The accumulated data can inform policy development by providing empirical evidence about AI system characteristics and trends. Instead of broad, one-size-fits-all rules, they can tailor their approach based on the specific capabilities and risks of different models. Finally, in contexts where AI capabilities have strategic significance, registries can help governments keep track of who's developing what, potentially informing export controls or other national security measures.

Governments around the world have already started to implement model registries. The U.S., for example, has taken a relatively light-touch approach so far, focusing primarily on the most advanced AI models. In October 2023, President Biden signed an Executive Order on AI that included provisions for a model registry. The United States has adopted an initially targeted approach to model registration, focusing oversight on the most advanced AI systems while maintaining flexibility for future expansion. This strategy, formalized in the October 2023 Executive Order, establishes clear compute-based thresholds for registration requirements. Systems exceeding 10^26 floating point operations in training must provide comprehensive documentation of their capabilities and limitations. They also need to disclose measures taken to protect their models from unauthorized access or theft.

China has taken yet another approach, focusing on algorithmic recommendation systems rather than AI models per se. Their Internet Information Service Algorithmic Recommendation Management Provisions, which came into effect in 2022, target systems based on their potential influence on public discourse and social behavior. This framework requires detailed registration of algorithms used across various digital platforms, with particular emphasis on algorithms with "public opinion properties" or "social mobilization capabilities.". Companies must disclose not just technical details but also the underlying principles and intended purposes of their algorithms, creating transparency around both capabilities and intentions.

Challenges

As you might imagine, the implementation of model registries hasn't been without its challenges:

1. Defining the Scope: One of the biggest challenges is determining which models should be subject to registration requirements. Set the bar too low, and you risk stifling innovation with excessive bureaucracy. Set it too high, and you might miss potentially risky systems.
2. Protecting Intellectual Property: AI companies invest enormous resources in developing their models and are understandably reluctant to share too much detail about their inner workings. Striking a balance between transparency and IP protection is a delicate act.
3. Enforcement and Compliance: How do you ensure companies actually comply with registration requirements? And what are the consequences for non-compliance?

A Know Your Customer regime for AI - In the financial sector, banks are required to implement Know Your Customer (KYC) schemes to identify and verify client identities. This helps prevent money laundering and other financial crimes. Similarly, we could implement a KYC scheme for frontier AI (Egan & Heim 2023). Under this scheme, compute providers would be required to implement KYC-like processes for their clients developing frontier AI models. If a company suddenly starts using an unusually large amount of compute power, this could trigger a reporting requirement. The compute provider would need to gather information about the nature of the project and report it to the relevant regulatory body.

This approach provides early warning of potentially problematic or sudden advancements in AI capabilities. It allows for nuanced and targeted export controls. It also offers more precise control over compute quantities and the flexibility to suspend access if necessary.

Implementing this regime would involve establishing a dynamic threshold of compute that effectively captures high-risk frontier model development, setting clear requirements for compute providers to keep records and report high-risk entities, and creating a government capacity to co-design, implement, administer, and enforce the scheme.

Incident reporting - AI incident reporting is a process where developers, companies, and sometimes even users report significant issues, near-misses, or incidents related to AI systems. These could range from privacy breaches and security vulnerabilities to unexpected biases in decision-making or large-scale material or human harms.

Incident reporting frameworks foster information-sharing about what went wrong (or almost went wrong), and thus creates a feedback loop that helps companies improve their systems and prevent similar issues in the future.

<iframe
src="https://ourworldindata.org/grapher/annual-reported-ai-incidents-controversies?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Global annual number of reported artificial intelligence incidents and controversies (Giattino et al., 2023)
</caption-iframe>

Learning from Other Industries: Aviation Safety

The Aviation Safety Reporting System (ASRS) in the United States is often held up as a gold standard for incident reporting (Cheng 2024). It's confidential, voluntary, and – crucially – non-punitive. This means that pilots, air traffic controllers, and other aviation professionals can report near-misses or safety concerns without fear of repercussions. The results speak for themselves: since the ASRS was implemented, aviation fatalities have plummeted.

This approach has fostered a culture of openness that enables continuous improvement through comprehensive data collection on near-misses and potential risks. The system's success stems from its focus on identifying systemic issues rather than assigning individual blame, creating a model that could be adapted for AI safety.

AI presents unique challenges that make incident reporting particularly tricky (Farrell 2024):

1. Defining an "incident": In aviation, it is clear what constitutes an incident or near-miss. But with AI, the lines can be blurry. Is an AI chatbot giving misleading information an incident? What about subtle algorithmic bias? Clear, agreed-upon definitions are needed to ensure the viability of incident reporting systems (OECD 2024).

2. Attribution and responsibility: AI systems often involve multiple stakeholders – developers, data providers, platform operators, and end-users. Determining who's responsible for reporting an incident (and potentially facing consequences) is not always straightforward.

3. Proprietary concerns: Companies invest millions in developing cutting-edge AI. They're understandably wary of sharing too much information about their systems.

Towards a Comprehensive AI Incident Reporting Framework

Implementing such a framework requires careful design to balance multiple competing needs (Farrell, 2024). The foundation must be built on precise, tiered definitions of incidents ranging from minor technical issues to catastrophic failures. This classification system would support a dual-channel reporting structure: mandatory reporting for severe incidents causing significant harm, and confidential channels for near-misses and minor incidents, providing a way for AI professionals to report concerns and minor incidents without fear of repercussions, potentially managed by a neutral third party to ensure confidentiality. The framework's effectiveness depends on standardized reporting formats that facilitate analysis while enabling rapid dissemination of critical information. This might include fields for system specifications, incident description, root cause analysis, and mitigation steps taken. Throughout the system, careful balance must be maintained between public transparency and commercial sensitivity to ensure both broad learning and continued industry participation.

Mechanisms for ensuring compliance

Licensing regime - One approach to compliance enforcement could be to implement a licensing regime for frontier AI models, similar to how nuclear power plants or pharmaceutical companies must be licensed to operate. Under this system, companies developing frontier AI models would need to obtain a license by demonstrating compliance with established safety standards.

This process would integrate detailed technical documentation requirements with ways to demonstrate the implementation of required safety measures (e.g. through a safety case, see Buhl et al. 2024), creating a continuous cycle of compliance and verification. Regular audits and inspections would ensure ongoing adherence to safety standards.

Another, complementary approach could be to grant enforcement powers to supervisory authorities. These authorities would have the power to conduct investigations, issue fines for non-compliance, and even halt the development or deployment of models deemed too risky. Let’s say a company is found to be developing a frontier AI model without implementing the required safety protocols. The supervisory authority could issue a cease-and-desist order, requiring the company to halt development until they can demonstrate compliance with safety standards.

Governing effectively often requires looking to other domains that have grappled with similar regulatory challenges. One particularly relevant example is the Federal Select Agent Program (FSAP) in the biosecurity domain (Anderson-Samways 2023).

The FSAP was established to regulate the possession, use, and transfer of biological select agents and toxins that could pose a severe threat to public health and safety. Like frontier AI, the biosecurity field deals with rapidly evolving technologies, potentially severe risks, and the need to balance safety concerns with scientific progress.

