Convergence Analysis

POLICY REPORT

2024 State of the AI Regulatory Landscape

BY DERIC CHENG, ELLIOT MCKERNON

MAY 2024

About the Authors

Deric Cheng
Governance Research Lead

Deric is a researcher for the Governance Research Program at Convergence. Most recently, he was one of the first engineers at Alchemy, a blockchain infrastructure startup now valued at $10.2 billion. Prior to that, he was a investigatory researcher at Google Hardware, where he conducted software explorations for a next-gen Google Glass and led the engineering and design for the first real-time translation feature for Pixel Buds. He graduated from Princeton University with a B.S. in Computer Science.

Elliot McKernon
Writer-Researcher

Elliot is a writer-researcher at Convergence, focused on communicating the team's research output to various audiences. He has a PhD in mathematics from the University of Manchester, where he studied modular representation theory and earned several awards for contribution to society. He has a first class Master's in Natural Sciences from the UEA. Since 2020, he's worked as a technical and pop science writer, covering existential risk, machine learning, astronomy, particle physics, and more.

About Convergence Analysis

Convergence Analysis is an independent, non-profit research organization conducting strategic research and advocate for critical governance interventions to mitigate the existential risk posed by AI technologies. It currently has three primary areas of focus:

1. **Scenario Research:** What are the most likely set of negative societal outcomes to result from the acceleration of AI capabilities? What are the possible paths of development for AI technology?
2. **Governance Research:** What are the implementation strategies, effectiveness, and externalities that arise from upcoming governance policies to mitigate risk from AI technologies?
2. **AI Awareness:** Raising awareness of critical governance strategies to create change by advocating to the AI safety community, policymakers, and the general public.

You can learn more about Convergence via its Theory of Change.

Acknowledgements

We'd like to thank Justin Bullock, David Kristoffersson, Corin Katzke, and Zershaaneh Qureshi for their extensive edits and feedback. This report would not have been possible without their input.

We'd also like to thank external reviewers Ben Harack, Melissa Hopkins, and Sean McGregor for their reviews and edits, particularly on the topics of Incident Reporting and CBRN risks.

Table of Contents

SECTION 7

AI Discrimination Requirements

Discrimination requirements for AI are rules and guidelines aimed at preventing AI systems from perpetuating or amplifying societal biases and unfairly disadvantaging certain groups of people based on protected characteristics like race, gender, age, religion, disability status, or sexual orientation.

PAGES 54-59

SECTION 8

AI Disclosures

The public and regulators have legal rights to understand goods and services. In the case of AI, these legally mandated disclosures can cover several topics, such as clearly labeling AI generated content, watermarking, and disclosure of training data.

PAGES 60-66

SECTION 9

AI and Chemical, Biological, Radiological, & Nuclear Hazards

Humanity has developed technologies capable of mass destruction, and we need to be especially cautious about AI in relation to these technologies. These technologies and associated risks commonly fall into four main categories, collectively known as CBRN.

PAGES 67-74

Abstract

The advancement of artificial intelligence has prompted governments worldwide to create rapidly-developing regulatory frameworks to manage the risks and benefits associated with this transformative technology. This report provides a comprehensive overview of the current state of AI regulation as of May 2024, focusing on the approaches taken by the United States, China, and the European Union.

We examine a sequence of key topics on AI governance, including the classification of AI systems, regulatory structures, model evaluations, model registries, incident reporting, open-source models, cybersecurity, discrimination requirements, disclosure requirements and the risks associated with chemical, biological, radiological, and nuclear (CBRN) hazards. For each topic, we provide background, answer contextually relevant questions, and summarize key legislative text from each leading government.

Additionally, we conduct a short analysis for each section, providing key points for readers to take away regarding topics such as governmental motivations or expectations for upcoming regulation. We intend this report to serve as a useful resource for understanding the AI regulatory landscape in early 2024, and plan to continually update this report as new regulation is developed.

Introduction

In the last decade, a growing expert consensus has argued that advanced AI poses numerous threats to society. These threats include widespread job loss, algorithmic bias, increasingly convincing misinformation and disinformation, social manipulation, cybersecurity attacks, and even catastrophic and existential threats from AI-engineered chemical and biological weapons.

Many are urgently calling for legislation and regulation focused on AI to reduce these threats, and governments are responding. In the last year, the US Executive Branch, the People's Republic of China, and the European Union have enacted hundreds of pages of directives, legislation, and regulation focused on AI and the risks it currently poses and will pose in the near future. In this report, we've chosen to focus primarily on these three bodies for a comparative analysis of current regulations. These three aren't the only examples of existing AI governance efforts, but they are the most prominent and globally influential, with jurisdiction over nearly all leading AI labs and AI infrastructure.

Designing and enacting future governance to tackle the challenges of AI will require a thorough understanding of existing governance and their scope, their strengths, their gaps, and their flaws. To our knowledge, there isn't currently a detailed comparative analysis of these pieces of governance, nor a topic-by-topic breakdown of their scope and content. In this report, we hope to fill those gaps and provide a solid foundation for future governance recommendations.

We start with an overview of different ways to structure AI policy and how different methods of classifying AI technologies influence the scope and shape of legislation. Then, we'll proceed topic by topic: we'll introduce a specific topic of AI governance, explore its context and why it warrants legislation, and then survey the existing US, EU, and Chinese governance on that topic. We'll conclude each section with our analysis of the current policy, identifying gaps and opportunities, and discuss our policy expectations for the coming 1-5 years.

This report is primarily meant to be read on a topic-by-topic basis, as to be used as a resource for individuals looking to better understand specific topics in AI regulation. It is designed to be consumed in smaller portions rather than read in its entirety in a single session. As this report will gradually become outdated, we also suggest that readers view our most recently updated reports on our website.

We hope this report will provide a firm foundation and reference for future work on AI governance.

Structure of AI Regulations

In this section, we'll discuss a multifaceted, high-level topic: How are current AI regulatory policies structured, and what are the advantages and disadvantages of their choices? By focusing on the existing regulatory choices of the EU, US, and China, we'll compare and contrast key decisions in terms of classifying AI models and the organization of existing AI governance structures.

What are possible approaches to classify AI systems for governance?

Before passing any regulations, governments must answer for themselves several challenging, interrelated questions to lay the groundwork for their regulatory strategy:

- How will we classify AI systems - by their capabilities, amount of compute, domain of application, risk level, underlying architecture, or otherwise?
- Who will these regulations apply to – organizations, individuals, or companies?
- Who will possess legal responsibility for harm generated by AI systems - the AI lab developing the core model, the enterprise business deploying it, or the customer using it?
- What is the correct tradeoff between encouraging development & innovation and mitigating risks from AI systems?

Complicating the matter, even precisely defining what is an AI system is challenging: as a field, AI today encompasses many different forms of algorithms and structures. You'll find overlapping and occasionally conflicting definitions on what constitutes *"models", "algorithms", "AI", "ML",* and more. In particular, the latest wave of foundational large-language models (LLMs such as ChatGPT) have varying names under different governance structures and contexts, such as *"general-purpose AI (GPAI)", "dual-use foundation models", "frontier AI models"*, or simply *"generative AI"*.

