CHAPTER 4

Governance

Charles Martinet
French Center for AI Safety (CeSIA)

Markov Grey
French Center for AI Safety (CeSIA)

Su Cizem
French Center for AI Safety (CeSIA)

Charbel-Raphaël Segerie
French Center for AI Safety (CeSIA)

Contents

1. Introduction

> *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.*
>
> **The Bletchley Declaration** **2023**
> Signed by 28 countries, including all AI leaders, and the EU, 2023

Artificial Intelligence has the potential to revolutionize numerous aspects of society, from healthcare to transportation to scientific research. Through the previous chapters you have seen AI's ability to defeat world champions at Go, generate photorealistic images from text descriptions, and even 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). Technical AI safety research is necessary to ensure AI behaves reliably and aligns with human values, especially as it becomes more capable and autonomous. Even though technical research is necessary it alone is not sufficient 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. As a reminder frontier AIs are highly capable models that could possess dangerous capabilities sufficient to pose severe risks to public safety (Anderljung et al., 2023). Although in recent history many state of the art advancements have been driven by LLMs or foundation models , frontier AI as a term is not limited to just these types of models. 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. We will focus 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.

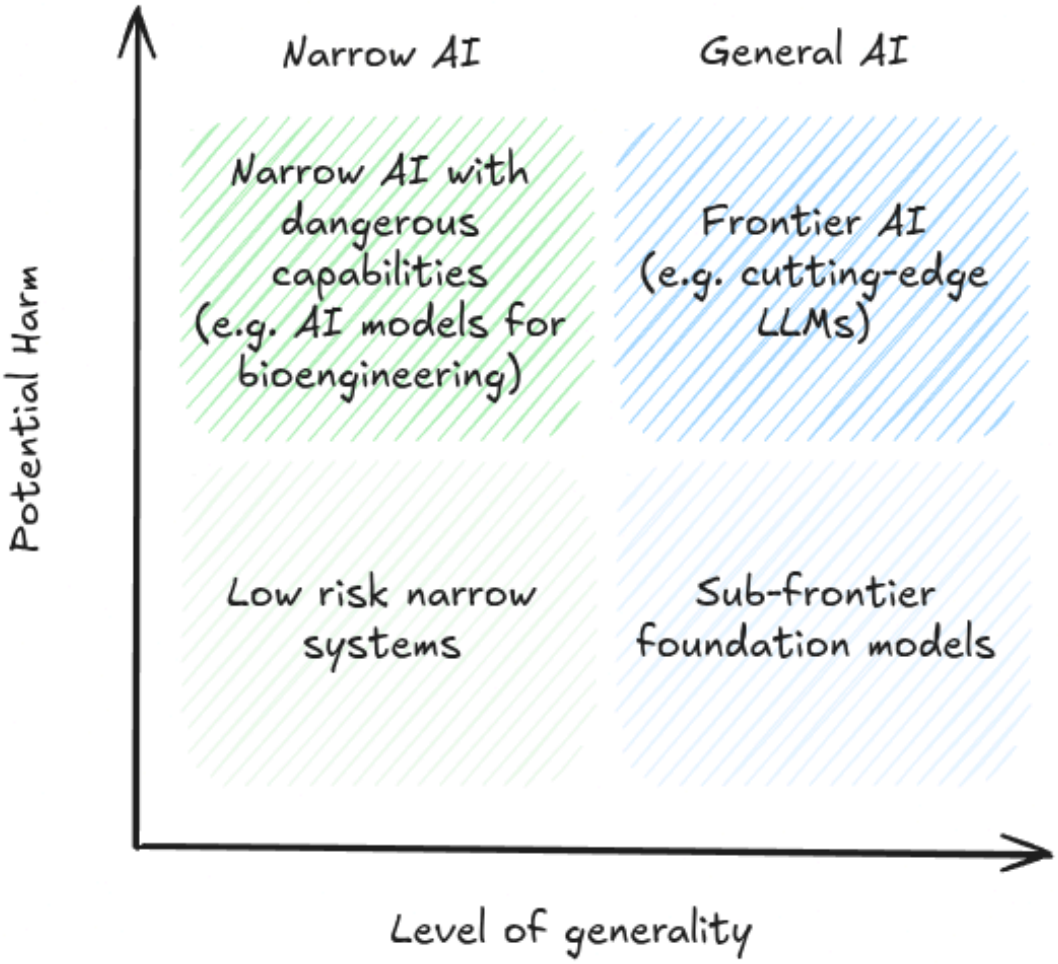

Figure 1: *Distinguishing AI models according to their level of potential harm and generality. We focus here on frontier AI models (U.K. government, 2023).*

2. Governance Problems

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.

The combination of AI as a general-purpose, information, intelligence technology creates unique 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.

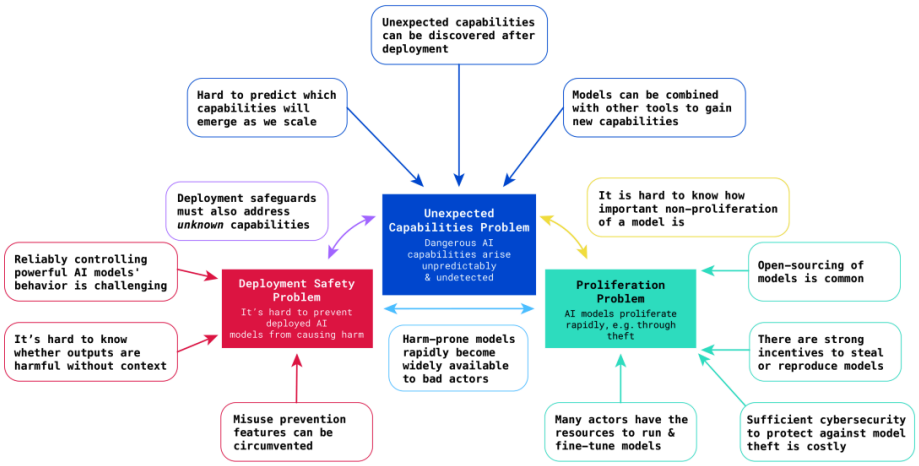

Figure 2: *Summary of the three regulatory challenges posed by frontier AI (Anderljung, 2023)*

2.1 Unexpected Capabilities

AI systems develop surprising abilities that weren't part of their intended design. Through several of our chapters now, we have shown that foundation models can show "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). Evaluations have found that frontier models can autonomously conduct basic scientific research, hack into computer systems, and manipulate humans through persuasion, none of which were explicitly trained for (Phuong et al., 2024 ; Boiko et al., 2023 ; Turpin et al., 2023 ; Fang et al., 2024).

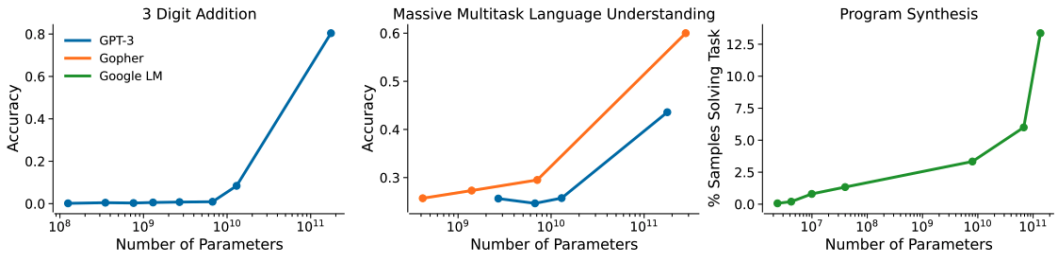

Figure 3: *Example of unexpected capabilities. Graphs showing several metrics that improve suddenly and unpredictably as models increase in size (Ganguli et al., 2022)*

AI evaluations are still in their early stages in 2025. 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).

2.2 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).

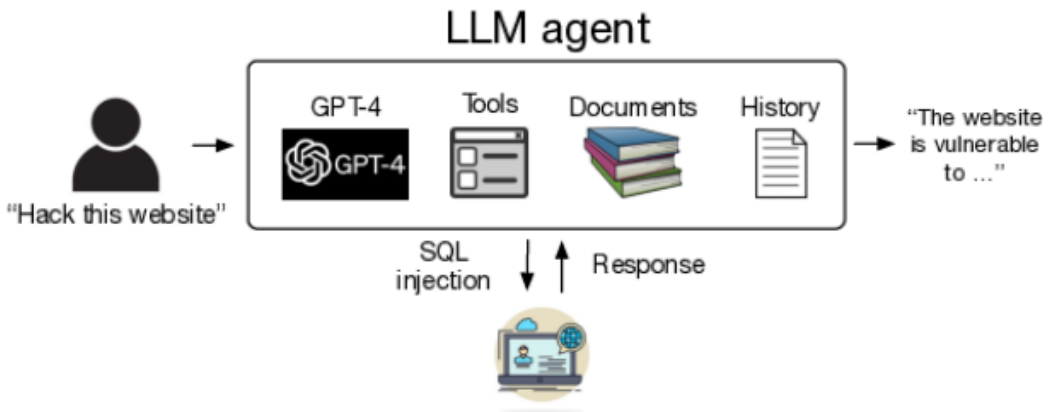

Figure 4: *A schematic of using autonomous LLM agents to hack websites (Fang et al., 2024). Once a dual-purpose technology is public, it can be used for both beneficial and harmful purposes.*

AI agents amplify 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).

2.3 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 5: *Examples of Proliferation (Özcan, 2024).*

2.4 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)

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.

3. Compute Governance

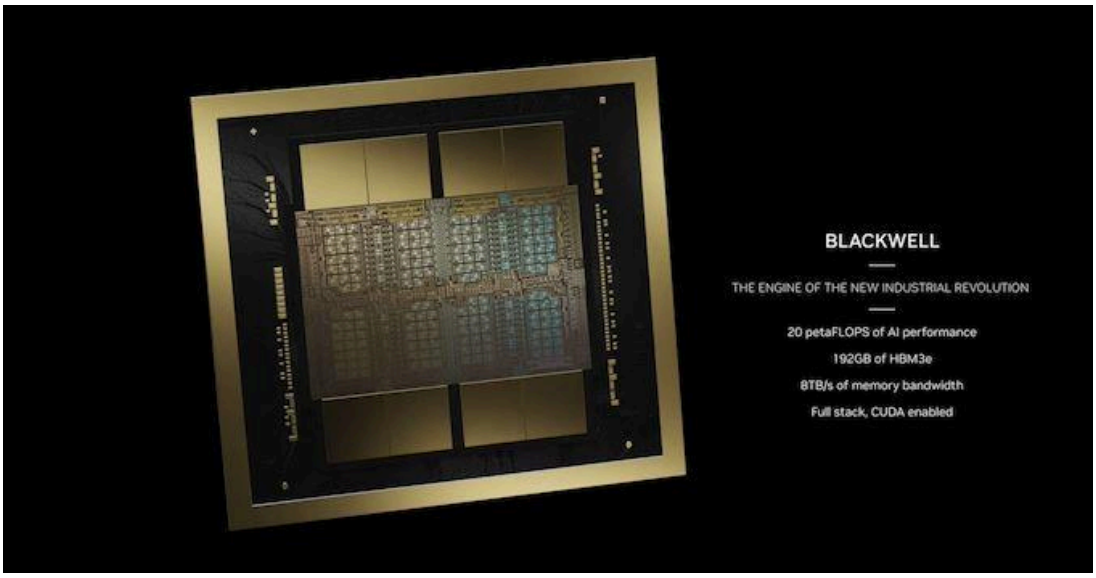

Figure 6: *Example of an NVIDIA Blackwell B100 accelerator (2025). Each B100 carries 192 GB of HBM3e memory and delivers nearly 20 PFLOPS of FP4 throughput, roughly doubling the performance of the H100 from 2024 (NVIDIA, 2025).*

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).