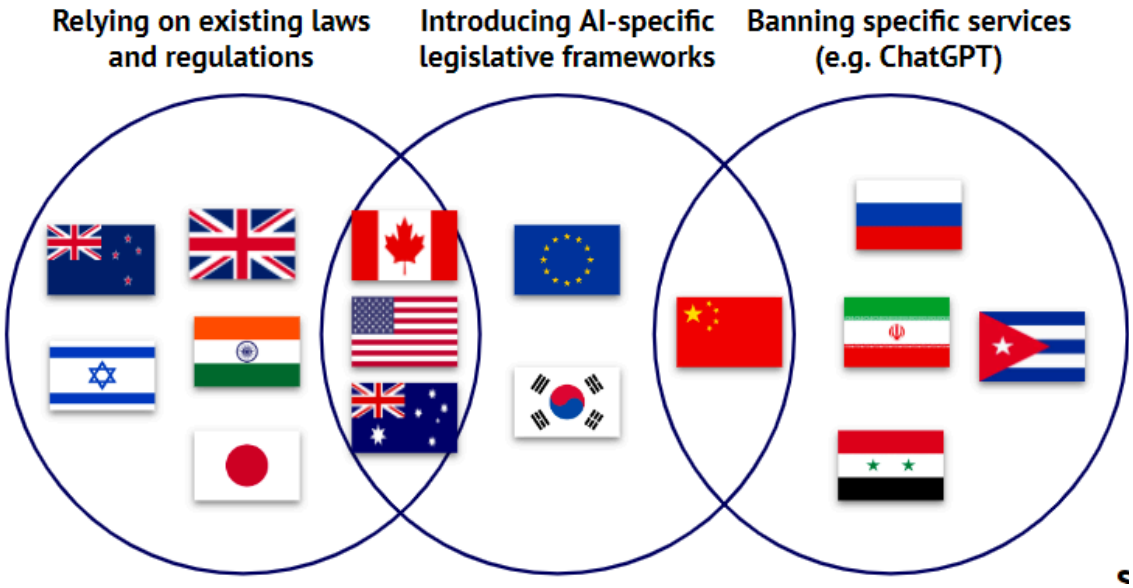
The FSAP employs a sophisticated risk-based regulatory system that begins during the research and development phase. Rather than waiting until biological agents are ready for use, the program requires registration and licensing early in the process - a model particularly relevant for AI governance, where early intervention may be crucial for managing risks.

Through continuous monitoring and regular inspections, the FSAP maintains ongoing visibility into research activities, enabling quick responses to evolving risks. This is complemented by a tiered regulatory framework that applies different levels of oversight based on an agent's risk profile. Such an approach could be particularly valuable for AI governance, where the vast spectrum of AI systems demands varying levels of scrutiny. The most powerful models would face stringent controls, while less capable systems could operate under lighter oversight, creating an efficient allocation of regulatory resources.

However, the FSAP also offers cautionary tales. Its reliance on checklist-based compliance in some areas has been criticized for potentially missing novel risks. This underscores the importance of maintaining a flexible, adaptive approach in AI governance.

The Architecture of AI Regulations

Creating AI-specific laws or relying on existing sectoral frameworks



<caption>
([State of AI report, 2023])
</caption>

Ex ante and ex post measures

A key consideration in AI governance is the balance between ex ante and ex post measures. Ex ante governance focuses on preemptive actions, setting rules and guidelines before potentially harmful AI systems are developed or deployed. This approach is particularly relevant for frontier AI, where the stakes are high and the potential for irreversible harm exists. Ex post governance, conversely, deals with the consequences of AI deployment, including liability frameworks and remediation measures. Effective AI governance requires a judicious mix of both approaches, anticipating potential issues while remaining flexible enough to address unforeseen challenges.

Vertical vs horizontal governance

The scope of governance measures also varies, with some targeting specific sectors (vertical regulation) and others applying broadly across multiple domains (horizontal regulation). Vertical approaches might focus on AI applications in healthcare or finance, tailoring governance to the unique challenges of each sector. Horizontal measures, such as data protection regulations or algorithmic transparency requirements, cut across sectors to address overarching concerns.

No single function or lever can adequately address the multifaceted challenges posed by frontier AI. Instead, effective governance requires a carefully orchestrated interplay of various mechanisms, adapting to the evolving capabilities of AI systems and the shifting societal and ethical landscapes they inhabit.

International Governance

The need for international governance

!!! quote "António Guterres (UN Secretary-General)"
<tab>

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

Can't individual countries just regulate AI within their own borders? The short answer is: no, not effectively.

There are several reasons why domestic governance alone is insufficient:

1. No monopoly on development: No single country has a monopoly on AI development. Even if the United States, for example, were to implement stringent regulations, AI developers in countries with laxer standards could still potentially create and deploy dangerous AI systems that could affect the entire world.
2. Global impact: The potential risks of advanced AI - from large-scale cyberattacks to economic disruption - are inherently global in nature. As James Cleverly, the UK Foreign Secretary, put it when discussing China's participation in the Bletchley AI Safety summit: "We cannot keep the UK public safe from the risks of AI if we exclude one of the leading nations in AI tech."
3. Race to the bottom: Without international coordination, countries may be reluctant to implement strict regulations unilaterally, fearing that they'll be left behind in the AI race. This can lead to a "race to the bottom" in terms of safety standards. International governance can help align incentives between nations, encouraging responsible AI development without forcing any one country to sacrifice its competitive edge.

<iframe
src="https://ourworldindata.org/grapher/cumulative-number-of-large-scale-ai-systems-by-country?tab=chart" loading="lazy" style="width: 100%; height: 600px; border: 0px none;" allow="web-share; clipboard-write"></iframe>
<caption-iframe>
Cumulative number of large-scale AI systems by country (Giattino et al., 2023)
</caption-iframe>

Current initiatives

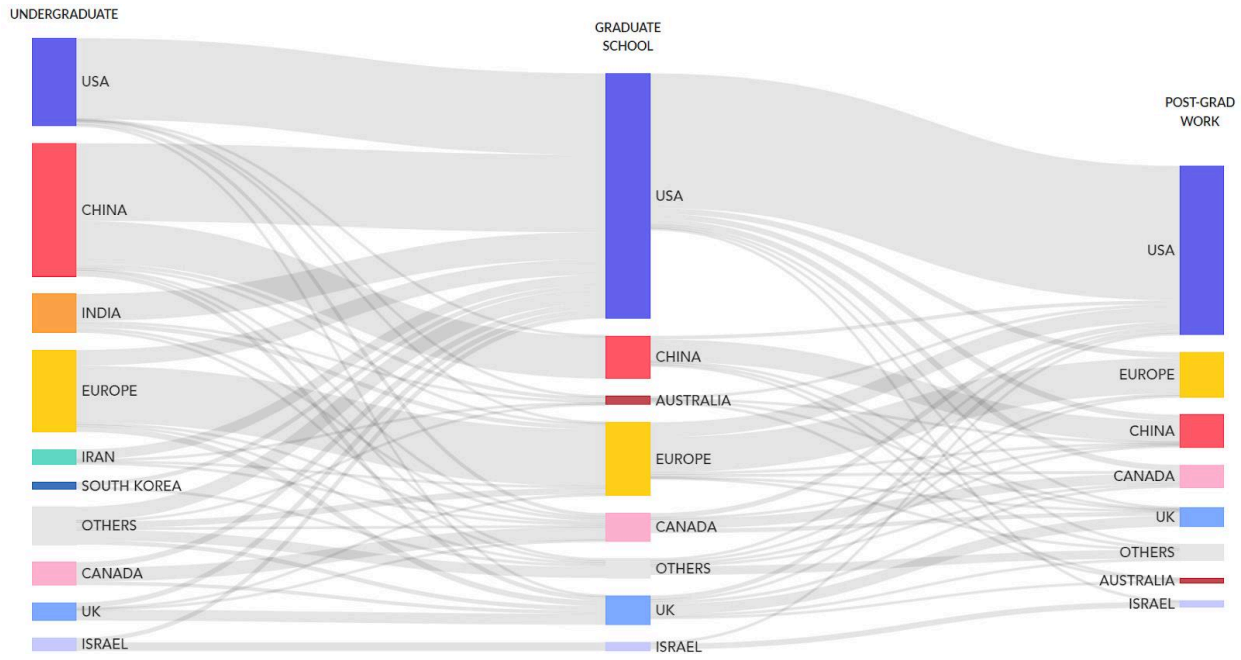
Global Impacts of National Regulations

!!! quote "Kamala Harris (Former US Vice President)"
<tab>
"[...] just as AI has the potential to do profound good, it also has the potential to cause profound harm. From AI-enabled cyberattacks at a scale beyond anything we have seen

before to AI-formulated bio-weapons that could endanger the lives of millions, these threats are often referred to as the "existential threats of AI" because, of course, they could endanger the very existence of humanity. These threats, without question, are profound, and they demand global action."
</tab>

The inherently global nature of technology development means that national policies can have far-reaching effects.