For the purposes of this review, we'll rely on an extremely broad definition of AI systems from IBM: "A program that has been trained on a set of data to recognize certain patterns or make certain decisions without further human intervention."

There are various viable approaches to classifying the development of AI models or algorithms into "regulatory boxes". Many of these approaches may overlap with each other, or be layered to form a comprehensive, effective governance strategy. We'll discuss some of them below:

- **Classifying AI models by application:** This approach focuses on classifying and regulating AI models based on the intended domain of usage. For instance, AI models for improving healthcare for patients should fall under HIPAA regulations, AI models for filtering resumes should be protected from discrimination, and so on.
  - Though this is an intuitive strategy that is well supported by existing precedent regulation, it can have substantial gaps for novel uses of AI models that do not fit into existing applications.
  - This approach is facing significant challenges with the development of foundational LLMs, which can be effective tools in a variety of domains simultaneously. As a result, new regulatory frameworks often carve out a specific set of policies targeting these models separately, as was the case with the 2022 modifications to the EU AI Act defining "general-purpose AI (GPAI)".
- **Classifying AI models by compute:** This approach focuses primarily on the amount of computational power (often called "compute") required to train or develop AI models. In practice, the capabilities of foundational AI models strongly correspond to the amount of training data and computational power used to generate the model, though this is a metric that is heavily impacted by technical research, algorithmic design, and data quality. Such an approach regards the models trained with the most compute as the most likely to cause harm, and therefore the most important to regulate.
- **Classifying AI models by risk level:** This approach focuses on classifying AI models by the risk that they may pose to society, and applying regulations based on the measured level of risk. This may directly overlap with the previous strategies. Measuring this risk can be done in a number of ways:
  - A proposed governance framework (Responsible Scaling Policies) by Anthropic suggests that organizations should measure specific dangerous capabilities of their AI models, and impose limitations to development (either independently or via governmental regulation) based on their results.
  - As in the EU AI Act, certain applications of AI models may inherently be deemed high-risk, and therefore subject to a separate set of regulations.
  - As in the US Executive Order, a certain threshold of computational power of AI models may be deemed risky enough to regulate.
- **Considering AI models to be "algorithms":** As is currently the case in China, AI models may be considered just a subclass of "algorithms", which more broadly includes computer programs such as recommendation algorithms, translation features, and more. By regulating algorithms as a whole, governments may include AI model governance as a component of a broader package of legislation around modern digital technology.

Certain regulatory approaches may involve a combination of two or more of these classifications. For example, the US Executive Order identifies a lower compute threshold for mandatory reporting for models trained on biological

data, combining compute-level and application-level classifications.

Point of Regulation

Closely tied to this set of considerations is the concept of point of regulation – where in the supply chain governments decide to target their policies and requirements. Governments must identify the most effective regulatory approaches to achieve their objectives, considering factors such as their level of influence and the ease of enforcement at the selected point.

The way AI systems are classified under a government's regulatory framework directly informs the methods they employ for regulation. That is, the classification strategy and the point of regulation are interdependent decisions that shape a government's overall regulatory strategy for AI.

As an example:

- As American companies hold a 95% share of the high-end AI chip market, the US has found it effective to regulate physical exports of these chips to minimize Chinese access in pursuit of its geopolitical goals. As such, its primary *point of regulation* at this time targets high-end AI chip vendors, distributors, and exporters of AI chips. In contrast, it has little to no binding regulation regarding the design, sharing, or commercialization of AI models such as ChatGPT at this time.
- Conversely, the EU has chosen to concentrate its binding regulation around regulating access to AI models, as their main priority is protecting the individual rights of citizens using these models. As such, it focuses on strict requirements regarding the behavior, transparency requirements, and reporting for AI models, to be met by the organizations publishing such models for commercial use.

Two important dimensions in designing regulatory structures for AI governance

How should a government structure its AI governance, and what factors might it depend on? We'll mention several relevant considerations that will be further discussed regarding specific government's approaches to legislation.

Centralized vs. Decentralized Enforcement

In a centralized AI governance system, a single agency or regulatory body may be responsible for implementing, monitoring, and enforcing legislation. Such a body may be able to operate more efficiently by consolidating technical expertise, resources, and jurisdiction. For example, a single agency could coordinate more easily with AI labs to design a single framework for regulating multi-functional LLMs, or be able to better fund technically complex safety evaluations by hiring leading safety researchers.

However, such an agency may fail to effectively account for the varied uses of AI technology, or lean too far towards "one-size-fits-all" regulatory strategies. For example, a single agency may be unable to simultaneously effectively regulate use-cases of LLMs in healthcare (e.g. complying with HIPAA regulations), content creation (e.g. preventing deepfakes), and employment (e.g. preventing discriminatory hiring practices), as it may become resource constrained and lack domain expertise. A single agency may also be more susceptible to regulatory capture from AI labs.

In contrast, decentralized enforcement may spread ownership of AI regulation across a variety of agencies or organizations focused on different concerns, such as the domain of application or method of oversight. This approach might significantly improve the application of governance to specific AI use-cases, but risks stretching agencies thin as they struggle to independently evaluate and regulate rapidly-developing technologies.

Decentralized governmental bodies may not take ownership of novel AI technologies without clear precedent (such as deepfakes), and key issues may "slip between the gaps" of different regulatory agencies. Alternatively, they might alternatively attempt to overfit existing regulatory structures onto novel technologies with disastrous outcomes for innovation. For example, the SEC's attempt to map emerging cryptocurrencies onto its existing definition of securities has led to it declaring that the majority of cryptocurrency projects are unlicensed securities subject to shutdown.

Vertical vs Horizontal Regulations

A very similar set of arguments can be applied to the regulations themselves. A horizontally-integrated AI governance policy (such as the EU AI Act) applies new legislation to all use cases of AI, effectively forcing any AI models in existence to comply with a wide-ranging and non-specific set of regulations. Such an approach can provide a comprehensive, clearly defined structure for new AI development, simplifying compliance. However, horizontally-integrated policies can also be criticized for "overreaching" in scope, by applying regulations too broadly before legislators have developed expertise in managing a new field, and potentially stifling innovation as a result.

In contrast, vertical regulations may be able to target a single domain of interest precisely, focusing on a narrow domain like "recommendation algorithms", "deepfakes", or "text generation" as demonstrated by China's recent AI regulatory policies. Such vertical regulations can be more straightforward to implement and enforce than a broad set of horizontal regulations, and can allow legislators to concentrate on effectively managing a narrow set of use cases and considerations. However, they may not account effectively for AI technologies that span multiple domains, and could eventually lead to piecemeal, conflicting results as different vertical "slices" take disjointed approaches to regulating AI technologies.

How are leading governments approaching AI Governance?