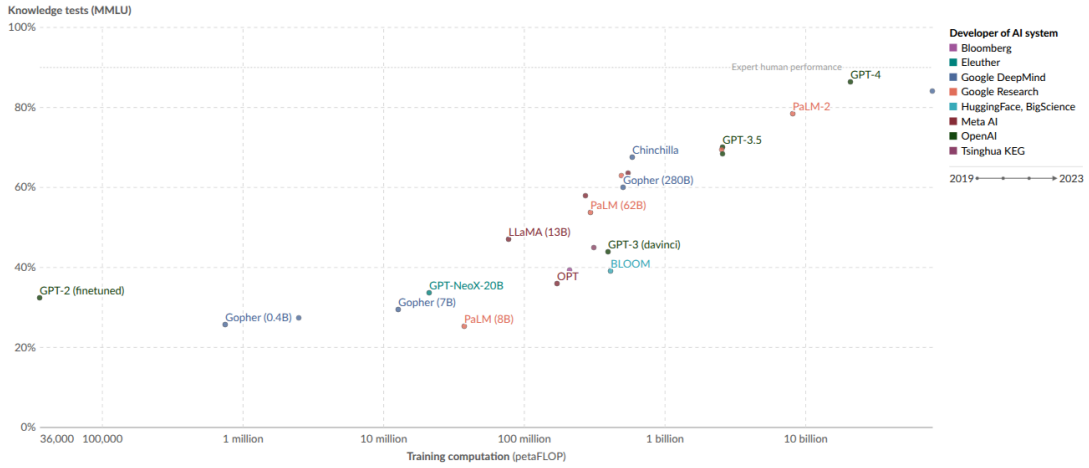

Figure 7: *Performance on knowledge tests vs. training computation. Performance on knowledge tests is measured with the MMLU benchmark, here with 5-shot learning, which gauges a model's accuracy after receiving only five examples for each task. Training computation is measured in total petaFLOP, which is 1e15 floating-point operations (Giattino et al., 2023).* (interactive version on website)

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.

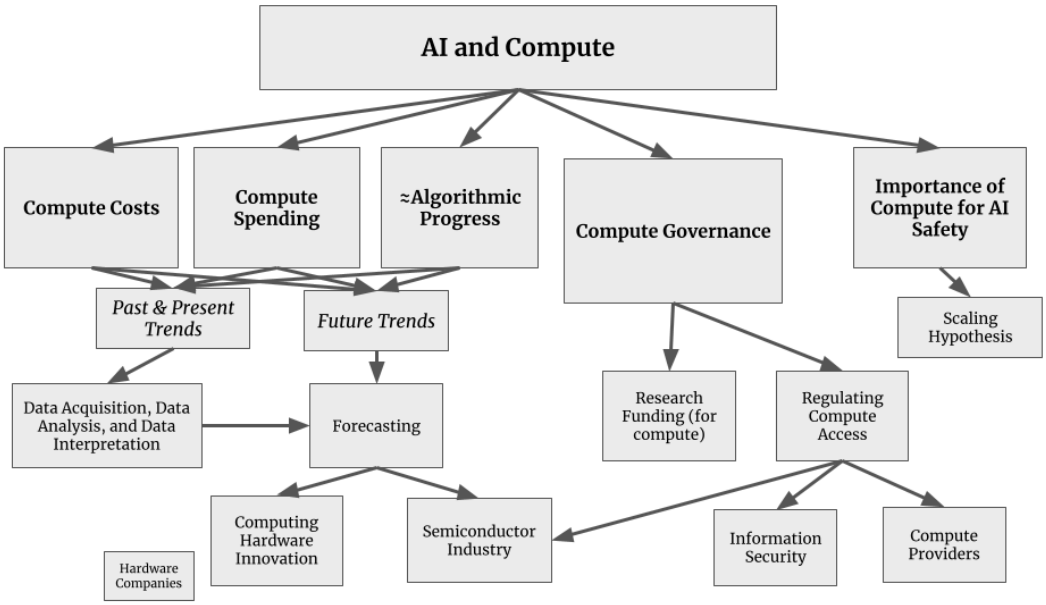

Figure 8: *Sketch of research domains for AI and Compute (Heim, 2021).*

3.1 Tracking

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).

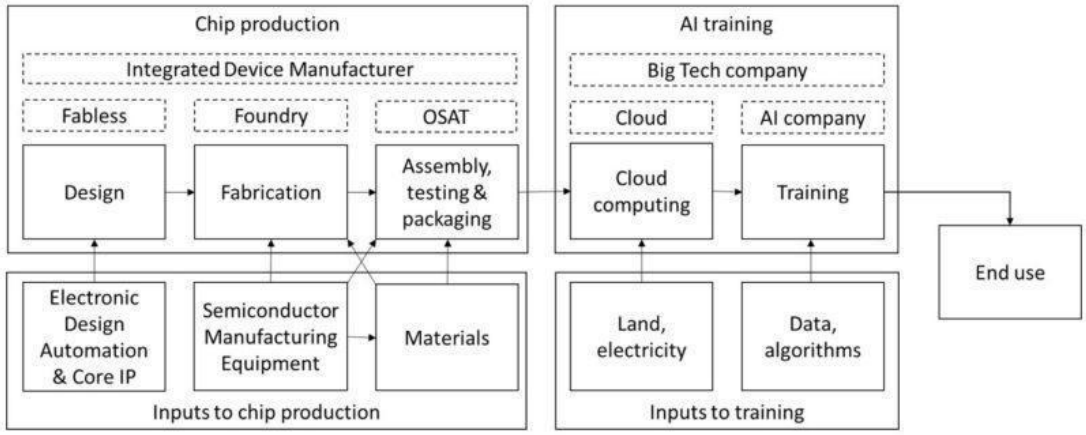

Figure 9: *The compute supply chain (Belfield & Hua 2022).*

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).

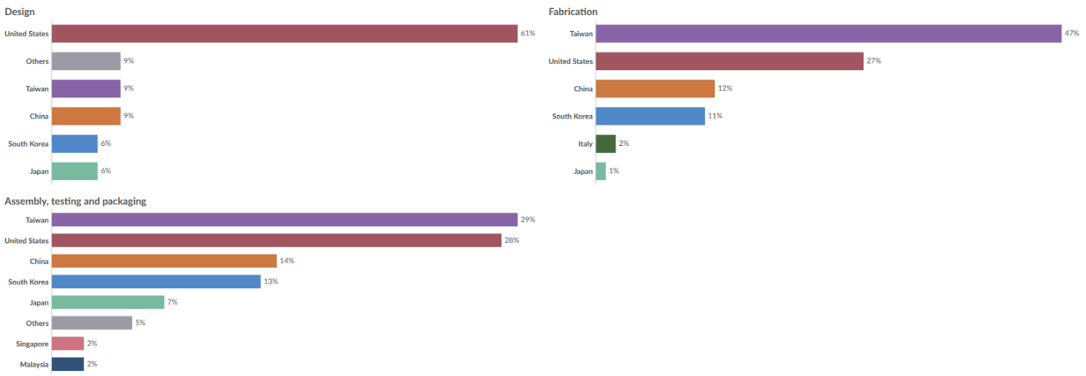

Figure 10: *Market share for logic chip production, by manufacturing stage (Giattino et al., 2023).*
(interactive version on website)

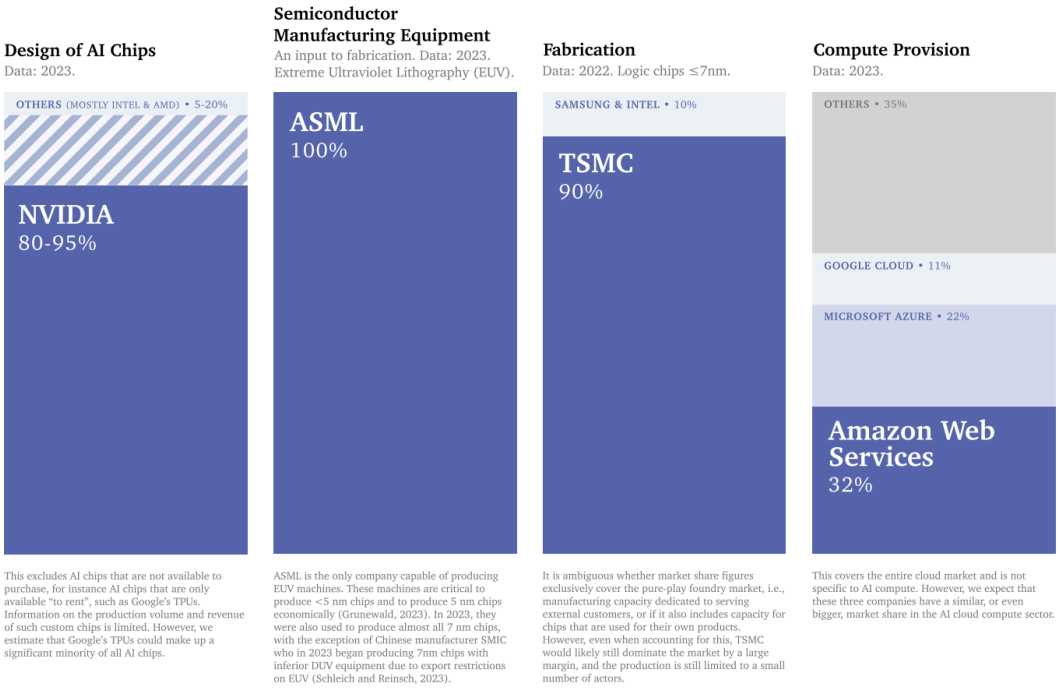

Figure 11: *Concentration of the AI Chip Supply Chain Expressed as percentage of total market share (Sastry et al., 2024).*

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.