Even immigration policy is important:



<caption>
What are the career paths of top-tier AI researchers? (source: MacroPolo)
</caption>

For example, the United States' Executive Order on AI imposes reporting obligations on cloud providers, and export controls aimed at limiting China's access to advanced AI technologies. These actions, while originating from a single nation, have global implications.

Similarly, the European Union's AI Act is poised to have an impact far beyond the EU's 27 member states. Companies worldwide, eager to maintain access to the lucrative European market, often find it more cost-effective to adopt EU standards across their entire operations rather than maintaining separate standards for different regions.

For example, a U.S. tech company developing a new AI-powered facial recognition system for use in public spaces may see this system being classified as "high-risk"

under the EU AI Act. This would subject it to strict requirements around data quality, documentation, human oversight, and more. Companies then have a choice to make: develop two separate versions of your product – one for the EU market and one for everywhere else – or simply apply the EU standards globally. Many will be tempted to choose the second option, to minimize their cost of compliance. This illustrates what's known as the "Brussels Effect" (Bradford 2020): EU regulations can end up shaping global markets, even in countries where those regulations don't formally apply.

The Brussels Effect can manifest in two ways (Siegmann & Anderljung 2022):

1. De facto: Companies voluntarily adopt EU standards globally to avoid the complexity and cost of maintaining different standards for different markets.

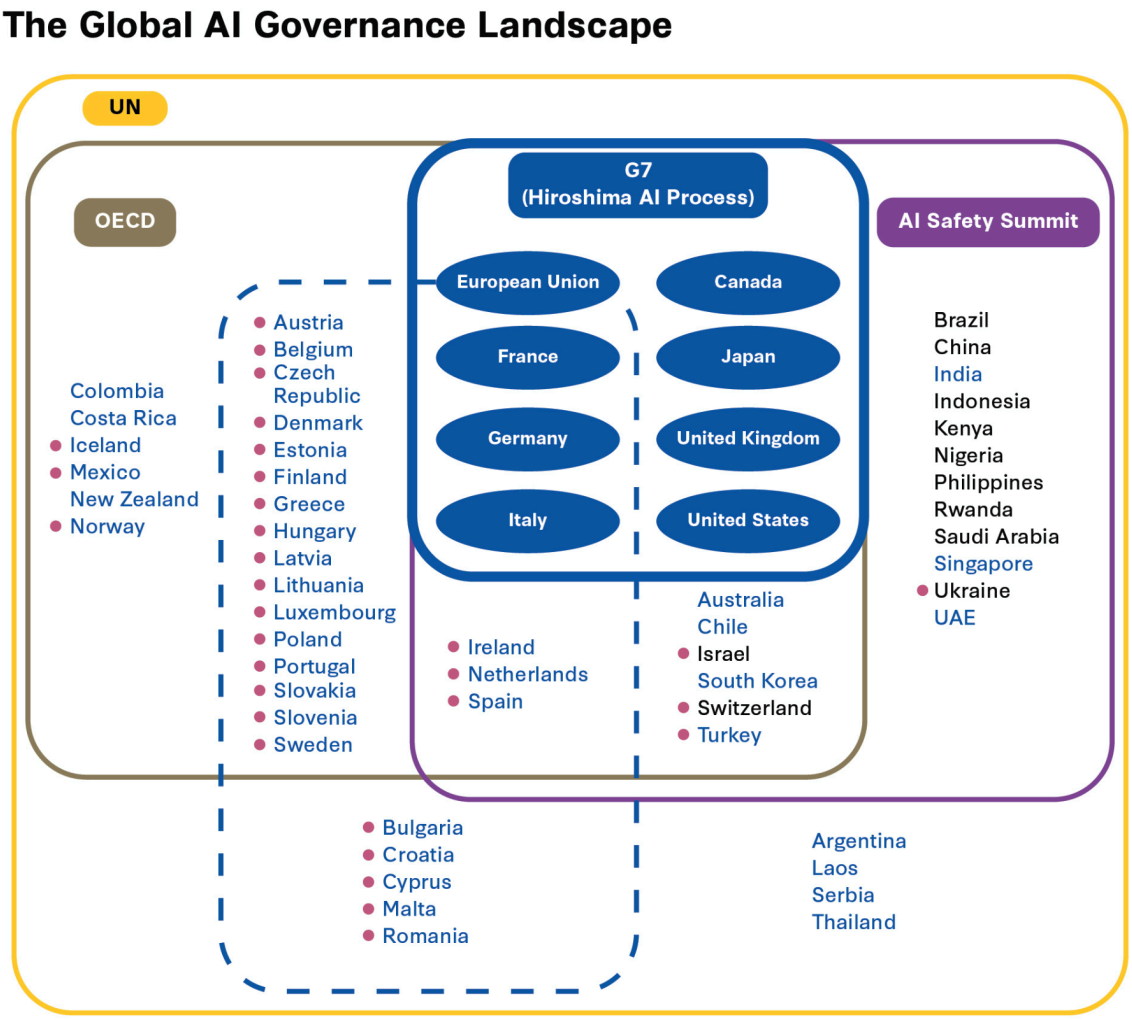
2. De jure: Other countries adopt regulations similar to the EU's, either to maintain regulatory alignment or because they view the EU's approach as a model to emulate.

For frontier AI, the Brussels Effect could be particularly significant. The EU's regulations might offer the first widely-adopted and mandated operationalization of concepts like "risk management" or "systemic risk" in the context of frontier AI. As other countries grapple with how to regulate advanced AI systems, they may look to the EU's framework as a starting point.

!!! quote "Ursula von der Leyen (Head of EU Executive Branch)"
<tab>
"[We] should not underestimate the real threats coming from AI [...] It is moving faster than even its developers anticipated [...] We have a narrowing window of opportunity to guide this technology responsibly."
</tab>

International initiatives

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



<caption>
The global ai governance landscape
</caption>

But it's not just individual nations taking action. A patchwork of international initiatives has emerged to address the governance of AI on a global scale:

- **The AI Safety Summit**: Held in the UK in 2023, this event brought together 28 nations and the EU to discuss AI safety. It resulted in the Bletchley Declaration, established AI Safety Institutes, and set the stage for future summits.
- **The Hiroshima AI Process**: Launched by the G7 nations, this initiative aims to promote responsible AI development and use.
- **United Nations efforts**: The UN is working on a report due in mid-2024 that will examine international institutions for AI governance.
- **OECD guidelines**: The Organisation for Economic Co-operation and Development has been particularly influential in shaping AI governance principles.

- **Council of Europe AI treaty**: This proposed treaty aims to protect human rights in the context of AI development and use.
- **China's Global AI Governance Initiative**: Demonstrating that AI governance is a priority even for nations often at odds with Western powers, China has put forth its own proposal for international AI governance.