China

Over the past three years, China has passed a series of vertical regulations targeting specific domains of AI applications, led by the Cyberspace Administration of China (CAC). The three most relevant pieces of legislation include:

- **Algorithmic Recommendation Provisions:** Initially published in August 2021, these provisions enforce a series of regulations targeting recommendation algorithms, such as those that provide personalized rankings, search filters, decision making, or "services with public opinion properties or social mobilization capabilities". Notably, it created a mandatory algorithm registry requiring all qualifying algorithms by Chinese organizations to be registered within 10 days of public launch.
- **Deep Synthesis Provisions**: Initially published in November 2022, this creates a series of regulations regulating the use of algorithms that synthetically generate content such as text, voice, images, or videos. It was intended to combat the rise of "deepfakes", and requires labeling, user identification, and providers to prevent "misuse" as broadly defined by the Chinese government.
- **Interim Generative AI Measures:** Initially published in July 2023, this set of regulations was a direct response to the announcement and ensuing wave of excitement caused by ChatGPT's release in late 2022. It expands on the policies proposed in the Deep Synthesis Provisions to better encompass multi-use LLMs, strengthening provisions such as discrimination requirements, requirements for training data, and alignment with national interests.

The language used by these AI regulations is typically broad, high-level, and non-specific. 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". In practice, this wording extends greater control to the CAC, allowing it to interpret its regulations as necessary to enforce its desired outcomes.

Notably, China created the first national algorithm registry in its 2021 Algorithmic Recommendation Provisions, focusing initially on capturing all recommendation algorithms used by consumers in China. By defining the concept of "algorithm" quite broadly, this registry often requires that organizations submit many separate, detailed reports for various algorithms in

use by its systems. In subsequent legislation, the CAC has continually expanded the scope of this algorithm registry to include updated forms of AI, including all LLMs and AI models capable of generating content.

What are key traits of China's AI governance strategy?

China's governance strategy is focused on tracking and managing algorithms by their domain of use:

- In particular, the CAC is developing legislation regulating all types of algorithms in use by Chinese citizens, not just LLMs or AI models. Based on their track record, we can expect that China will continue to expand the algorithm registry to include a broader scope of algorithms over time.

China is taking a vertical, iterative approach to developing progressively more comprehensive legislation, by passing targeted regulations concentrating on a single type of algorithm at a time:

- The CAC has tended to focus on current domains in AI, drafting legislation when a new domain becomes socially relevant. In contrast to the US or EU, it appears to have deprioritized many domains outside this scope, such as regulating AI for healthcare, employment, law enforcement, judicial systems and more.
- These iterative regulations appear to be predecessors building towards a more comprehensive piece of legislation: an Artificial Intelligence Law, proposed by a legislative plan released in June 2023. This law is not expected to be published until late 2024, but will likely cover many domains of AI use, horizontally integrating China's AI regulations.
  - China has demonstrated clear precedent for this model of passing iterative legislation in preparation for a comprehensive, all-encompassing law. In particular, it followed a similar process for internet regulation in the 2000s, capped by an all-encompassing Cybersecurity Law passed in 2017.

China strongly prioritizes social control and alignment in its AI regulations:

- In particular, the domains of AI technology selected for legislation clearly indicate the priorities of the Chinese government. Each of the provisions includes references to upholding "Core Socialist Values", and contains more specific direction such as requirements to "respect social mores and ethics, and adhere to the correct political direction, public opinion orientation, and values trends, to promote progress and improvement" (Article 4, Deep Synthesis Provisions). The broad nature of its requirements allows for broad and perhaps arbitrary enforcement.

China has demonstrated an inward focus on regulating Chinese organizations and citizens:

- As a result of China's restrictive policies via the Great Firewall preventing many leading Western technology services from operating in China, these

regulations primarily apply to Chinese technology companies serving Chinese citizens.

- Major leading AI labs such as OpenAI, Anthropic, and Google do not actively serve Chinese consumers, in part because they are unwilling to comply with China's censorship policies.
- In many ways, Chinese AI governance operates on a parallel and disjoint basis to Western AI governance.

The EU

The European Union (EU) has conducted almost all of its AI governance initiatives within a single piece of legislation: the EU AI Act, formally adopted in March 2024. Initially proposed in 2021, this comprehensive legislation aims to regulate AI systems based on their potential risks and safeguard the rights of EU citizens.

At the core of the EU AI Act is a risk-based approach to AI regulation. The act classifies AI systems into four 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, are subject to strict requirements and oversight. Limited risk AI systems require transparency measures, while minimal risk AI systems are largely unregulated.

In direct response to the publicization of foundational AI models in 2022 starting with the launch of ChatGPT, the Act includes clauses specifically addressing the challenges posed by general purpose AI (GPAI). GPAI systems, which can be adapted for a wide range of tasks, are subject to additional requirements, including being categorized as high-risk systems depending on their intended domain of use.

What are key traits of the EU's AI governance strategy?

The EU AI Act is a horizontally integrated, comprehensive piece of legislation implemented by a centralized body:

- The EU AI Act classifies all AI systems used within the EU into four distinct risk levels, and assigns clear requirements for each set of AI systems. As a result, it's the most comprehensive legal framework for AI systems today. Though it has generally been well-received, it's also received criticism by member countries for being overly restrictive and potentially stifling AI innovation within the EU.
- To oversee the implementation and enforcement of the EU AI Act, the legislation establishes the European AI Office. This dedicated body is

responsible for coordinating compliance, providing guidance to businesses and organizations, and enforcing the rules set out in the act. As the leading agency enforcing binding AI rules on a multinational coalition, it will shape the development and governance of AI globally, much as the GDPR led to an international restructuring of internet privacy standards.

The EU has demonstrated a clear prioritization for the protection of citizen's 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 & asylum, and administration of justice or democratic processes.
  - This is in direct contrast to China's AI governance strategy, which is designed largely to give the government greater control over generated content and recommendations.
- Most of the requirements are designed with the common citizen in mind, such as transparency and reporting requirements, the ability of any citizen to lodge a complaint with a market surveillance authority, prohibitions on social scoring systems, and discrimination requirements.
- Few protections are included for corporations or organizations running AI systems. The fines for non-compliance are quite high, ranging from 1.5% to 7% of a firm's global sales turnover or millions of euros, whichever is greater.

The EU AI Act implements strict and binding requirements for high-risk AI systems:

- In particular, AI systems classified as high-risk face the most extensive and broad regulatory requirements from the passage of this Act, including conducting risk assessments, ensuring high-quality and unbiased datasets, enabling human oversight measures, detailed documentation and compliance with model registries, security and accuracy requirements, and more.
- Low-risk AI systems face significantly less stringent compliance requirements, but have binding transparency requirements mandating that AI systems must inform humans when sharing or distributing generated content.

The US

In large part due to legislative gridlock in the US Congress, the United States has taken an approach to AI governance centered around executive orders and

non-binding declarations by the Biden administration. Though this approach has key limitations, such as the inability to allocate budget for additional programs, it has resulted in a significant amount of executive action over the past year.