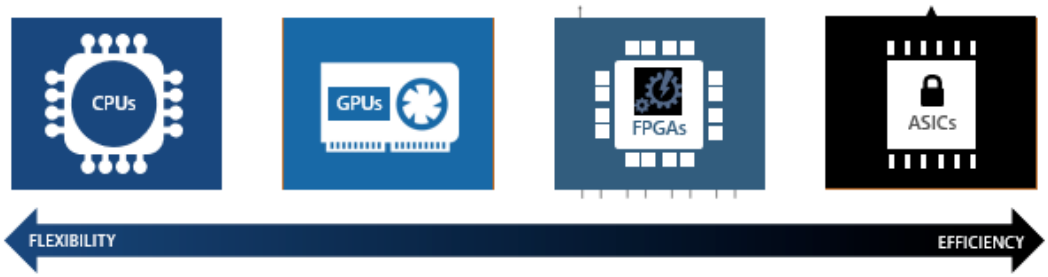

Figure 12: *The spectrum of chip architectures with trade-offs in regards to efficiency and flexibility.*

Rather than trying to control all computing infrastructure, governance should 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 hardware.

3.2 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.

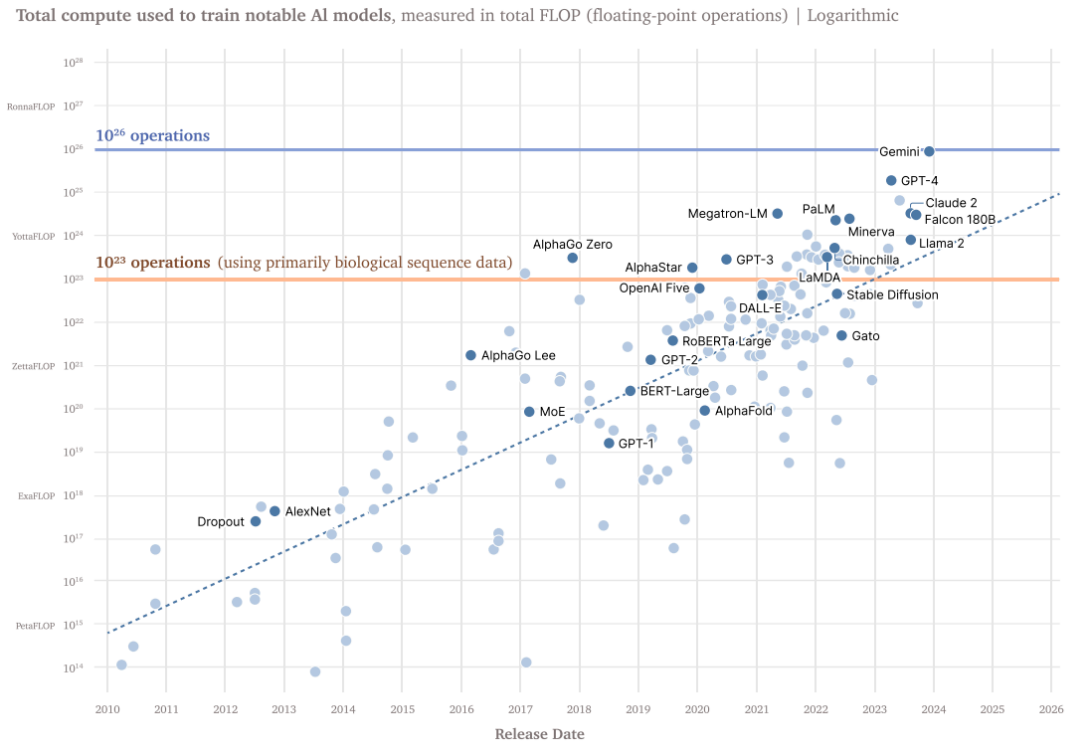

Figure 13: *Compute Thresholds as specified in the US executive order 14110 (Sastry et al., 2024).*

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.

3.3 On-Chip Controls

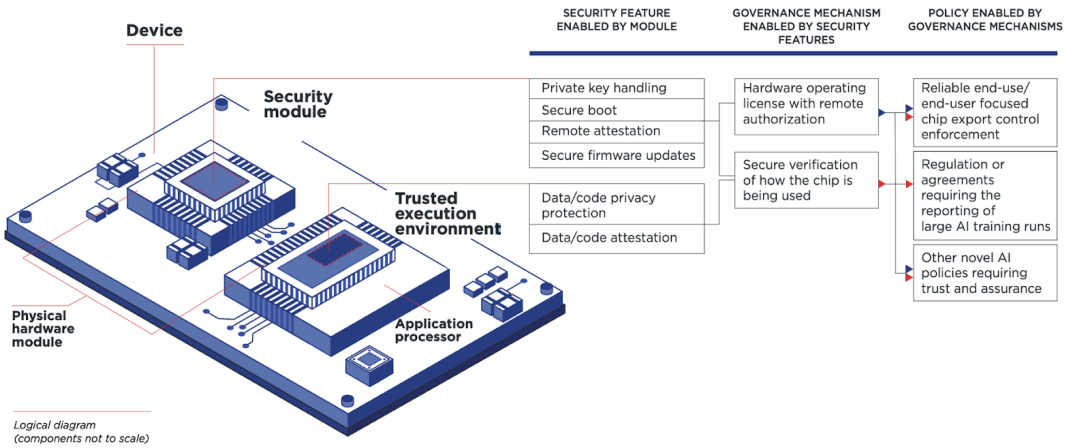

Figure 14: *Current AI chips already have some components of this architecture, but not all. These gaps likely could be closed with moderate development effort as extensions of functionality already in place (Aarne et al., 2024).*

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. It is worth noting that on-chip controls are highly speculative.

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).

3.4 Limitations

The trend over the last decade has involved more compute, but this will not last forever. We spoke at length about scaling laws in previous chapters. Research suggests scaling based returns to AI capabilities are still possible through 2030 (Sevilla et al., 2024). Algorithmic improvements also 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, reasoning models (LRMs) and inference-time scaling (e.g. OpenAI o3, Claude 4, 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).

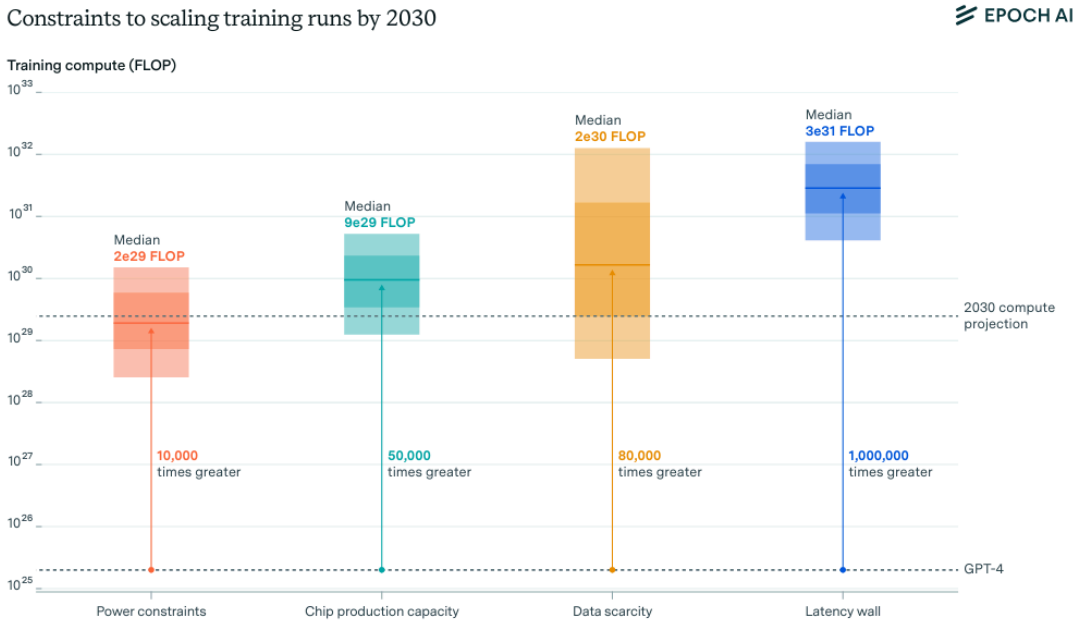

Figure 15: *Estimates of the scale constraints imposed by the most important bottlenecks to scale. Each estimate is based on historical projections. The dark shaded box corresponds to an interquartile range and light shaded region to an 80 percent confidence interval. The four boxes showcase four constraints that might slow down growth in the future: power, chips (compute), data and latency (Sevilla et al., 2024).*

Smaller more specialized models can 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).

Compute governance can help manage AI risks, but overly restrictive controls can accelerate power concentration. 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 training methods are being researched, but have not really caught up to centralized methods (Douillard et al., 2023 ; Jaghouar et al., 2024).[1] 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).

Compute monitoring and compute thresholds should primarily operate as an initial screening mechanism. These approaches should be used mainly 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.

Technical governance measures need to coordinate with corporate, national and international initiatives. We focused on compute governance as our primary technical example, though coordination challenges apply equally to data governance, model governance, and other technical measures. Each approach faces the same fundamental limitation: technical measures alone cannot address systemic risks that emerge from competitive dynamics and global deployment. This is why technical measures must be embedded within corporate, national and international governance frameworks that align incentives with coordinated safety standards. Before we talk about those however, we need to explore broader concepts like decision making under uncertainty, game theoretic collective action problems and other systemic forces that shape the governance landscape. We will talk about this in the next section.

[1] Example models trained using Decentralized methods include the INTELLECT-1 and INTELLECT-2 (Prime Intellect, 2025)

4. Systemic Challenges

4.1 Race dynamics

> *[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.*
>
> **John Schulman**
> Co-Founder of OpenAI

We already talked about race dynamics in the chapter on AI risks as amplifying factors for all risks. We mention them here again, because governance initiatives might have special leverage to be able to mitigate race dynamics.

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).