<caption>
Cartoon highlighting a discrepancy between countries' statements and their true intentions in the context of the U.K.'s november 2023 AI Safety Summit (The Economist)
</caption>

Stages of International Policymaking

International policymaking typically progresses through several stages (Badie et al., 2011):

1. Agenda setting: Identifying the issue and getting it on the international agenda.
2. Policy formulation: Developing potential solutions and approaches.
3. Decision making: Choosing a course of action.
4. Implementation: Putting the chosen policy into practice.
5. Evaluation: Assessing the effectiveness of the policy and making adjustments as needed.

In the case of AI governance, we're still largely in the early stages of this process. The AI Safety Summit, for instance, represents a crucial step in agenda setting and initial policy formulation. But the real work of crafting binding international agreements and implementing them still lies ahead.

Policy options

Institutional Models. Various institutional arrangements could support international AI governance, from scientific consensus-building bodies to emergency response networks. These range from lighter-touch coordination mechanisms to more comprehensive frameworks for standard-setting and enforcement.

Non-proliferation. Drawing from nuclear weapons control strategies, non-proliferation approaches aim to limit access to advanced AI systems and critical resources like specialized chips. While these measures can help slow dangerous proliferation, they face significant challenges around enforcement and potential counterproductive effects on innovation.

Regulatory Agreements. International regulatory frameworks offer a collaborative path forward, where countries agree to develop AI safely and verify compliance through monitoring at the model, organizational, and jurisdictional levels. The jurisdictional certification approach provides one concrete model, leveraging market access as an incentive for participation.

Containment. For those concerned about catastrophic risks, more dramatic measures like the MAGIC plan propose centralizing advanced AI development in a single international facility. While politically challenging, historical precedents like early nuclear weapons control proposals suggest such radical approaches shouldn't be dismissed entirely.

Institutional Models for International AI Governance

As the international community grapples with how to govern frontier AI, a variety of institutional models have been proposed (Maas & Villalobos 2024):

- **Scientific Consensus-Building**: The Intergovernmental Panel on Climate Change (IPCC) was tasked with informing governments about the state of knowledge of climate change and its effects. A similar body could provide regular reports on AI capabilities and risks to policymakers and the public. Given the rapid pace of AI development, this body would need to be nimbler than traditional scientific consensus-building organizations.
- **Political Consensus-Building and Norm-Setting**: Building on scientific consensus, we might envision a forum for political leaders to discuss AI governance issues and develop shared norms and principles. This could take the

form of an AI-focused analogue to the United Nations Framework Convention on Climate Change (UNFCCC). Such a body could facilitate ongoing dialogue, negotiate agreements, and adapt governance approaches as the technology evolves.

- **Coordination of Policy and Regulation**: As countries develop their own AI regulations, there's a risk of a fragmented global landscape that could hinder innovation and create regulatory arbitrage opportunities. An international body focused on policy coordination could help address this challenge. Such an institution could work to harmonize AI regulations across countries, perhaps starting with areas of broad consensus and gradually tackling more contentious issues.
- **Enforcement of Standards and Restrictions**: For any international AI governance regime to be effective, there needs to be a mechanism for monitoring compliance and enforcing agreed-upon standards. This is where proposals like the jurisdictional certification approach discussed above come into play.
- **Stabilization and Emergency Response**: As we've discussed, the potential for "normal accidents" in AI systems is a serious concern. An international body focused on AI stability and emergency response could play a crucial role in mitigating these risks. This could consist in a global network of companies, experts and regulators, ready to assist in case of a major AI system failure or unexpected behavior. This group could also work proactively to identify potential vulnerabilities in global AI infrastructure and develop contingency plans. The International Atomic Energy Agency's Incident and Emergency Centre provides a potential model for this type of institution. However, given the potential speed of AI-related incidents, this body would need to operate on much faster timescales.
- **International Joint Research**: Collaborative international research could play a key role in ensuring that frontier AI development prioritizes safety and beneficial outcomes for humanity. An institution dedicated to facilitating such research could help pool resources, share knowledge, and ensure that safety considerations are at the forefront of AI development. CERN, the European Organization for Nuclear Research, offers one example for how such collaboration could work.
- **Distribution of Benefits and Access**: As frontier AI systems become more powerful, ensuring equitable access to their benefits will be crucial. An international institution focused on this challenge could work to prevent a harmful concentration of AI capabilities and ensure that the technology's benefits are widely distributed. This body might manage a global fund for AI development assistance, help facilitate technology transfers, or work to ensure that AI systems are developed with diverse global perspectives in mind.

!!! note "Learning from Nuclear Arms Control: Three Lessons for AI Governance"
<tab>

As we contemplate how to govern frontier AI on a global scale, it's instructive to look at how the international community has handled other powerful, potentially destructive technologies. Nuclear weapons provide a particularly relevant case study.

At first glance, nuclear weapons and AI might seem like very different technologies. One is a physical weapon of mass destruction, the other a general-purpose technology with immensely varied applications. But both share key characteristics: they're dual-use technologies with both civilian and military applications, and they have the potential to dramatically alter the global balance of power and pose significant risks.

So, what can we learn from decades of nuclear arms control efforts? Let's consider three key lessons (Maas 2019):

The Power of Norms and Institutions

In the early days of the nuclear age, many feared that nuclear weapons would proliferate rapidly, leading to widespread use. Yet today, nearly 80 years after the first nuclear detonation, only nine countries possess nuclear weapons, and they've never been used in conflict since World War II.

This outcome was the result of a taboo and concerted efforts to build global norms against nuclear proliferation and use. The Nuclear Non-Proliferation Treaty (NPT), signed in 1968, created a framework for preventing the spread of nuclear weapons while promoting peaceful uses of nuclear technology. We might envision similar norm-building efforts for AI.

The Role of Epistemic Communities

The development of nuclear arms control agreements wasn't solely the work of diplomats and politicians. It relied heavily on input from scientists, engineers, and other technical experts who understood the technology and its implications.

These experts formed what political scientists call an "epistemic community" – a network of professionals with recognized expertise in a particular domain. They played a crucial role in shaping policy debates, providing technical advice, and even serving as back-channel diplomats during tense periods of the Cold War.

One challenge to leveraging such networks for global AI governance will be ensuring that epistemic communities can effectively inform policy decisions. Unlike nuclear physicists, who were often employed directly by governments, many AI experts work in the private sector.

The Persistent Challenge of "Normal Accidents"

Despite decades of careful management, the nuclear age has seen several close calls – incidents where human error, technical malfunctions, or misunderstandings nearly led to catastrophe. Sociologist Charles Perrow termed these "normal accidents," arguing that in complex, tightly-coupled systems, such incidents are inevitable.

Applying the concept to AI, we could see unexpected interactions and cascading failures increase as AI systems become more complex and interconnected. Moreover, the speed at which AI systems operate could mean that a "normal accident" in AI might unfold too quickly for human intervention.

This reality challenges the notion of "meaningful human control" often proposed as a safeguard for AI systems. While human oversight is crucial, we must also design governance systems that are robust to the possibility of rapid, unexpected failures.