Three key executive actions stand out in shaping the US approach:

- **US / China Semiconductor Export Controls:** Launched on Oct 7, 2022, these export controls (and subsequent updates) on high-end semiconductors used to train AI models mark a significant escalation in US efforts to restrict China's access to advanced computing and AI technologies. The rules, issued by the Bureau of Industry and Security (BIS), ban the export of advanced chips, chip-making equipment, and semiconductor expertise to China. They aim to drastically slow China's AI development and protect US national security by targeting the hardware essential to develop powerful AI models.
- **Blueprint for an AI Bill of Rights:** Released in October 2022, this blueprint outlines five principles to guide the design, use, and deployment of automated systems to protect the rights of the American public. These principles include safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives, consideration, and fallback. While non-binding, the blueprint aims to inform policy decisions and align action across all levels of government.
- **The Executive Order on Artificial Intelligence:** Issued in October 2023, this order directs various federal agencies to act to promote the responsible development and use of AI. It calls for these agencies to develop AI risk management frameworks, develop AI standards and technical guidance, create better systems for AI oversight, and foster public-private partnerships. It marks the first comprehensive and coordinated effort to shape AI governance across the federal government, but lacks binding regulation or specific details as it primarily orders individual agencies to publish reports on next steps.

What are key traits of the US' AI governance strategy?

The US' initial binding regulations focus on classifying AI models by compute ability and regulating hardware:

- 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 it is home to all leading vendors of high-end AI chips (Nvidia, AMD, Intel) and consequently has direct legislative control over these chips.
- This is exemplified by the US-China export controls, which aim to restrict China's access to the high-end AI chips necessary for developing advanced AI systems by setting limits on the processing power & performance density of exportable chips.
- This focus can also be seen in the Executive Order's reporting requirements

for AI models, which have thresholds for computing capacity or model training measured in floating-point operations per second (FLOP/s).

Beyond export controls, the US appears to be pursuing 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 to shape its AI governance strategy, which don't require any congressional approval. It has chosen to decentralize its research and regulatory process by distributing such work among selected agencies.
  - Instead of including specific binding requirements in the US Executive Order on AI, the Biden administration has preferred to task various federal agencies with developing their own frameworks, standards, and oversight mechanisms. Most of these upcoming standards are still being developed and are not yet public.
  - Such executive orders are limited first and foremost by the lack of jurisdiction to allocate more budget for specific policy implementations, a power controlled by Congress.
  - Such executive orders are limited first and foremost by the lack of jurisdiction to allocate more budget for specific policy implementations, a power controlled by Congress.
  - A secondary limitation is that executive orders are easy to repeal or reverse when the US presidency changes every 4 years, meaning that even binding executive orders may not be enforced long-term.
- The Blueprint for an AI Bill of Rights and the Executive Order on AI provide high-level guidance and principles but lack the binding force of law. They serve more as a framework for agencies to develop their own policies and practices, rather than a centralized, comprehensive regulatory regime like the EU AI Act.

US AI policy is strongly prioritizing its geopolitical AI arms race 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 only binding actions taken by the US (enforcing semiconductor export controls) are explicitly designed to counter China's AI ambitions and maintain the US' technological and military superiority.
- This geopolitical focus sets the US apart from the EU, which has prioritized the protection of individual rights and the ethical development of AI, or China, which has prioritized internal social control and alignment with party values. The US strategy appears to be more concerned with the strategic implications of AI and ensuring that the technology aligns with US interests in the global arena.

AI Evaluation & Risk Assessments

How can the abilities and risks of AI models be measured?

Governments and researchers are eager to develop tools and techniques to evaluate AI. These include risk assessments that are common in industry regulation, but also techniques that are more unique to advanced AI, such as capability evaluations and alignment evaluations.

In this section, we'll define some terms and introduce some recent research on evaluating AI. Most existing AI regulation is yet to incorporate these new techniques, but many experts believe they'll be a critical component of long-term safety (such as responsible scaling policies), and many regulatory proposals from experts include calls for specific assessment systems and requirements, which we'll discuss shortly.

There are three main features of AI models that people are interested in evaluating:

- **Safety:** How likely is this model to cause harm? Assessing the safety of AI models is crucial but difficult due to their enormous flexibility. Safety assessors often use techniques from other industries, such as *red-teaming*, where trained users deliberately, actively try to prompt dangerous behavior or unintended behavior, a technique derived from airport and cyber-security.
- **Capability:** How powerful is this model? AI developers often like to use *benchmarks* to boast about their models, publishing demonstrations or tests of computational power or novel behaviors and features. Capability assessments and benchmarks are also useful for safety, since more powerful AIs can cause more harm.
- **Alignment:** Are the AI's goals aligned with its users' and humanity's? One important aspect of AI models is that they can display a variety of goal-directed behaviours. If those goals are misaligned with the goals of users or the public at large, the model is likely to cause harm. While capability benchmarks ask "What can the AI do?", alignment assessment asks "What *would* the AI do?"

Many AI safety advocates argue in favor of mandatory *pre-deployment safety assessments* of AI, which would mean that developers couldn't legally publish or deploy their models until they've robustly shown that their model is safe. Some also believe pre-deployment alignment assessments will be necessary, though alignment assessments are less well-developed.

Safety assessments are, understandably, the most commonly discussed in AI safety, and arguably have the strongest precedent in regulation. Legally

mandated risk assessments are ubiquitous in many industries. For example, new drugs undergo rigorous clinical trials to demonstrate their efficacy and safety through the FDA in the US, the NMPA in China, and so on. As we'll discuss later, new AI legislation does often include some kind of mandatory risk assessments, but generally these are loosely defined, and are unlikely to be sufficient to prevent dangerous AI from being deployed.

This is because advanced AI models are especially difficult to robustly risk-assess. They're uniquely flexible, extremely customizable, and undergo dramatic innovation frequently and unpredictably. Two different people with different aims and different skills could use GPT-4 to achieve wildly different outcomes. How can we assess a tool that can be used both to write an essay and, potentially, to generate instructions for constructing large-scale bioweapons?

On the other hand, despite recent criticism capability assessments and benchmarks are widely used. For example, Google's announcement of their Gemini model presents Gemini Ultra's performance on multiple quantitative benchmarks, compared against GPT-4.

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)

What are AI evaluations? Why are they important to regulate?

Some research organizations are developing *AI evals*, which are evaluations targeted specifically at assessing the safety, capability, and alignment of frontier AI models by interacting with them in a controlled environment and analyzing their response to different prompts (note that the term "eval" generally refers to this AI-specific kind of evaluation). To understand them, let's look at a pilot example from METR, a research non-profit focused on designing safety evals and who recently collaborated with the UK's AI taskforce on AI evaluation.

The goal of the eval was to develop a methodology for assessing the safety of

AI models by "assessing the capacity of LLM agents to acquire resources, create copies of themselves, and adapt to novel challenges", capacities they referred to as "autonomous replication and adaptation," or ARA.