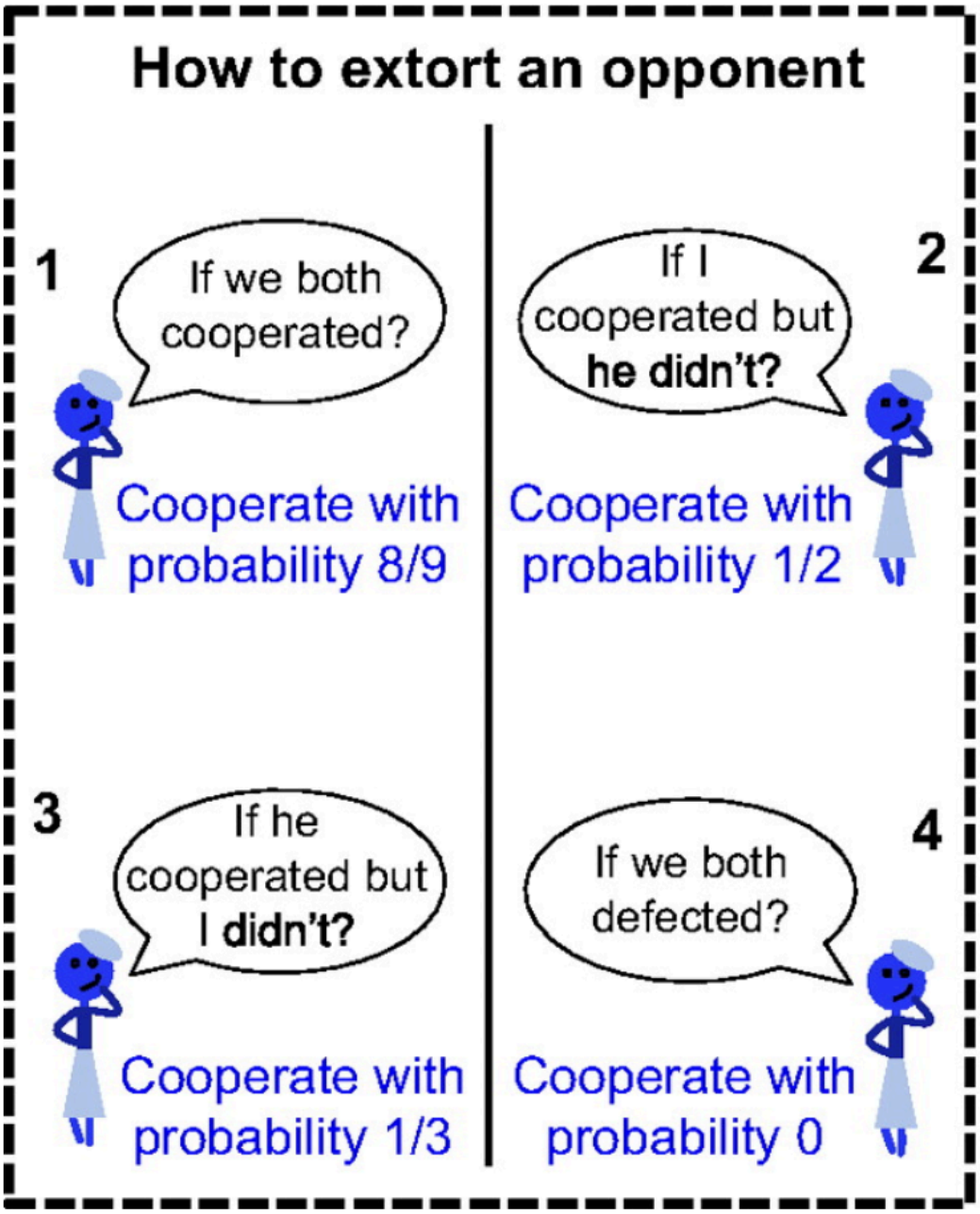

Figure 16: *How to extort your opponent, and what you stand to gain by extortion (Stewart & Plotkin, 2012).*

The AI race creates a 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).

Competitive pressure leads to safetywashing, cutting corners on testing, skipping external red-teaming, and rationalizing 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.

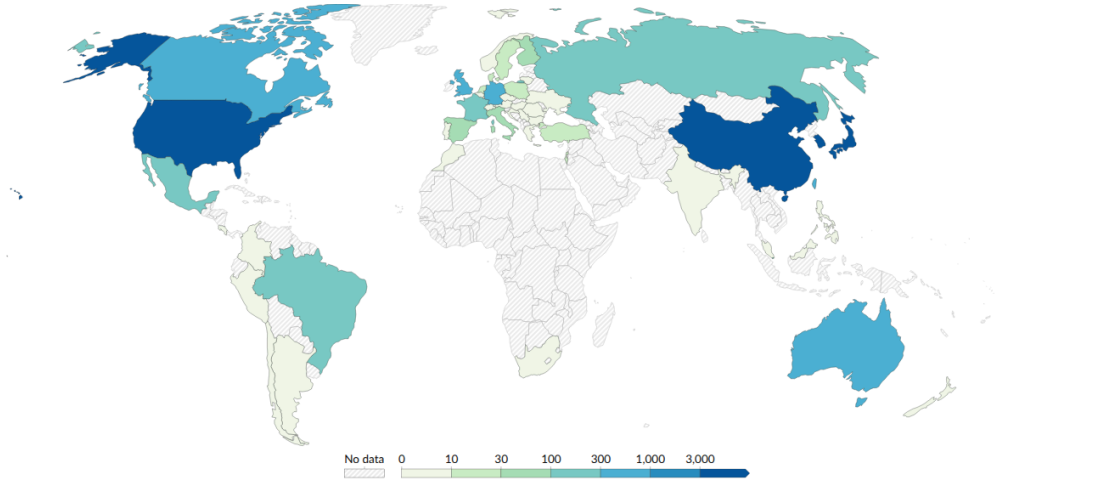

Figure 17: *Annual patent applications related to artificial intelligence, 2019. Patents submitted in the selected country's patent office (Giattino et al., 2023).* (interactive version on website)

National security concerns intensify race 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 (AP News, 2017). 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.

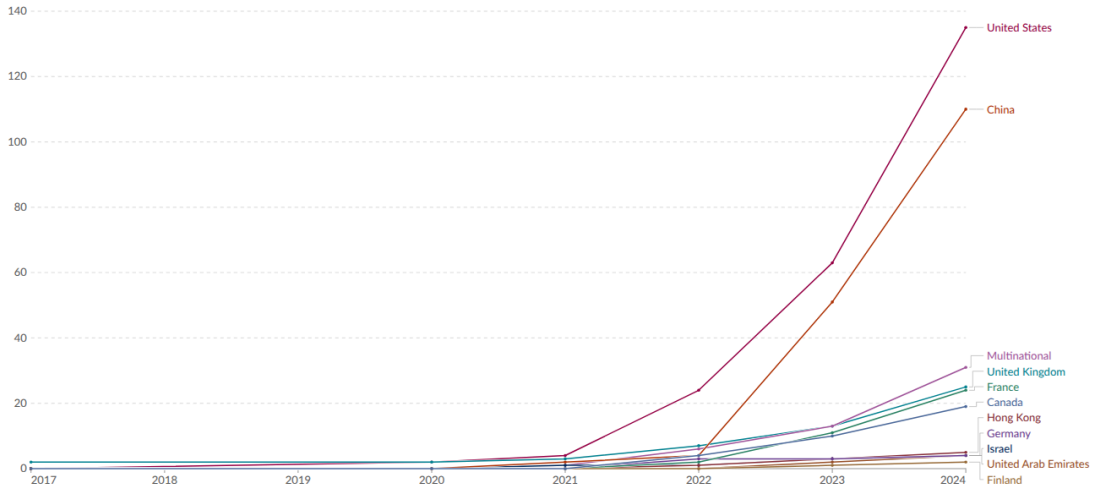

Figure 18: *Cumulative number of large-scale AI systems by country since 2017. Refers to the location of the primary organization with which the authors of large-scale AI systems are affiliated (Giattino et al., 2023).* (interactive version on website)

Race dynamics make collective action and 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.

AI governance needs innovative approaches to break out of race dynamics. Traditional arms control offers limited lessons, since AI development happens in private companies, not government labs. We need new approaches (Trajano & Ang, 2023 ; Barnett, 2025). Several ideas have been proposed. 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 (think about it like 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).

4.2 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. Let's think about this scenario - 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.

Balancing Proliferation mitigations and power concentration

OPTIONAL NOTE

Open-source releases face the same fundamental hardware constraints as proprietary development. While releasing model weights or training code might seem like it democratizes AI capabilities, the underlying compute requirements that we have discussed throughout the book remain unchanged. Anyone even with access to Llama's weights still needs hundreds if not millions of dollars in specialized hardware. Even fine-tuning frontier models for specific tasks requires significant GPU clusters that remain out of reach for most actors. This creates an interesting paradox: we can copy the "recipe" instantly, but we still can't afford the "kitchen."**The proliferation risk from open-source releases primarily comes from actors who already have substantial compute access - not from truly democratizing dangerous capabilities to resource-constrained adversaries.** Individual threat actors, bioterrorism or other catastrophic misuse scenarios would still need multi million dollar compute infrastructure to run frontier models capable of such harms. This hardware bottleneck means that the most concerning dual-use capabilities remain concentrated in the hands of major corporations and governments who control massive GPU clusters. While this concentration may provide some near-term safety benefits by limiting access to dangerous capabilities, it simultaneously accelerates concerning power dynamics where only a handful of entities can access the most capable AI systems. Until breakthroughs in model distillation, new architectures, or dramatically cheaper hardware make local hosting feasible, we face a fundamental trade-off between democratized access and concentrated control.

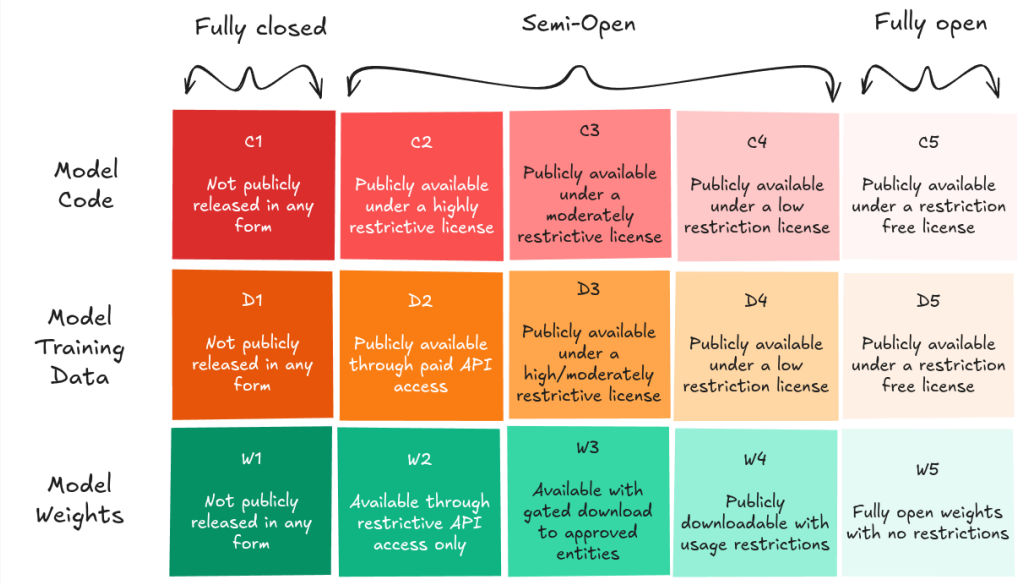

Figure 19: *A proposed gradient of access focusing on both model code and training data (Eiras et al., 2024) to carefully mitigate both proliferation and power concentration risks. We can see combinations of levels of access e.g. DeepSeek-V3 might roughly be considered C5-D1 (DeepSeek, 2025).*

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).

AI proliferation poses unique challenges - 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.