</tab>

Non-proliferation

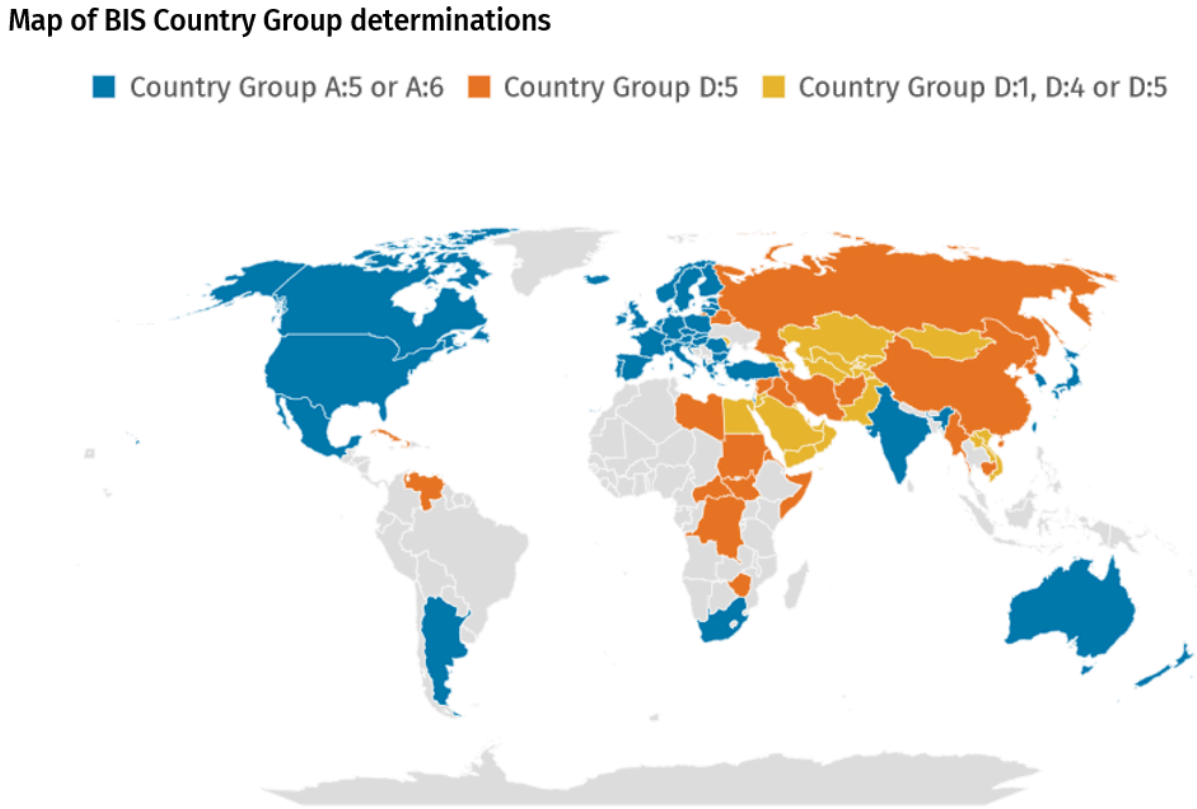
!!! quote "Demis Hassabis (Co-Founder and CEO of DeepMind)"
<tab>
"We must take the risks of AI as seriously as other major global challenges, like climate change. It took the international community too long to coordinate an effective global response to this, and we're living with the consequences of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

Non-proliferation, a term most commonly associated with nuclear weapons, refers to efforts to prevent the spread of dangerous technologies or materials.

In the context of AI, non-proliferation strategies aim to limit or control access to potentially dangerous AI systems or to the resources (like advanced computer chips) needed to develop them. This approach can be applied at both the national and international levels.

At the national level, this might mean only allowing companies with robust risk management procedures to access large-scale computing resources or training data. Internationally, it could involve preventing countries that lack adequate AI safety regulations from acquiring advanced AI capabilities.

This approach can help slow the spread of potentially dangerous AI technologies, giving responsible AI labs more time to develop safety methods and defensive technologies. It allows for a "pick a champion" strategy, where support is concentrated on responsible actors who are more likely to develop AI in a safe and beneficial manner.



<caption>
Map Of BIS (Bureau of Industry and Security) Country Group determinations (Rhodium Group)
</caption>

BIS is the "Bureau of Industry and Security", an entity within the US Department of Commerce in charge of export control policy. Depending on which category a country belongs to, it will have easier (in blue) or harder (in yellow and orange) access to US-made chips and chip-making equipment.

Non-proliferation strategies in AI can take several forms:

1. **Unilateral prevention**: This involves a country or group of countries taking steps to prevent other actors from acquiring AI models or key AI inputs. This could be applied to entire countries, specific entities like terrorist groups, or individual labs that don't meet certain safety standards.
2. **Protection against theft**: This strategy focuses on safeguarding AI models and technologies against stealing and unwanted tech transfer. Methods might include enhanced information security measures, security clearances for AI researchers, and strict controls on the sharing of sensitive AI research.
3. **Collaborative prevention**: This approach involves countries working together to prevent proliferation, primarily to non-state actors but potentially to other states

as well. An example of this could be a compute reporting regime, where cloud providers collect and share information about large-scale compute usage with regulators, who then share this information internationally to raise awareness of unwanted AI development activities.

Non-Proliferation in Action: U.S. Export Controls Towards China

A real-world example of non-proliferation strategies in AI is the United States' implementation of export controls targeting China's AI development capabilities (Allen 2022). Since October 2022, the U.S. has been working to block China's access to high-end chips from the U.S. and other countries, chip design software, semiconductor manufacturing equipment (SME), and even components required for producing SME.

These controls are being enforced with the cooperation of the Netherlands and Japan, who control key nodes in the global semiconductor supply chain.

It's important to note that these export controls aren't primarily about AI safety or even direct misuse of AI. They seem to be largely motivated by concerns about the use of advanced chips in weapons systems and a desire to prevent China from gaining economic (and thus geopolitical) dominance through AI.

While this is currently a unilateral policy, it has the potential to evolve into a bilateral or even multilateral arrangement through the implementation of verification mechanisms, such as through audits and inspections, which could be used to determine which companies might be added to a "white list" and thus allowed to receive advanced chips (NCUSCR 2023).

Non-Proliferation: Limitations and Challenges

Non-proliferation strategies in AI governance face complex challenges that stem from both technical and geopolitical realities. Historical evidence suggests these measures can produce unintended consequences that undermine their effectiveness. The U.S. experience with satellite technology export controls in the 1990s serves as a cautionary tale - restrictive policies led to a dramatic decline in U.S. market share from 73% to 25% over a decade, while simultaneously accelerating Chinese domestic capability development (Hwang & Weinstein 2022).

The technical landscape presents additional complications. Ongoing improvements in AI efficiency threaten to erode the effectiveness of compute-based controls as a governance mechanism (Pilz et al. 2023). Even assuming that compute-based controls remain relevant, it can be challenging to determine in advance which states will behave responsibly (for example by implementing adequate AI safety measures), making it hard to decide where to apply non-proliferation measures. Rather than preventing proliferation, restrictive measures can sometimes catalyze development races, as

evidenced by China's response to U.S. export controls through increased domestic AI investments and reciprocal control measures.

These practical challenges intersect with important moral considerations. Non-proliferation strategies often face criticism for their potentially discriminatory impact on technological and economic development across different nations. This perceived inequity can generate significant backlash, potentially undermining international cooperation necessary for effective AI governance. The challenge lies in developing approaches that can effectively manage proliferation risks while maintaining fairness and avoiding counterproductive outcomes in the global AI landscape.

Regulatory agreements

Given the limitations of unilateral non-proliferation strategies, many experts argue for a more collaborative approach through international regulatory agreements. The basic idea is simple: countries agree to develop AI safely and prove to each other that they're complying with agreed-upon safety standards and regulations.

These agreements can take many forms, varying in their level of legalization, number of participating states, and whether they involve the creation of new international organizations. The key is that they provide a framework for states to offer reliable evidence that they and their companies are developing AI responsibly.

When designing regulatory agreements for AI, there are three key levels to consider:

1. Model level: This involves setting standards and verification processes for individual AI models.
2. Organization level: This focuses on the AI development organizations themselves, ensuring they have proper safety protocols and risk management procedures in place.
3. Jurisdiction level: This is about the broader regulatory environment in a country or region, including laws, enforcement mechanisms, and oversight bodies.