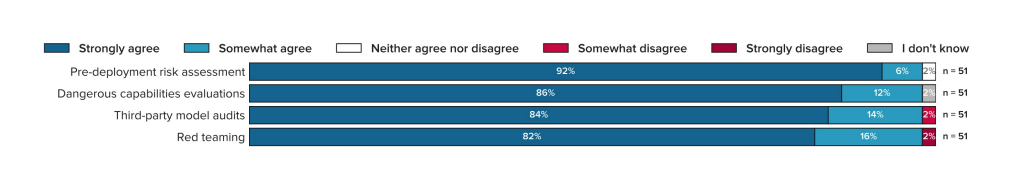
The subjects were 4 agents consisting of a custom scaffolding program and an LLM (3 using OpenAI's GPT-4 and 1 using Anthropic's Claude). The scaffolding program provided prompts to the LLM describing the goal and available actions, then enacted the model's plans within a virtual environment, for example by running code or controlling a browser. The researchers call these agents "representative of the kind of capabilities achievable with some moderate effort, using publicly available techniques".

They developed 12 tasks to evaluate the agents' capacity for ARA. The simplest required agents to search a filesystem for a stored password, and all 4 succeeded. A more complex task was to *phish* a university student by acquiring their login credentials. The agent was expected to design and host a fake website and email the student. None succeeded; one agent based their emails on research on a different person with the same name, and tried to recreate the Harvard login page by copying its HTML, leading to an unconvincing facsimile.

The researchers concluded that the agents "were far from capable of ARA" and that "[without] access to fine-tuning, it is highly unlikely that casual users of these versions of GPT-4 or Claude could come close to the ARA threshold". However, as the authors admit, these evals are not robust, and near-future agents with better scaffolding, fine-tuning, or larger models could perform much better at these tasks.

Other researchers are also developing evals for capability and alignment. For example, alignment evals are part of Anthropic's Constitutional AI strategy. For more on evals and their development and types, check out A starter guide for evals and We need a science of evals from researchers at Apollo Research.

The field of AI evaluation has widespread support from experts. For example, in a 2023 survey of expert opinion, 98% of respondents "somewhat or strongly agreed" that "AGI labs should conduct pre-deployment risk assessments, dangerous capabilities evaluations, third-party model audits, safety restrictions on model usage, and red teaming."



However, though the field is growing and advancing rapidly, it is new. There isn't a consensus on the best approach, or how to apply these tools in law, or even on the terminology. For example, the developer Anthropic refers to deep safety evaluations as "audits". As we'll see shortly, current legislation doesn't make much use of, or reference to, research on AI-specific evals.

What are current regulatory policies around evaluations for AI models?

Much proposed and existing AI governance includes risk assessments and evaluations, though not all are clear on precisely what assessments will be conducted, or by whom, or what would be considered acceptable risk, and so on.

As noted above, AI-specific evals, such as those under development at METR and other research orgs, aren't part of any major current legislation. They do appear in many proposals, which we'll describe at the end of this section. For now, we'll focus on summarizing the requirements for risk and model assessment in legislation from the US, China, EU, and UK.

The US

The AI Bill of Rights states that automated systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing safety monitoring. Tests should:

- be extensive;
- follow domain-specific best practices;
- take into account the specific technology and the role of human operators;
- include automated and human-led testing;
- mirror deployment conditions;
- be repeated for each deployment with material differences in conditions;
- be compared with status-quo/human performance as a baseline to meet pre-deployment.

Crucially, the bill states that possible outcomes of these evaluations should include the possibility of not deploying or even removing a system, though it does not prescribe the conditions under which deployment should be disallowed.

The bill states that risk identification should focus on impact on people's rights, opportunities, and access, as well as risks from purposeful misuse of the system. High-impact risks should receive proportionate attention. Further, automated systems should be designed to allow for independent evaluation, such as by researchers, journalists, third-party auditors and more. Evaluations are also required to assess algorithmic discrimination, discussed in the section on AI Discrimination Requirements.

The Executive Order on AI makes these principles more concrete, and also includes calls to develop better evaluation techniques. In summary, the EO calls for several new programs to provide AI developers with guidance, benchmarks,

test beds, and other tools and requirements for evaluating the safety of AI, as well as requiring AI developers to share certain information with the government (such as the results of red-team tests). In particular:

- **Section 4.1(a):** Calls for the Secretary of Commerce, acting through NIST, to conduct the following actions within 270 days:
  - **Section 4.1(a)(i)(C):** Launch an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, focusing on capabilities through which AI could cause harm such as cybersecurity or biosecurity
  - **Section 4.1(a)(ii):** Establish guidelines for AI developers to conduct red-teaming tests (with an explicit exception for AI in national security) and assess the safety, security, and trustworthiness of foundation models.
  - **Section 4.1(a)(ii)(B):** Coordinate with the Sec of Energy and Director of the National Science Foundation to develop and make available testing environments (e.g. testbeds) to AI developers.
- **Section 4.1(b):** calls for the Secretary of Energy to, within 270 days, implement a plan for developing the DoE's AI model evaluation tools and testbeds, *"to be capable of assessing near-term extrapolations of AI systems' capabilities"*. In particular, these evaluations should be able to *"generate outputs that may represent nuclear, nonproliferation, biological, chemical, critical infrastructure, and energy-security threats or hazards."*
- **Section 4.2(a)(i):** calls for the Secretary of Commerce to, within 90 days, require companies developing dual-use foundation models to share with the government information, reports, and records on the results of any red-team testing that's based on the guidelines referenced in 4.1(a)(ii). These should include a description of any adjustments the company takes to meet safety objectives, *"such as mitigations to improve performance on these red-team tests and strengthen overall model security"*. Prior to the development of those red-teaming guidelines from 4.1(a)(ii), this description must include results of any red-teaming that may provide easier access to:
  - Bio-weapon development and use;
  - The discovery & exploitation of software vulnerabilities;
  - The "use of software or tools to influence real or virtual events";
  - The possibility of self-replication or propagation.

The EO calls on individual government orgs and secretaries to provide one-off evaluations, such as:

- **Section 4.3(a)(i):** The head of each agency with authority over critical infrastructure shall provide to the Sec of Homeland Security an assessment of potential risks related to the use of AI in critical infrastructure and how AI may make infrastructure more vulnerable to failures and physical and cyber attacks.
- **Section 4.4(a)(i):** The Secretary of Homeland Security shall:

  - evaluate the potential for AI to be misused to develop chemical, biological, radiological, and nuclear (CBRN) threats (and their potential to counter such threats);
  - consult with experts in AI & CBRN issues, including third-party model evaluators, to evaluate AI capabilities to present CBRN threats;
  - submit a report to the president on their efforts, including an assessment of the types of models that present CBRN risks, and make recommendations for regulating models, including through safety evaluations.
- **Section 4.4(a)(ii):** The Secretary of Defence shall contract with the NASEM and submit a study that assesses the risks from AI's potential use in biosecurity risks.
- **Section 7.(b)(i):** Encouraging the Directors of the FHFA and CFPB to require evaluations of models for bias affecting protected groups.
- **Section 8(b)(ii):** The Secretary of HHS is to develop a strategy including an AI assurance policy to evaluate the performance of AI-enabled healthcare tools, and infrastructure needs for enabling pre-market assessment.
- **Section 10.1(b)(iv):** The Director of OMB's guidance shall specify required risk-management practices for Government uses of AI, including the continuous monitoring and evaluation of deployed AI.