Verifying that someone if not developing harmful AI capabilities is extremely 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). This would require a combination of many different technical, and national measures.

Method	Evasion technique
Energy monitoring	Masking datacenter energy use, placement of datacenter in power plant
Financial intelligence	Use of shell corporations, other financial reporting evasion techniques
Remote sensing	Concealing datacenters underground, concealed cooling (e.g. pumping heat into a large body of water)
Whistleblowers	Secrecy measures, minimizing organisation size
Customs data analysis	Local manufacture of chips, use of older chips

Figure 20: *Table of evasion techniques to avoid verification methods under current national technical means. (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).

4.3 Uncertainty

> *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.*
>
> **Greg Brockman**
> Co-Founder and Former CTO of OpenAI

Expert predictions consistently fail to capture AI's actual trajectory. If you read media coverage of ChatGPT — which called it 'breathtaking', 'dazzling', 'astounding' — you'd get the sense that large language models (LLMs) took the world completely by surprise. Is that impression accurate? Actually, yes. (Cotra, 2023) 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 technical or economic 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.

4.4 Accountability

> *[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.*

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

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).

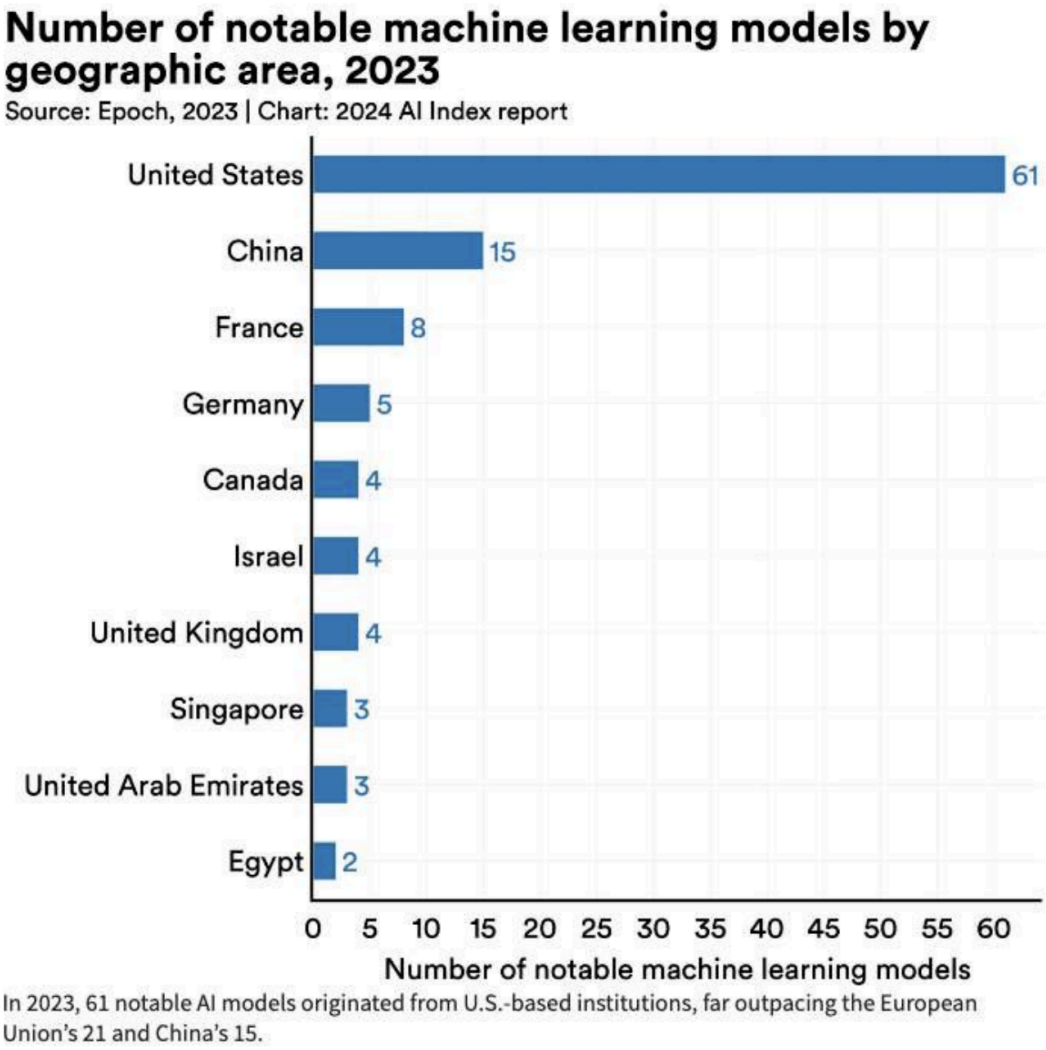

Figure 21: *In 2023, most of the notable AI models originated from U.S. institutions (Stanford, 2024).*

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).

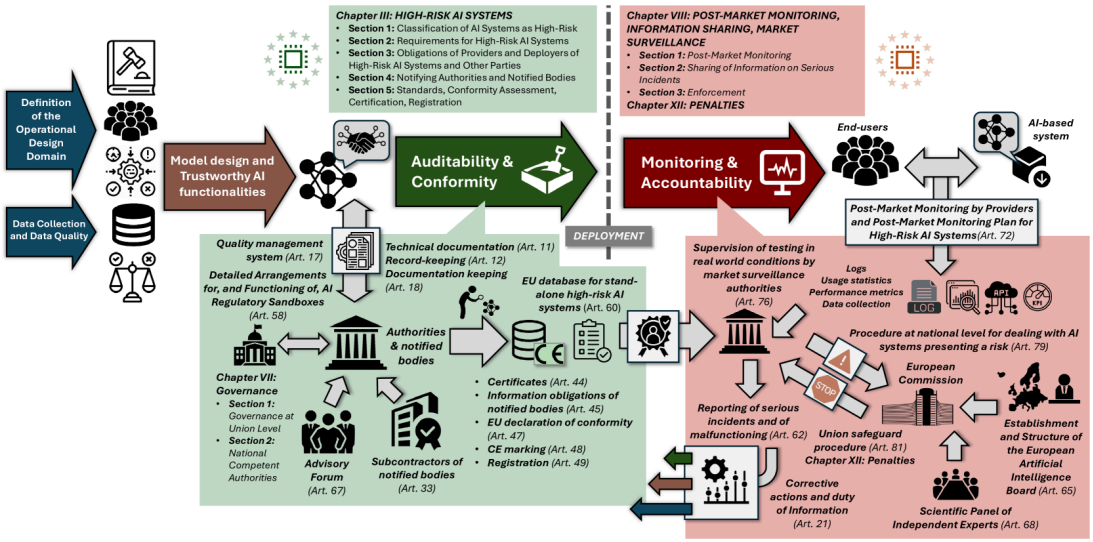

Figure 22: *Diagram showing the path from auditability to auditability (ex-ante) to accountability (post-hoc) (Herrera-Poyatos et al., 2025)*

4.5 Power and Wealth Concentration

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).

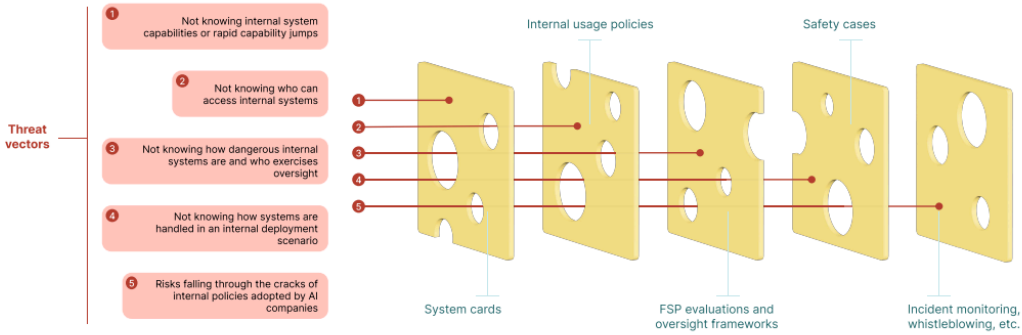

Figure 23: *'Swiss cheese model'model representing one recommended defense-in-depth strategy against the risk of undetected and unconstrained power accumulation. Threat vectors are in red (Stix et al., 2025)*

Democratic governance faces existential challenges. When information itself is controlled by private entities, traditional democratic institutions struggle to remain effective (Kreps & Kriner, 2023). Some empirical evidence 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).

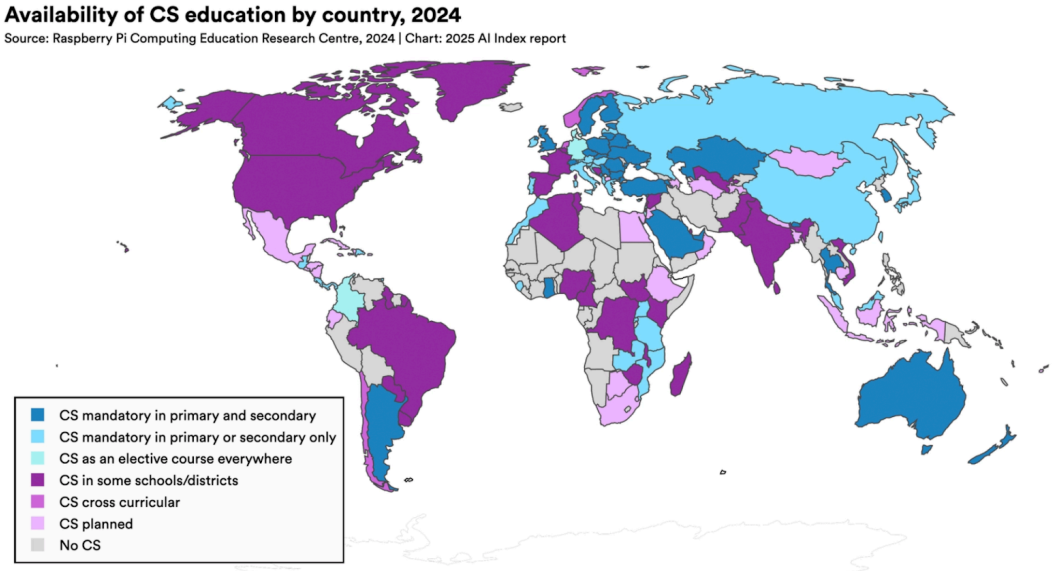

Figure 24: *In the U.S., the number of graduates with bachelor's degrees in computing has increased 22 percent over the last 10 years. Yet access remains limited in many African countries due to basic infrastructure gaps like electricity (Stanford HAI, 2025).*

5. 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 25: *The three levels of AI governance.*

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 26: *How the levels interact and reinforce.*

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.