LEVEL	Model	Organization	Jurisdiction
ACCOUNTABILITY TARGET	• "Model characteristics: • Information security • Performance • Truthfulness • Robustness • [...] • Documentation of model limitations and user instructions"	• "Quality management systems • Risk management systems • Organizational accountability & incentive structures • Data sourcing & model training • Testing & verification procedures • Documentation of design choices • Model access and dissemination strategies"	• Legislative and regulatory framework for accountable AI development (notably with regards to model- and organization-level auditing), including internationally agreed-upon rules and standards • Effective enforcement of relevant laws and regulations, including effective state authority over AI labs
VERIFICATION PROCESS	• "Performance-oriented methods: • Formal verification and benchmarking • Adversarial methods including red teaming and honeypotting"	• "Process-oriented methods: • Review of internal documentation • Interview with managers and software developers"	Process-oriented methods: • On site inspections of public authorities • Interviews with relevant stakeholders • Analysis of relevant legislation and regulations

<caption>
Accountability Targets and Verification Processes for Auditing AI models, organizations, and jurisdictions (Mökander et al. 2023)
</caption>

Most international agreements, especially in high-stakes domaines, operate at the jurisdiction level: it's typically easier for states to negotiate with each other than to directly regulate individual companies or products across borders.

!!! note "A Proposal for AI Regulatory Agreements: the jurisdictional certification approach"
<tab>

One potential model for AI regulatory agreements would involve the creation of an international organization that certifies jurisdictions for compliance with international AI safety standards, as proposed by Trager et al. 2023. These standards might include requirements for licensing AI developers, liability frameworks, the establishment of national AI regulators, and specific safety standards for AI development and deployment.

Under this model, AI labs would be monitored primarily by their national regulators. However, the international organization could also directly certify AI firms in countries that lack the resources or technical capacity to effectively regulate on their own. This approach has the advantage of encompassing all three levels (model, organization, and jurisdiction) while still allowing for some flexibility in how different countries implement the agreed-upon standards.

For any such agreement to be effective, there need to be strong incentives for countries to participate and comply. One powerful approach is to tie compliance to market access. For example, states could ban the import of goods that integrate AI from non-certified jurisdictions. They could also ban the export of AI inputs (like specialized chips) to non-certified jurisdictions.

To further strengthen enforcement, the agreement could require that states embed these enforcement provisions in their domestic laws as a condition of certification. This would provide all states with a strong incentive to join the regime and stay in compliance, as the economic costs of non-participation would be significant.
</tab>

While the idea of a global AI regulatory regime might seem far-fetched, there are actually existing international agreements that provide useful models.

The International Civil Aviation Organization (ICAO), a UN agency, audits state aviation oversight systems and publishes reports on each state's compliance with ICAO standards. In the U.S., the Federal Aviation Administration enforces these standards and can prohibit airlines from non-compliant countries from operating in the U.S.

The Financial Action Task Force (FATF) combats money laundering and terrorism financing. States agree on a set of standards, and the FATF monitors progress. Countries that don't have or enforce the necessary regulations can be put on a blacklist, significantly impacting their ability to attract international investment.

These examples show that it's possible to create effective international regulatory regimes, even in areas that touch on sensitive issues of national security and economic competitiveness.

The Security-Transparency Tradeoff

One of the key challenges in designing any international regulatory agreement for AI is balancing the need for verification with concerns about revealing sensitive information. This is known as the security-transparency tradeoff (Coe & Vaynman 2019).

On one hand, ensuring adherence to safety measures requires some form of verification. This might involve inspectors checking safety measures in a country's labs, inspecting AI models, or monitoring compute usage. There's also a need for broader monitoring to prevent evasion of the rules – for example, tracking the locations of data centers or the sale of specialized AI chips.

On the other hand, states may be reluctant to accept such intrusive inspections. There are concerns about sovereignty costs – the idea that allowing foreign inspectors into sensitive facilities impinges on a state's independence. There are also worries about proliferation risks: inspectors could potentially gain access to valuable intellectual property and transfer this information to other countries or companies.

This security-transparency tradeoff is a key reason why arms control agreements have been relatively rare historically (Coe & Vaynman 2019). Finding the right balance between verifying compliance and protecting sensitive information is crucial for the success of any AI governance agreement.

The jurisdictional certification approach described earlier offers one potential solution to this dilemma by allowing states to monitor their own labs while still providing assurance to the international community. However, more innovative technical solutions may also help to reduce this tradeoff.

!!! note "A Proposal for a Verification Mechanism: Catching a Chinchilla"
<tab>
One intriguing proposal for verifying compliance with AI development agreements while maintaining privacy comes from the paper "What Does It Take to Catch a Chinchilla?" (Shavit 2023).

The goal of this proposal is to "provide governments high confidence that no actor uses large quantities of specialized ML chips to execute a training run in violation of agreed rules" while maintaining the privacy and confidentiality of models and data.

The proposal has three main components:

1. Using on-chip firmware to occasionally save snapshots of the neural network weights stored in device memory, in a form that an inspector could later retrieve.
2. Saving sufficient information about each training run to prove to inspectors the details of the training run that resulted in the snapshotted weights.

3. Monitoring the chip supply chain to ensure that no actor can avoid discovery by amassing a large quantity of untracked chips.

While this proposal is not yet technically feasible, the authors argue that it presents only "narrow technical challenges" and could potentially provide a way to verify compliance with AI development agreements without revealing sensitive information about models or training data.
</tab>

While regulatory agreements offer a promising approach to international AI governance, they're not without their limitations.

The relationship between agreement effectiveness and political feasibility creates a central dilemma - the more robust the safety measures an agreement proposes, the more resistance it typically encounters from participating nations. This tradeoff between feasibility and effectiveness echoes throughout the history of international technology governance, particularly in cases like nuclear non-proliferation.

The timeline challenge compounds these difficulties. The development of the International Atomic Energy Agency's oversight capabilities serves as a sobering example - it required over two decades from the first use of nuclear weapons to establish meaningful inspection powers. In the context of AI's rapid advancement, such lengthy implementation periods could render agreements obsolete before they become operational.

The inherent difficulty of verifying compliance without exposing sensitive technological information creates additional complexity. Unlike physical technologies, AI development often leaves few observable traces, making traditional verification approaches insufficient. Finally, AI is a rapidly evolving field, and any regulatory agreement needs to be flexible enough to adapt to new developments.

Containment

For those who believe that catastrophic risks from AI are likely in the near future, more radical approaches to governance might seem necessary. One such approach is the idea of containment or technological restraint. The basic idea behind containment is to slow down or pause the development of advanced AI. This could serve two strategies (Maas 2022):

- **Delay**: giving more time for society to adapt and for alignment research to catch up with capabilities
- **Restraint**: if safe alignment is deemed very unlikely, or if there's no way to ensure alignment techniques will be used, restraint might be necessary to prevent catastrophic outcomes.

The 'MAGIC' Plan

One specific proposal for containment is the "MAGIC" (Multinational AGI Consortium) plan (Hausenloy et al. 2023). The core idea of MAGIC is to monopolize the development of advanced AI above a given compute threshold in a single facility, combined with a moratorium on development outside of this facility.

Under this plan, signatory countries would mandate cloud computing providers to prevent any training runs above a specific size within their national jurisdictions. The rationale is that advanced AI systems can be dangerous even before deployment, due to risks like theft, deceptive alignment, or power-seeking behavior.

The MAGIC plan proposes several key features to address the challenges of advanced AI development. At its core, it would establish a single, exclusive facility with a global monopoly on advanced AI model creation. This centralization aims to prevent a dangerous proliferation of powerful AI systems. The facility would prioritize safety, focusing on developing AI architectures that are inherently secure and exploring methods to constrain existing AI systems within safe boundaries. To protect its critical work, the facility would implement stringent security measures. Down the line, as safe advanced AI systems are developed, the consortium could distribute equitably the benefits of AI advancements among all participating nations.