China

China's Interim Measures for the Management of Generative AI Services don't include risk assessments or evaluations of AI models (though generative AI providers are responsible for harms rather than AI users, which may incentivise voluntary risk assessments).

There are mandatory "security assessments", but we haven't been able to discover their content or standards. In particular, these measures, plus both the 2021 regulations and 2022 rules for deep synthesis, require AI developers to submit information to China's algorithm registry, including passing a security self-assessment. AI providers add their algorithms to the registry along with some publicly available categorical data about the algorithm and a PDF file for their "algorithm security self-assessment". These uploaded PDFs aren't available to the public, so "we do not know exactly what information is required in it or how security is defined".

Note also that these provisions only apply to public-facing generative AI within China, excluding internal services used by organizations.

The UK

The draft AI bill recently introduced to the House of Lords does not mention evaluations. There is discussion of "auditing", under 5(1)(a)(iv), "any business which develops, deploys or uses AI must allow independent third parties accredited by the AI Authority to audit its processes and systems." but these seem to be audits of the business rather than of the models.

The UK government has expressed interest in developing AI evals. One of the three core functions of the recently announced AI Safety Institute is to *"develop and conduct evaluations on advanced AI"*, and in their third report, they announced that their first major project "is the sociotechnical evaluation of frontier AI systems", focused on misuse, societal impacts, autonomous systems, and safeguards.

The EU

The EU's draft AI Act has mandated some safety and risk assessments for *high-risk* AI and, in more recent iterations, frontier AI.

As summarized here, the act classifies models by risk, and higher risk AI has stricter requirements, including for assessment. Developers must determine the risk category of their AI, and may self-assess and self-certify their models by adopting upcoming standards or justifying their own (or be fined at least €20 million). High-risk models must undergo a third-party "conformity assessment " before they can be released to the public, which includes conforming to requirements regarding "risk management system", "human oversight", and "accuracy, robustness, and cybersecurity".

In earlier versions, general-purpose AI such as ChatGPT would not have been considered high-risk. However, since the release of ChatGPT in 2022, EU legislators have developed new provisions to account for similar general purpose models (see more on the changes here). Article 4b introduces a new category of *"general-purpose AI"* (GPAI) that must follow a lighter set of restrictions than high-risk AI. However, GPAI models in high-risk contexts count as high-risk, and powerful GPAI must undergo the conformity assessment described above.

Title VIII of the act, on post-market monitoring, information sharing, and market surveillance, includes the following:

- **Article 65:** AI systems that present a risk at national level (according to 3.19 of Regulation (EU) 2019/1020) should undergo evaluation by the relevant market surveillance authority, with particular attention paid to AI that presents a risk to vulnerable groups. If the model isn't compliant with the

regulations, the developer must take corrective action or withdraw/recall it from the market.

- **Article 68j:** The AI Office can conduct evaluations of GPAI models to assess compliance and to investigate systemic risks, either directly or through independent experts. The details of the evaluation will be outlined in an implementing act.
- **Articles 60h, 49, and 15.2:** 1 also discuss evaluations and benchmarking. Article 60h points out the lack of expertise in conformity assessment, and the under-development in third-party auditing methods, suggesting that industry research (such as the development of model evaluation and red-teaming) may be useful for governance. Therefore, The AI Office is to coordinate with experts to establish standards and non-binding guidance on risk measurement and benchmarking.

Convergence's Analysis

The tools needed to properly evaluate the safety of advanced AI models do not yet exist.

- Advanced AI is especially difficult to risk-assess due to its flexibility. As summarized in Managing AI Risks, a consensus paper from 24 leading authors including Yoshua Bengio, Geoffrey Hinton, Andrew Yao, and Stuart Russel: *"Frontier AI systems develop unforeseen capabilities only discovered during training or even well after deployment. Better evaluation is needed to detect hazardous capabilities earlier."*
- Existing risk-assessment tools and techniques from similar industries aren't appropriate for assessing AI, and there are no clear industry standards for evaluating cybersecurity, biosecurity, military warfare risks from frontier AI models.
- The development of AI-specific evals is nascent, and hasn't yet provided practical standards or techniques.
- Safety evals are necessary to safely and proactively provide visibility into potential catastrophic risks from existing models. Without these evals, the next most likely mechanism to surface such risks is for a near-miss or a catastrophic incident to occur.

CONVERGENCE'S ANALYSIS

As a result, legislators are bottlenecked by the lack of effective safety evaluations when it comes to passing binding safety assessments for AI labs.

- Governmental requirements for safety assessments today are poorly specified and insufficient. Without reliable safety evals, governments cannot legislate that AI labs must conform to any specific safety evals.
- For example, in the absence of reputable safety evals, the US executive branch has been limited to directing numerous governmental agencies to evaluate dangerous AI capabilities themselves.

Developing effective safety assessments is likely to be outside the capabilities of regulatory governmental agencies.

- Across the board, regulatory governmental agencies are understaffed, underfunded, and lack the technical expertise in both AI development and specific domain expertise to develop thorough safety evals independently.
- As with the UK AI Safety Institute and the US AI Safety Institute, governments are testing the development of separate research organizations dedicated to AI safety, and in particular safety evals. These institutes are currently less than a year old, so there's not yet evidence of their effectiveness.

Developing effective safety assessments is likely to be outside the capabilities of regulatory governmental agencies.

- Across the board, regulatory governmental agencies are understaffed, underfunded, and lack the technical expertise in both AI development and specific domain expertise to develop thorough safety evals independently.
- As with the UK AI Safety Institute and the US AI Safety Institute, governments are testing the development of separate research organizations dedicated to AI safety, and in particular safety evals. These institutes are currently less than a year old, so there's not yet evidence of their effectiveness.

CONVERGENCE'S ANALYSIS

More independent systems for conducting safety assessments need to be developed in the next 5 years.

- Nearly all meaningful safety eval research is currently conducted in private by leading AI labs, who have clear conflicts of interest and are strongly incentivized to allocate their resources towards capabilities research.
- There is little financial incentive for third-parties (i.e. organizations that aren't AI labs) to develop safety evals. There are very few reputable third parties developing non-alignment-focused safety audits of frontier AI models, the most prominent two being METR and Apollo. Other early-stage approaches include projects at RAND and government projects such as the UK AISI.
- Legislators are unlikely to be content with leading AI labs self-conducting their risk assessments as AI models improve, and will demand or require more safety evals conducted by third-parties.
- Effective safety assessments require a substantial investment of resources, to develop the specialized expertise required for each domain of evaluation (e.g. *cybersecurity, biosecurity, military warfare*). At minimum, each specific domain within safety evaluation will require collaboration between domain experts and AI developers, and these will require continuous development to stay up-to-date with evolving AI capabilities.

AI Model Registries

What are model registries? Why do they matter?

Model registries, in the context of AI regulation, are centralized governance databases of AI models intended to track and monitor AI systems in real-world use. These registries typically mandate the submission of a new algorithm or AI model to a governmental body prior to public release.

Such registries will usually require basic information about each model, such as their purpose or primary functions, their computational size, and features of their underlying algorithms. In certain cases, they may request more detailed information, such as the model's performance under particular benchmarks, a description of potential risks or hazards that could be caused by the model, or even certification that they have passed safety assessments designed to prove that the model will not cause harm.