5.1 Corporate Governance

> *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.*
>
> **Elon Musk**
> Founder/Co-Founder of OpenAI, Neuralink, SpaceX, xAI, PayPal, CEO of Tesla, CTO of X/Twitter

> *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.*
>
> **Dario Amodei**
> Co-Founder/CEO of Anthropic, ex-president of research at OpenAI

In this section we'll look at 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.

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 corporate 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.

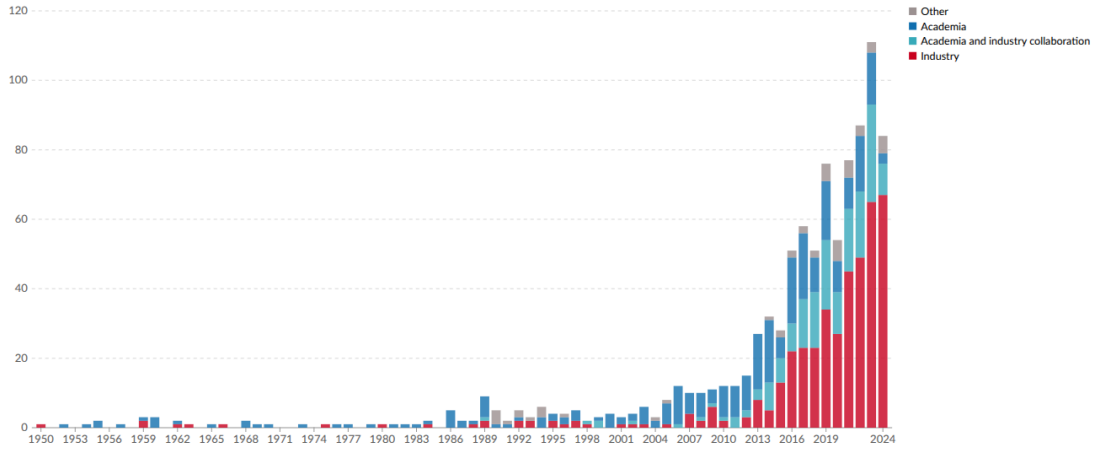

Figure 27: *Affiliation of research teams building notable AI systems, by year of publication. Describes the sector where the authors of a notable AI system have their primary affiliations (Giattino et al., 2023).* (interactive version on website)

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).

> *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.*
>
> **International AI Safety Report** (Bengio et al. 2025)

Internally deployed systems also need governance safeguards. Just because a model is not deployed publicly should not mean the corporate governance safeguards do not apply. We have seen in previous chapters that automating AI RnD is one of the core goals of several AI companies, this combined with proliferation safeguards and public release mitigations means that we can see many models that are heavily used internally but not available to the public. 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.

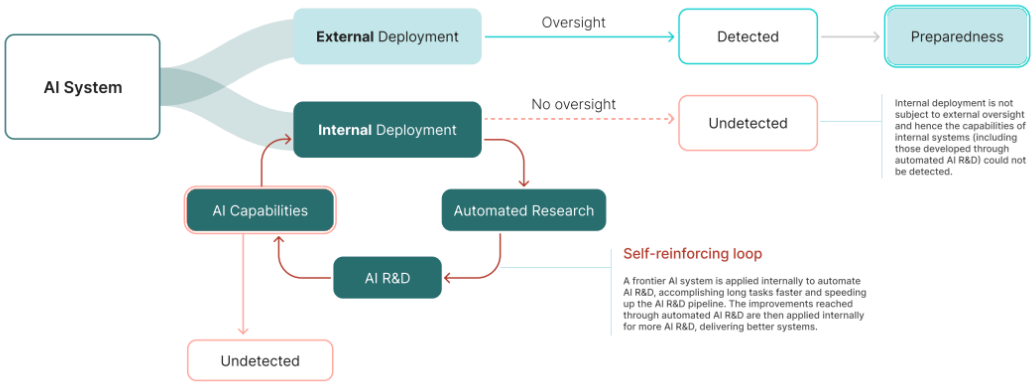

Figure 28: *The figure illustrates a self-reinforcing loop in which AI systems progressively automate AI research, leading to increasingly capable AI that further accelerates its own development (Stix, 2025).*

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 example, 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.

Corporate AI governance needs a combination of roles - board level oversight, AI risk executives, and technical safety teams. 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).

OpenAIs Corporate Restructuring

OPTIONAL NOTE

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).

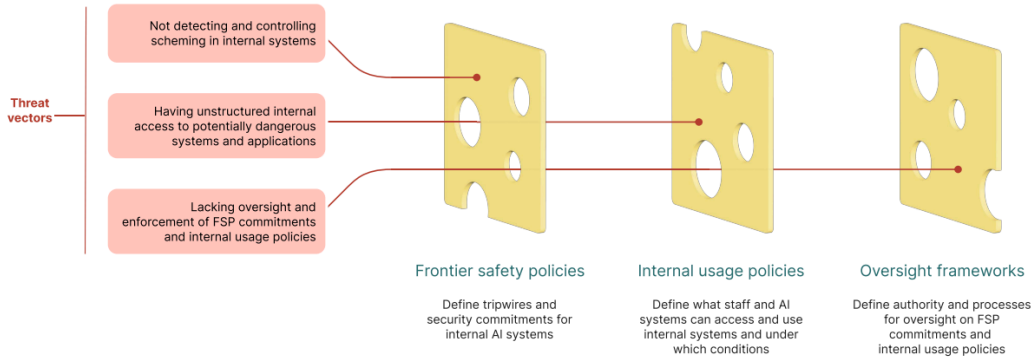

Figure 29: *Swiss cheese model representing our recommended defense-in-depth strategy against the risk of loss of control via internally deployed misaligned AI. Threat vectors are in red (Stix et al., 2025).*

5.1.1 Frontier Safety Frameworks

Frontier Safety Frameworks (FSFs) are one example of corporate AI governance. FSFs are 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. The concept of a FSF was first introduced in 2023. They 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). They go under different names, for example OpenAI calls their FSF the preparedness framework, and Anthropic calls them responsible scaling policies (RSPs). They are very similar in principle.

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).

A multi-layered internal auditing and governance approach helps operationalize safety frameworks in practice. When actually implementing the safety frameworks, organizations should ensure risks are identified and managed at multiple levels, reducing the chances of dangerous oversights. For example, 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). One approach 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.

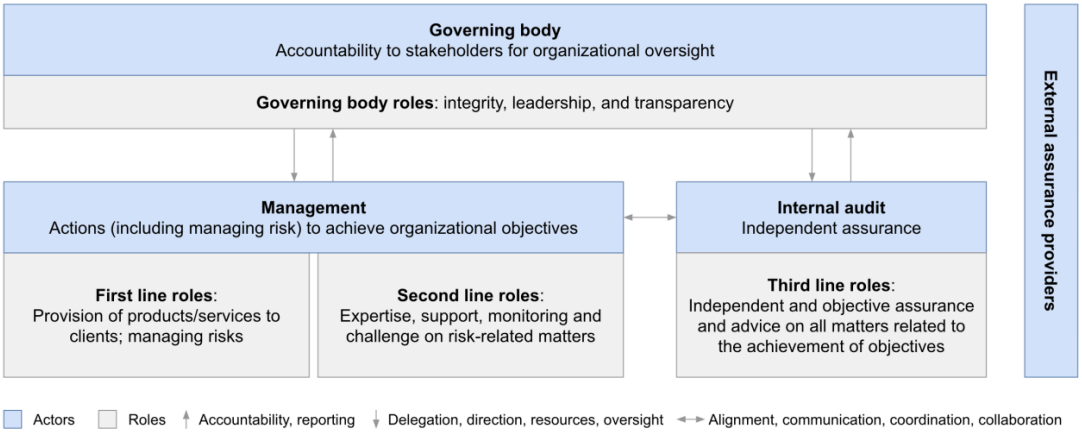

Figure 30: *The 3LoD model as described above (Schuett, 2023).*

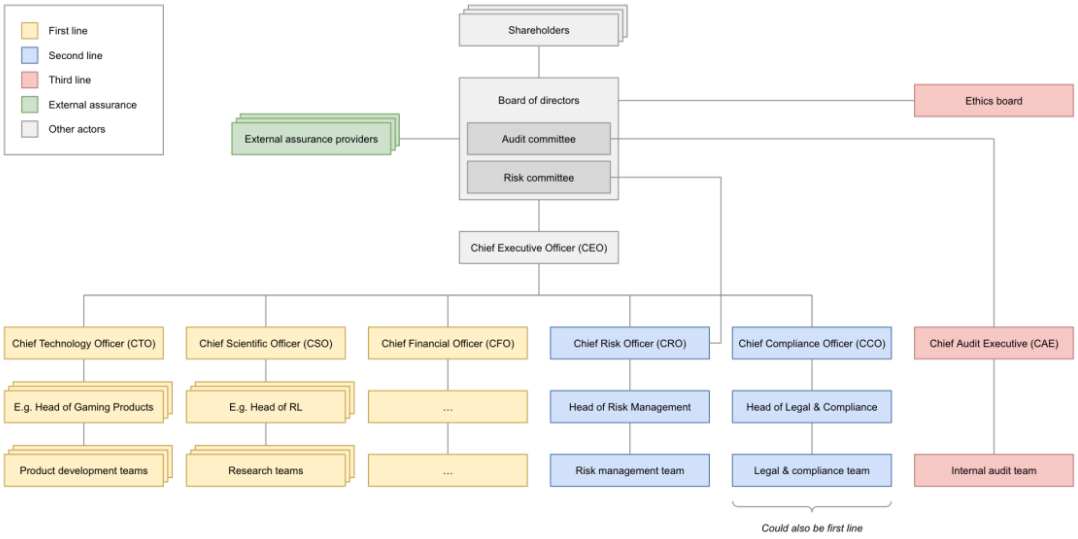

Figure 31: *Sample org chart of an AI company with equivalent responsibilities for each of the three lines (Schuett, 2023).*

FSFs need to account 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.