Despite its ambitious approach to mitigating AI risks, the MAGIC plan faces substantial hurdles. The most significant challenge lies in its political feasibility. Convincing nations to relinquish their independent AI development capabilities would be extraordinarily difficult, given the perceived strategic and economic advantages of leading in AI technology. The institutional design of such a facility presents another major obstacle. Creating a governance structure that remains impartial and resistant to the influence of competing national interests would require unprecedented levels of international cooperation and trust. There are also concerns about the concentration of power inherent in the plan. Centralizing advanced AI development in a single location could create a potential single point of failure or abuse, especially if the facility's management doesn't maintain true multilateral representation. Lastly, the plan's reliance on compute-based thresholds for defining "advanced" AI may prove problematic in the long term. As AI algorithms become increasingly efficient, the correlation between computational power and AI capability may weaken, potentially rendering this aspect of the plan less effective over time.

While proposals like MAGIC might seem far-fetched, history shows us that radical schemes for international control of dangerous technologies can gain surprising traction when the stakes are high enough. The development of nuclear weapons provides an illuminating parallel.

In the immediate aftermath of World War II, as the world grappled with the implications of atomic weapons, there was a serious push for international control of nuclear technology. The 1946 Acheson-Lilienthal Plan, which formed the basis of official U.S. policy at the time, proposed a radical solution: A new U.N. authority would "control all fissionable raw materials and have a monopoly on all dangerous, i.e., military activities" (Zaidi & Dafoe 2021) States would shut down all military nuclear activities, keeping only nuclear power plants, which would be inspected by the U.N. authority.

This plan, while ultimately not implemented, demonstrates that even the most powerful nations can seriously consider surrendering control of strategically crucial technologies in the face of catastrophic technological risks.

Moreover, as pointed out by Maas, "States can and will unilaterally forego, cancel, or abandon strategically promising technologies for a range of mundane reasons". (Maas, 2023) In the case of nuclear weapons, an estimated 14 to 22 nuclear weapons programs were considered but left unpursued, and 7 programs were pursued but later abandoned.

This historical precedent suggests that while containing AI development through international agreement would be extremely challenging, it's not entirely outside the realm of possibility, especially if the risks become more apparent and immediate.

Where Do We Go From Here?

As we've explored, there are several potential approaches to the international governance of frontier AI:

1. Non-proliferation: Limiting access to dangerous AI systems or the resources needed to develop them.
2. Regulatory agreements: Providing reliable evidence that states and companies are developing AI safely.
3. Containment: Monopolizing advanced AI development in a single, internationally controlled facility.

These approaches aren't mutually exclusive. In fact, managing advanced AI will likely require a combination of strategies operating at different levels. For example, governments could cooperate with like-minded states on regulatory agreements while simultaneously pursuing non-proliferation strategies to slow the spread of advanced AI capabilities to less responsible actors.

The path forward will depend on how the AI landscape evolves, how our understanding of AI risks develops, and how the international political climate shifts. Regardless of the specific approach, it is clear that some form of international

governance will be crucial for managing the risks and harnessing the benefits of frontier AI.

The design of effective AI governance frameworks must navigate several fundamental tradeoffs. A central tension exists between effectiveness and political feasibility - while stronger obligations might better mitigate risks, they become increasingly difficult for states to accept and implement. This challenge is mirrored in the relationship between participation and commitment depth, where broader participation often comes at the cost of weaker commitments. Deciding whether to prioritize wide participation or strong commitments is a key strategic choice.

These structural tensions are further complicated by dynamic considerations. Any governance framework must maintain legitimacy through inclusive stakeholder representation while remaining adaptable enough to respond to rapidly evolving AI capabilities. Finally, agreements must enable compliance monitoring without compromising sensitive information about AI development.

!!! note "Under Which Conditions Will States Desire and Accept International Governance?"
<tab>
Understanding when states might be willing to participate in international AI governance is crucial for designing effective arrangements. The factors influencing this willingness can be broadly categorized into desirability and feasibility factors. Desirability factors are those that determine a state's desire to be assured that AI is being developed safely in other countries. Feasibility factors are those that would prevent a state from fulfilling its desire for assurance, i.e. from accepting an international agreement, even if the desire for assurance exists.

In terms of desirability, several key elements come into play. First and foremost, states need to recognize that AI poses risks significant enough to warrant international cooperation. This awareness of extreme risks is fundamental to motivating action on a global scale. Additionally, states may want to ensure that other countries implement regulations, so that they can themselves regulate AI domestically without being left behind economically or technologically. Finally, a lack of trust in other countries' AI development practices could drive states towards international governance. If nations doubt the safety protocols or ethical standards of their counterparts, they may view collaborative oversight as a necessary safeguard.

Feasibility factors are equally important in determining the viability of international agreements for AI safety. The cost of risk-reducing measures plays a crucial role; the lower the economic and strategic costs of proposed safety standards and obligations, the more likely states are to accept them. Proposals that build on or align with existing regulatory frameworks or international agreements are also more likely to gain acceptance, as they require less dramatic shifts in policy and practice. Interestingly, the

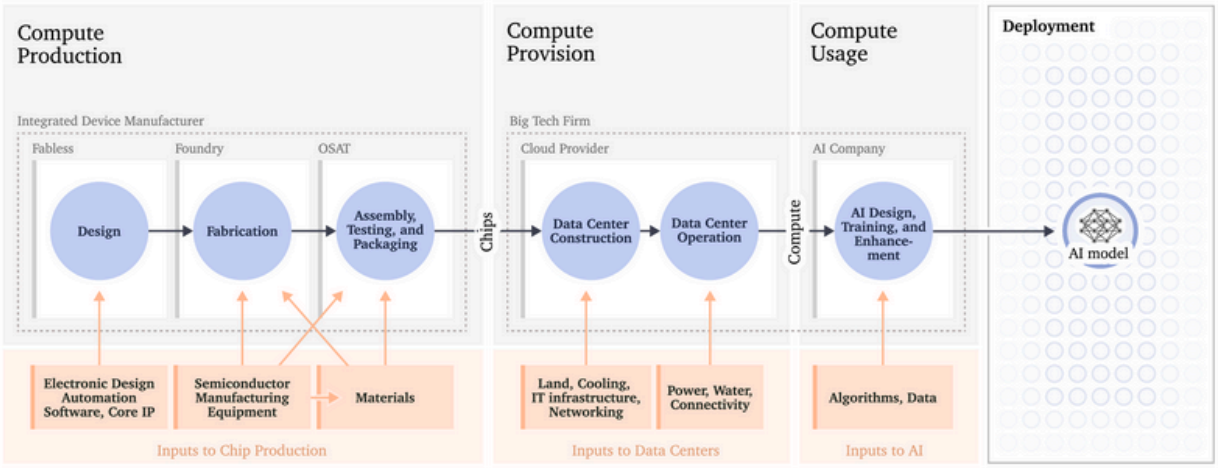
potential for competitive advantage can be a motivating factor. If states believe that adhering to safety regulations could give them an edge in the global market by fostering trust in their AI products, they may be more willing to participate. Verification costs and mechanisms represent another critical feasibility factor. The availability of verification methods that don't reveal strategically valuable information can make agreements more palatable to states concerned about maintaining their competitive edge or national security. Moreover, the expected compliance by other states significantly influences participation willingness. Nations are more likely to commit to international governance if they believe their counterparts will adhere to the agreed-upon standards.