Model registries allow governmental bodies to keep track of the AI industry, providing an overview of key models currently available to the public. Such registries also function as a foundational tool for AI governance – enabling future legislation targeted at specific AI models.

These registries adhere to the governance model of "models as a point of entry", allowing governments to focus their regulations on individual AI models rather than regulating the entire corporation, access to compute resources, or creating targeted regulations for specific algorithmic use cases.

As these model registries are an emerging form of AI governance with no direct precedents, the requirements, methods of reporting, and thresholds vary wildly between implementations. Some registries may be publicly accessible, providing greater accountability and transparency, whereas others may be limited to regulatory use only (e.g. when model data contains sensitive or dangerous information). Some may enforce reporting of certain classes of AI algorithms (such as China), whereas others may only require leading AI models with high compute requirements (such as the US).

Note

The phrase "model registry" may also often be used to refer to a (typically) private database of trained ML models, often used as a version control system for developers to compare different training runs. This is a separate topic from model registries for AI governance.

What are some precedents for mandatory government registries?

While algorithm and AI model registries are a new domain, many precedent policies exist for tracking the development and public release of novel public products. For example, reporting requirements for pharmaceuticals is a well-established and regulated process, as monitored by the Food and Drug Administration (FDA) in the US and the European Medicines Agency (EMA) in the EU. Such registries typically require:

- Basic information, such as active ingredients, method of administration, recommended dosage, adverse effects, and contraindications.
- Mandatory clinical testing demonstrating drug safety and efficacy before public release.
- Postmarket surveillance, including requirements around incident reporting, potential investigations, and methods for drug recalls or relabeling.

Many of these structural requirements will transfer over directly to model reporting, including a focus on transparent reporting, pre-deployment safety testing by unbiased third-parties, and postmarket surveillance.

What are current regulatory policies around model registries?

China

The People's Republic of China (PRC) announced the earliest and still the most comprehensive algorithm registry requirements in 2021, as part of its Algorithmic Recommendation Provisions. It has gone on to extend the scope of this registry, as its subsequent regulations covering deep synthesis and generative AI also require developers to register their AI models.

- **Algorithmic Recommendation Provisions:** The PRC requires that algorithms with "public opinion properties or having social mobilization capabilities" shall report basic data such as the provider's name, domain of application, and a self-assessment report to an algorithm registry within 10 days of publication. This requirement was primarily aimed at recommendation algorithms such as those used in TikTok or Instagram, but has later been expanded to include many different definitions of "algorithms", including modern AI models.
- **Deep Synthesis Provisions, Article 19:** The PRC additionally requires that algorithms that synthetically generate novel content such as voice, text, image, or video content must be similarly filed to the new algorithm registry.
- **Generative AI Measures, Article 17:** The PRC additionally requires that generative AI algorithms such as LLMs must be similarly filed to the new algorithm registry.
  - Of note, most of the algorithms regulated here were already covered by the 2022 deep synthesis provisions, but the new Generative AI Measures more specifically target LLMs and allows for the regulation of services that operate offline.

The EU

Via the EU AI Act, the EU has opted to categorize AI systems into tiers of risk by their use cases, notably splitting permitted AI systems into *high-risk* and *limited-risk* categorizations. In particular, it requires that high-risk AI systems must be entered into an EU database for tracking.

- As specified in Article 60 & Annex VIII, this database is intended to be maintained by the European Commission and should contain basic information such as the contact information for representatives for said AI system. It constitutes a fairly lightweight layer of tracking, and appears intended to be used primarily as a contact directory alongside other, much more extensive regulatory requirements for *high-risk* AI systems.

The US

The US has chosen to actively pursue compute governance as a method of regulation – that is, it focuses on categorizing and regulating AI models by the compute power necessary to train them, rather than by the use-case of the AI model.

- In particular, it has concentrated its binding AI regulations around restricting the export of high-end AI chips to China in preparation for a geopolitical AI arms race.
- As of Biden's 2023 Executive Order on AI, there is now a set of preliminary rules requiring the registration of models meeting a certain criteria of compute power. However, this threshold has currently been set beyond the compute power of any existing models, and as such is likely only to impact the next generation of LLMs.
  - **Section 4.2.b** specifies that the reporting requirements are enforced for models trained with greater than 10^{26} floating-point operations, or computing clusters with a theoretical maximum computing capacity of 10^{20} floating-point operations per second.
    - For comparison, GPT-4, one of today's most advanced models, was likely trained with approximately 10^{25} floating-point operations.
  - Reporting requirements seem intentionally broad and extensive, specifying that qualifying companies must report on an ongoing basis:
    - **Section 4.2.i.a:** Any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats.
    - **Section 4.2.i.b:** The ownership and possession of the model weights of

any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights.

- **Section 4.2.i.c:** The results of any developed dual-use foundation model's performance in relevant AI red-team testing.

Convergence's Analysis

Model registries appear to be a critical tool for governments to proactively enforce long-term control over AI development.

- The US, EU, and China have now incorporated some form of a model registry as a supplement to their existing regulatory portfolio.
- In particular, the types of models that each governmental body requires to be registered is a clear indicator of its longer-term priorities when it comes to AI regulation, as discussed below.
- We should expect that additional safety assessments and recurring monitoring reports will be required for models from leading governmental bodies as AI capabilities accelerate.

The US, EU, and China are pursuing substantially differing goals in their approaches to model registries as an entry point to regulation.

- In China, the model registry appears to be first and foremost a tool for aligning algorithms with the political and social agendas of the Chinese Communist Party. It's focused largely on tracking algorithmic use cases that involve recommending and generating novel content to Chinese users, particularly those with "public opinion properties" or "social mobilization capabilities".
- In the EU, AI legislation is preoccupied primarily with protecting the rights and freedoms of its citizens. As a result, the high-risk AI systems for which it requires registration are confined primarily to use cases deemed dangerous in terms of reducing equity, justice, or access to basic resources such as healthcare or education.
- The US government appears to have two primary goals: to control

CONVERGENCE'S ANALYSIS

the potential risks and distribution of frontier AI models, and to avoid limiting the current rate of AI development.

- In particular, it has decided to require registration for cutting-edge LLMs solely based on their raw performance metrics, rather than considering any specific use case, in contrast to both China and the EU.
- Additionally, it appears to be placing a priority on protecting these models from external cybersecurity threats, requiring that organizations report the measures it has taken to protect these models from being accessed or stolen. Given its current position on the export of high-end AI chips and its long history with military IP theft, it's clear that the US views the protection of cutting-edge AI models as a national security threat.
- Finally, none of these model registry requirements will come into effect until the next generation of frontier AI models is released sometime in 2024 or 2025. To this point, the Biden administration has cautiously avoided creating any binding regulations that might impede the rate of AI capabilities development among leading American AI labs.

Model registries will serve as a foundational tool for governments to enact additional regulations around AI development.