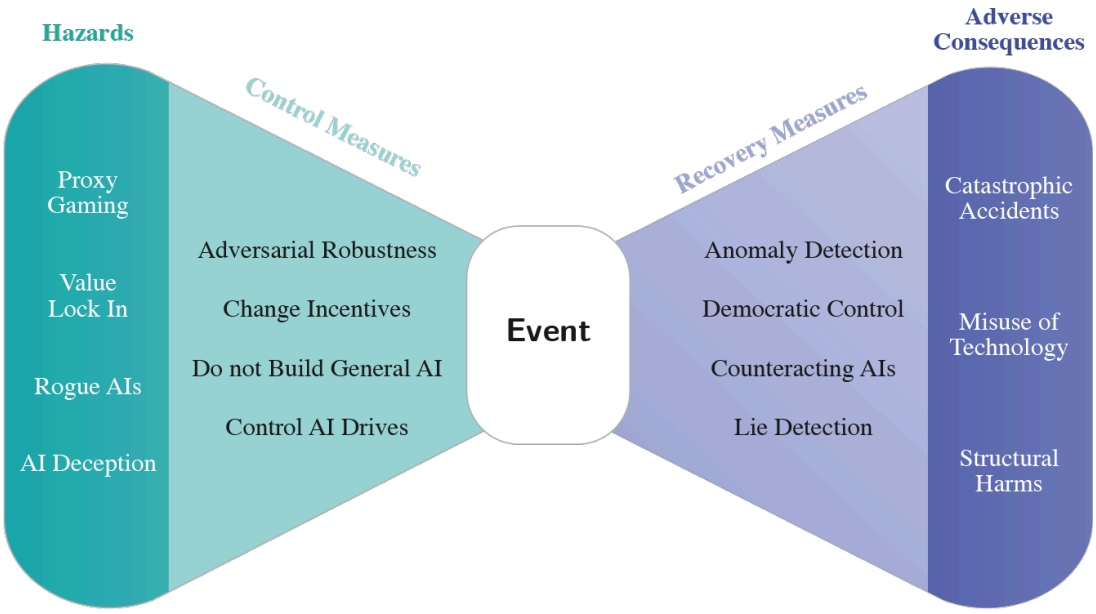

Figure 32: *Example of Bow-tie analysis technique (Hendrycks, 2024)*

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.

5.1.1.1 Limitations

> *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.*
>
> **Demis Hassabis**
> CEO and Co-Founder of DeepMind, Nobel Prize Laureate in Chemistry

FSFs represent a corporate self-regulation mechanism which represents progress but it might be insufficient. 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). 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).

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.

Systemic risks and collective action problems cannot be mitigated by corporate self-regulation of a single company. No one corporation can be trusted to serve the public interest alone. Corporate governance frameworks like FSFs show how companies can coordinate around shared safety standards. However, voluntary corporate coordination faces systematic pressures from market competition and regulatory arbitrage. When safety competes with speed or market share, even well-intentioned companies may defect from coordination agreements. This is why corporate governance requires the democratic legitimacy and enforcement power that only national governance can provide.

5.2 National Governance

> *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.*
>
> **Zhang Jun**
> China's UN Ambassador

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 33: *The state's role in governing artificial intelligence: development, control, and promotion through national strategies (Papyshev et al., 2023).*

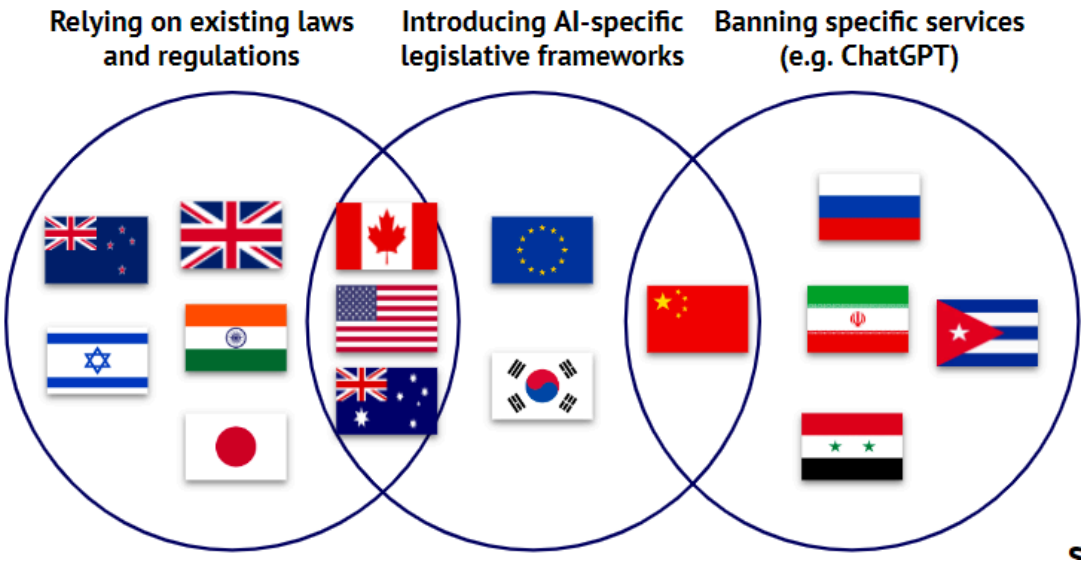

Figure 34: *(State of AI Report, 2023)*

National governance provides the enforcement mechanisms and democratic legitimacy necessary to make corporate coordination stable and binding. However, AI risks transcend national borders, and regulatory arbitrage allows development to shift to jurisdictions with weaker standards. No single nation can effectively govern global AI systems alone. This fundamental limitation drives the need for international coordination mechanisms.

5.3 International Governance

> *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.*
>
> **António Guterres**
> UN Secretary-General

> *[...] 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.*
>
> **Kamala Harris**
> Former US Vice President

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).

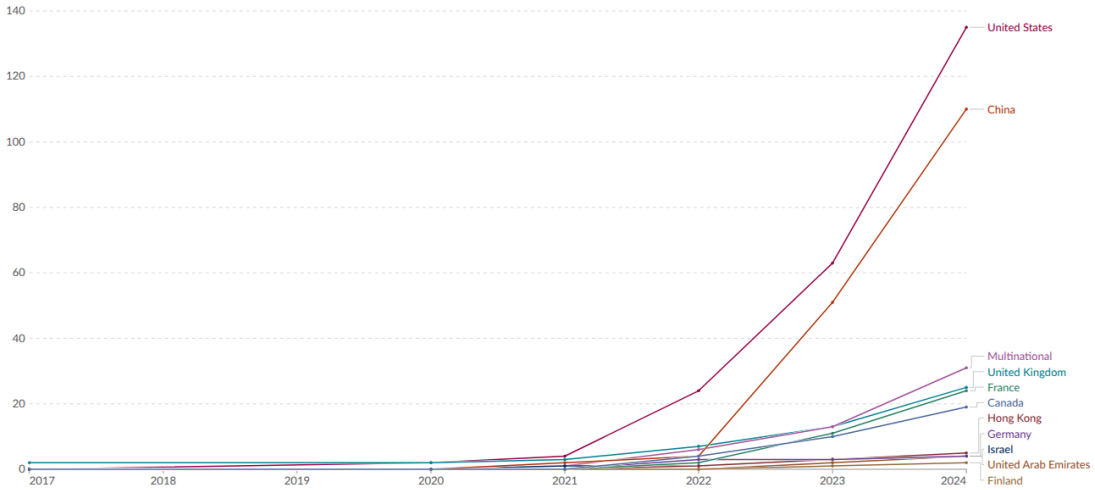

Figure 35: *Cumulative number of large-scale AI systems by country since 2017. Refers to the location of the primary organization with which the authors of a large-scale AI systems are affiliated (Giattino et al., 2023).* (interactive version on website)

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.

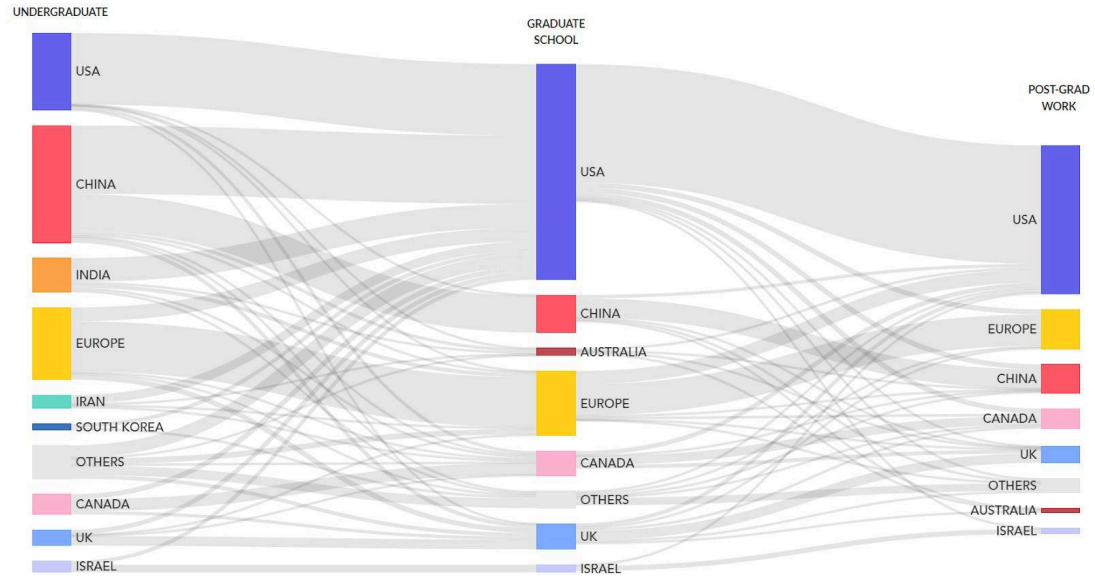

Figure 36: *What are the career paths of top-tier AI researchers? (MacroPolo)*

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).

> *[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.*
>
> **Ursula von der Leyen**
> Head of EU Executive Branch

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.

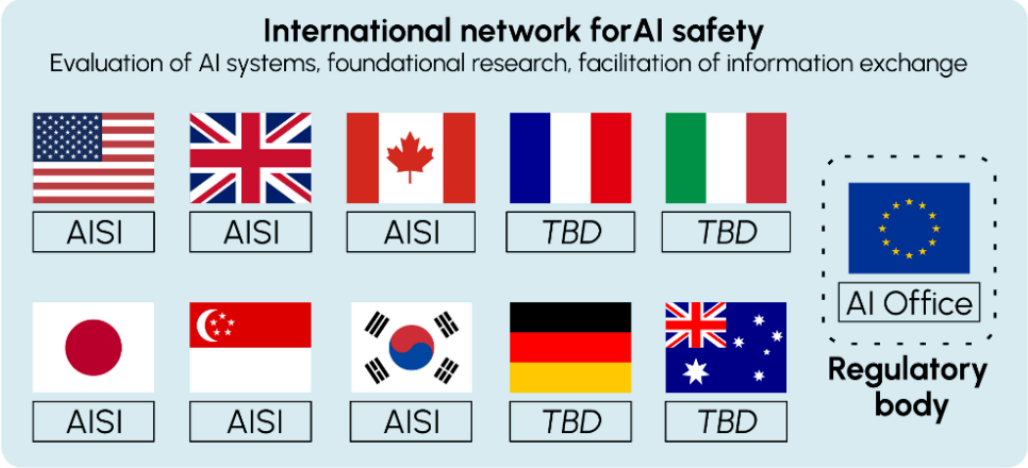

Figure 37: *These countries are part of the international network for AI safety, with their respective national bodies dedicated to AI safety (Variengien & Martinet, 2024).*

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 38: *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)*

Existing International Mechanisms (2025)

OPTIONAL NOTE

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.

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).

5.3.0.1 Policy Options

> *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.*
>
> **Demis Hassabis**
> Co-Founder and CEO of DeepMind

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 39: *An overview table of governance functions and their purpose.*

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).