Several other important factors can influence a state's willingness to engage in international AI governance. These include the number of actors involved, as broader participation can lend legitimacy and effectiveness to the effort. The presence of powerful states willing to take a leadership role can also be pivotal, as it can provide momentum and resources to the initiative. For less-resourced countries, the availability of technical aid can be a crucial factor in their ability and willingness to participate. Finally, the credibility of incentives or threats associated with participation can significantly impact a state's decision-making process: well-designed mechanisms can encourage other countries' compliance and deter their non-participation.
</tab>

There are also reasons for cautious optimism. Historical precedents like nuclear non-proliferation agreements show that international cooperation is possible even in areas of critical strategic importance. The emergence of various international AI initiatives demonstrates a growing recognition of the need for global coordination.

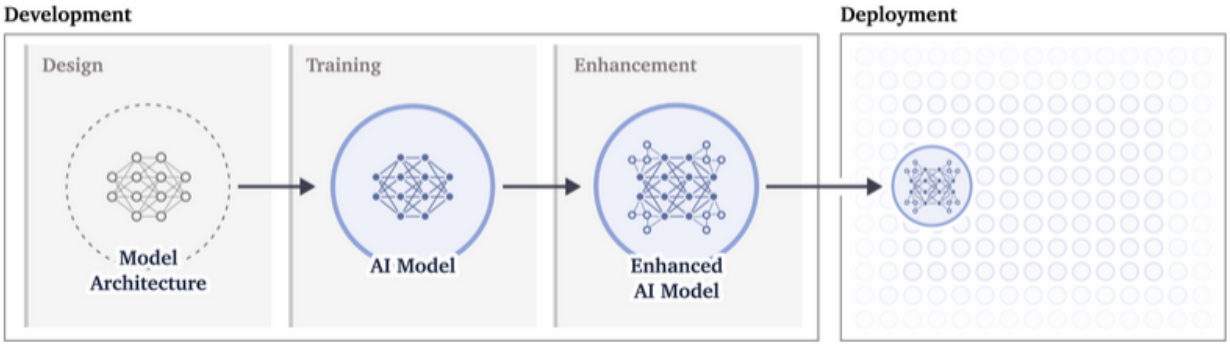
Moving forward, progress in AI governance will likely come through a combination of approaches: strengthening domestic regulations, fostering international cooperation through agreements and institutions, and potentially exploring more radical containment strategies if risks become more acute.

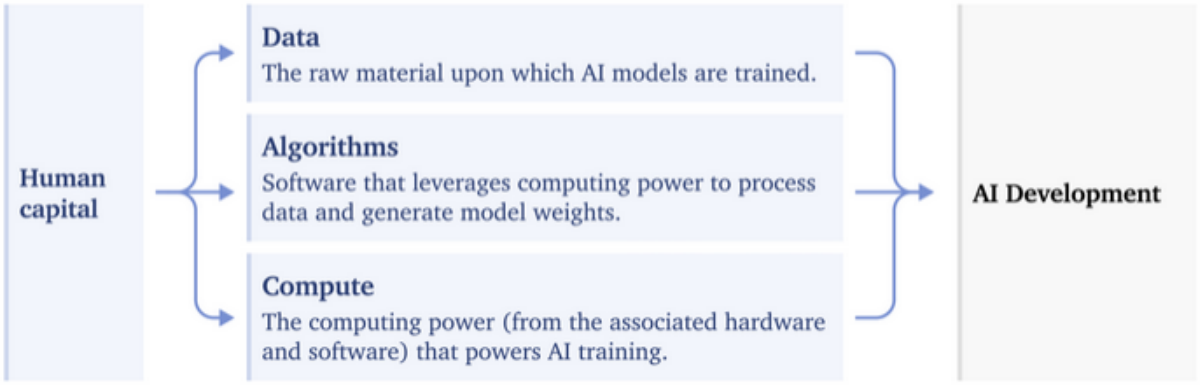
Images

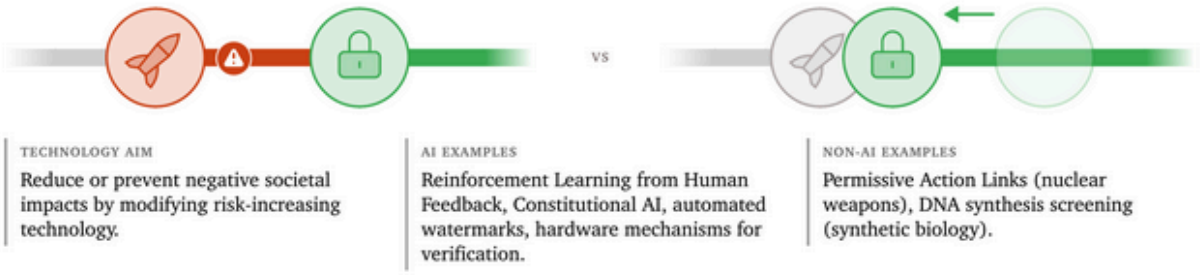


	Inputs				Output
	Compute	**Data**	**Algorithms**	**Talent** Secondary input	**Trained AI Model**
Detectability	High	Low	Low	Medium	Low
Excludability	High	Medium	Low	Medium	Medium
Quantifiability	High	Medium	Medium	Low	Low
Supply chain concentration	High	Low	Medium	Medium	High

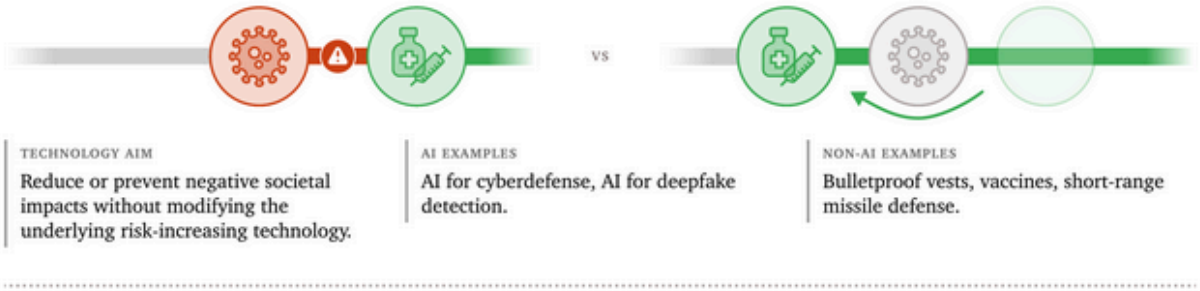
Properties



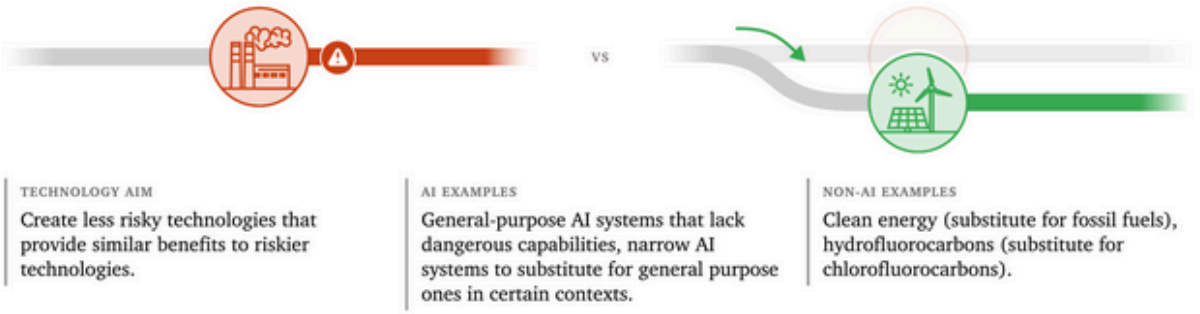




Defensive Technologies before risk-increasing technologies



Substitute Technologies instead of risk-increasing technologies



Can go into strategies, d/acc