6. Implementation

6.1 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).

6.2 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).

6.3 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).

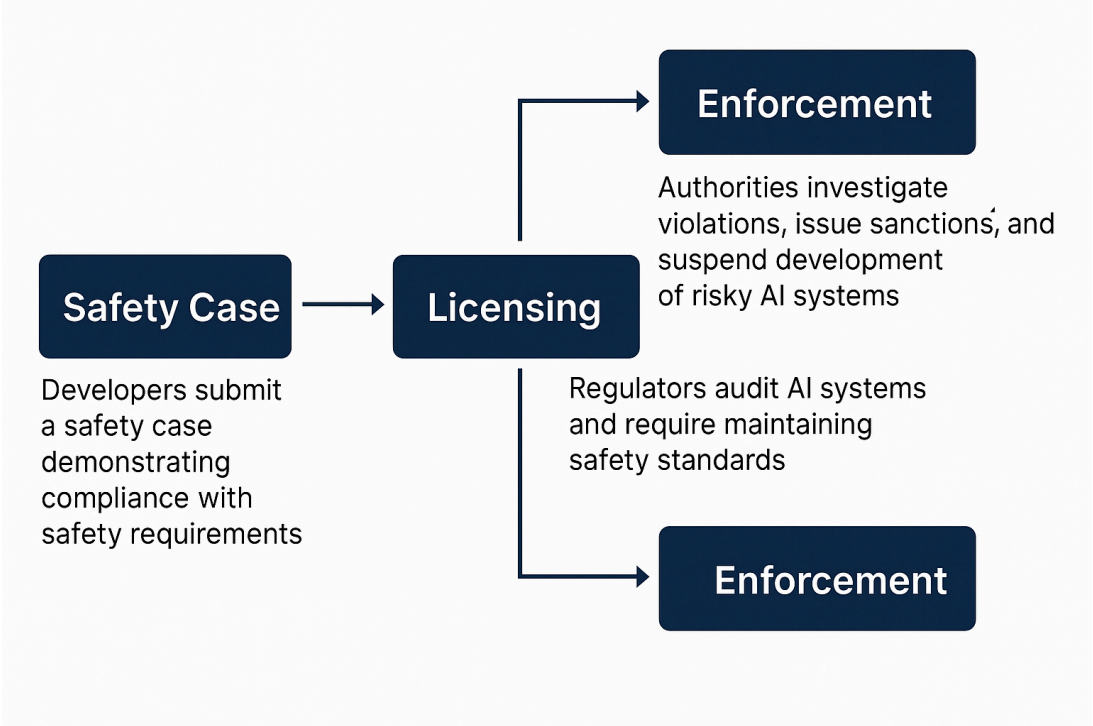

Figure 40: *The flow from safety cases to enforcement.*

6.4 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.

7. 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.

8. 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.

9. 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).

9.1 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).

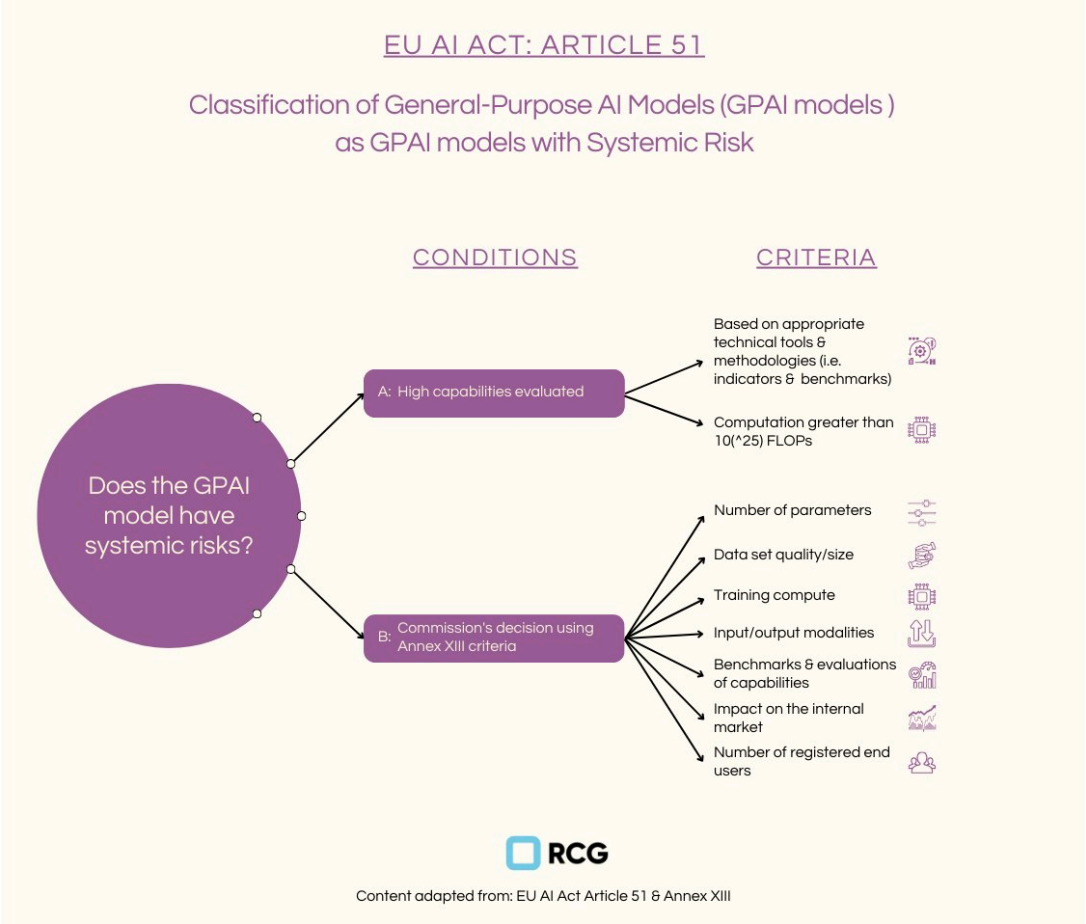

Figure 41: *The EU AI Act: Classification of general-purpose AI models with systemic risks (Observatorio de Riesgos Catastróficos Globales)*

9.2 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).

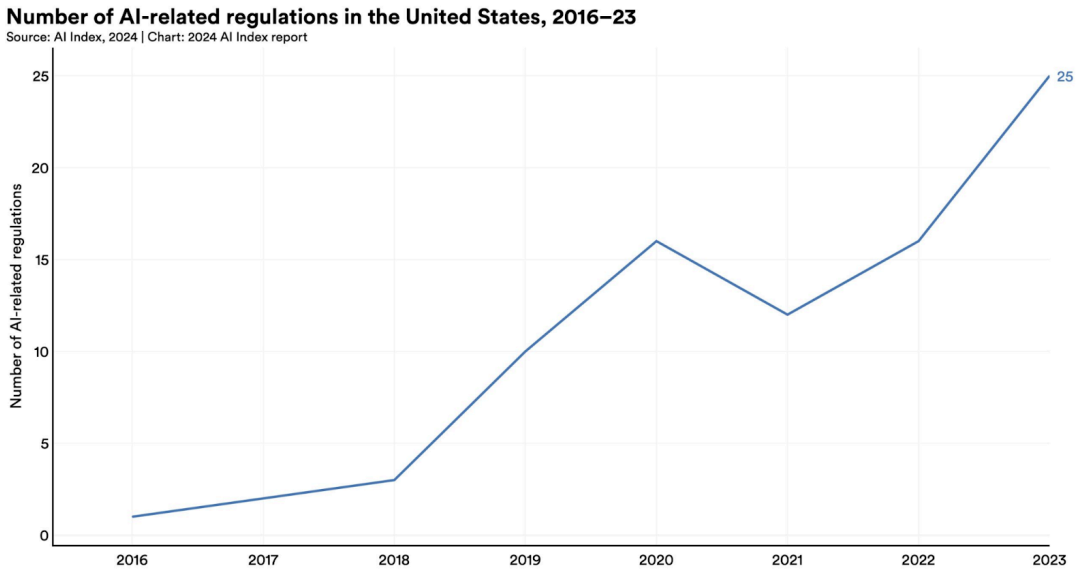

Figure 42: *Number of AI-related regulations in the United States, 2016-2023 (Stanford HAI, 2024)*

9.3 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).

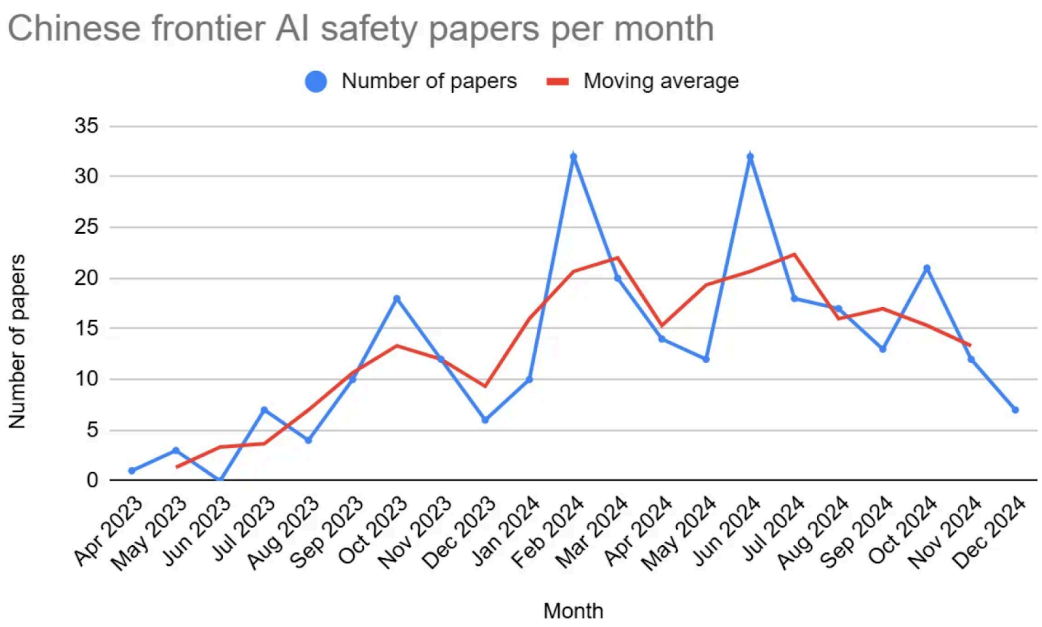

Figure 43: *In 2024, Chinese institutions significantly increased publication of frontier AI safety papers compared to 2023, from approximately seven papers per month in 2023 to 18 per month in 2024. (AI Safety in China, 2025)*

Acknowledgements

We would like to express our gratitude to Léo Karoubi and Ines Belhadj for their valuable feedback, discussions, and contributions to this work.