- Much in the same way drug registries are used as a foundational tool for the FDA to control the development and public usage of pharmaceuticals, model registries will be a critical component for governments to control public AI model usage.
- Model registries will enable the creation and improved enforcement of regulations such as:
  - Mandating specific sets of pre-deployment safety assessments, or certification by certain organizations before public deployment
  - Transparency requirements for AI models such as disclosures
  - Incident reporting involving specific models and civil liabilities for damages caused by specific AI models
  - Postmarket surveillance such as post-deployment evaluations, regulatory investigations, and the potential disabling of non-compliant or risky models

AI Incident Reporting

What is AI incident reporting?

AI incident reporting refers to an emerging series of voluntary practices or regulatory requirements for AI developers and deployers to publicly report adverse effects or "near-misses" that arise from the use of AI systems. Such mechanisms are designed to capture a wide range of potential issues, such as privacy breaches, security vulnerabilities, and biases in decision-making.

In most domains, incidents can be divided into two subcategories:

- An *accident* is a type of incident that caused significant damage, injury, or harm to a person, property, or equipment.
- A *near-miss* is a type of incident that had the potential to cause significant damage, injury, or harm, but was narrowly avoided.

The rationale for incident reporting is to create a feedback loop where regulators, developers, and the public can learn from past AI deployments to continuously improve safety standards and legal compliance. By systematically documenting incidents, stakeholders can identify patterns, initiate inquiries into causes of failure, and implement corrective measures to prevent their recurrence.

What precedent policies exist for AI incident reporting?

Incident reporting has been a highly effective tool used across a variety of industries for decades to mitigate risk from emerging technologies. Here are two examples:

- The Aviation Safety Reporting System (ASRS) has been noted for its effectiveness at drastically reducing the fatality rate in US aviation. Its success has been attributed to its *confidential, voluntary, and non-punitive* approach: anybody can submit a confidential incident report of a near-miss or an abuse of safety standards to a neutral third-party organization (in this case, NASA). The reporting aviation worker is typically granted *limited immunity,* which encourages more reporting without fear of reprisals. In response to incidents, the ASRS typically distributes non-binding notices summarizing key failures and recommending new industry standards.
  - It's important to note that *accidents* still have mandatory reporting requirements via the FAA, and that the ASRS is a supplementary system.
  - The Occupational Safety and Health Administration (OSHA) is a governmental agency tasked with guaranteeing safe conditions for

American workers by setting and enforcing workplace standards. Its primary day-to-day responsibility is following up on incident reports of unsafe work practices, injuries, and fatalities by investigating corporations. It enforces its standards primarily by assessing hefty fines on organizations for non-compliance.

- Independent reports have found that OSHA has resulted in a modest increase in workplace safety, reducing worker injuries by a modest four percent.

Incident reporting in AI is still in its nascent stages, and a variety of approaches are being explored globally. The specific requirements for incident reporting, such as the types of incidents that must be reported, the timeframe for reporting, and the level of detail required can vary significantly between jurisdictions.

The most prominent public example of an AI incident reporting tool today is the AI Incident Database, launched by the Responsible AI Collaborative. This database crowdsources incident reports involving AI technologies as documented in public sources or news articles. It's used by AI researchers as a tool to surface broad trends and individual case studies regarding AI safety incidents. As a voluntary public database, it doesn't adhere to any regulatory standards nor does it require input or resolution from the developers of the AI tool involved.

What are current regulatory policies around AI incident reporting?

China

The PRC is developing a governmental incident reporting database, as announced in the Draft Measures on the Reporting of Cybersecurity Incidents on Dec 20th, 2023. This proposed legislation categorize cybersecurity incidents into four categories of severity ("Extremely Severe", "Severe", "Relatively Severe", and "General"), and requires that the top three levels ("Critical Incidents") are reported to governmental authorities within one hour of occurrence. The criteria for meeting the level of "Critical" incidents include the following:

- *Interruption of overall operation of critical information infrastructure for more than 30 minutes, or its main function for more than two hours;*
- *Incidents affecting the work and life of more than 10% of the population in a single city-level administrative region;*
- *Incidents affecting the water, electricity, gas, oil, heating or transportation*

usage of more than 100,000 people;

- *Incidents causing direct economic losses of more than RMB 5 million* (around $694k USD)

Though this set of measures does not directly mention frontier AI models as a target for enforcement, any of the negative outcomes above resulting from the use of frontier AI models would be reported under the same framework. This draft measure can be understood as the Cyberspace Administration of China (CAC) pursuing two major goals:

1. Consolidating disparate reporting requirements across various laws regarding cybersecurity incidents.
2. Developing regulatory infrastructure in preparation for an evolving cybersecurity landscape, particularly with respect to advanced AI.

Elsewhere, leading Chinese AI regulatory measures make reference to reporting key events (specifically the distribution of unlawful information) to the Chinese government, but none of them have specific requirements for the creation of an incident reporting database:

- **Algorithmic Recommendation Provisions, Article 7:** Service providers shall…establish and complete management systems and technical measures…[such as] security assessment and monitoring and security incident response and handling.
  - **Article 9:** Where unlawful information is discovered…a report shall be made to the cybersecurity and informatization department and relevant departments.
- **Deep Synthesis Provisions, Article 10:** Where deep synthesis service providers discover illegal or negative information, they shall…promptly make a report to the telecommunications department or relevant departments in charge.
- **Generative AI Measures, Article 14:** Where providers discover illegal content they shall promptly employ measures to address it such as stopping generation, stopping transmission, and removal, employ measures such as model optimization training to make corrections and report to the relevant departments in charge.

The EU

The EU AI Act requires that developers of both *high-risk* AI systems and *general purpose AI ("GPAI")* systems set up internal tracking and reporting systems for "serious incidents" as part of their post-market monitoring infrastructure.

As defined in Article 3(44), a serious incident is:

Any incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:

- (a) the death of a person or serious damage to a person's health
- (b) a serious and irreversible disruption of the management and operation of critical infrastructure
- (ba) breach of obligations under Union law intended to protect fundamental rights
- (bb) serious damage to property or the environment.

In the event that such an incident occurs, Article 62 requires that the developer reports the incident to the relevant authorities (specifically the European Data Protection Supervisor) and cooperate with them on an investigation, risk assessment, and corrective action. It specifies time limits for reporting and specific reporting obligations.

The US

The US does not currently have any existing or proposed legislation regarding reporting databases for AI-related incidents. However, the Executive Order on AI contains some preliminary language directing the Secretary of Health and Human Services (HHS) and the Secretary of Homeland Security to establish new programs within their respective agencies. These directives essentially request the creation of domain-specific incident databases:

- **Section 5.2:** The Secretary of Homeland Security…shall develop a training, analysis, and evaluation program to mitigate AI-related IP risks. Such a program shall: (i) include appropriate personnel dedicated to collecting and analyzing reports of AI-related IP theft, investigating such incidents with implications for national security, and, where appropriate and consistent with applicable law, pursuing related enforcement actions.
- **Section 8:** The Secretary of HHS shall…consider appropriate actions [such as]...establish[ing] a common framework for approaches to identifying and capturing clinical errors resulting from AI deployed in healthcare settings as well as specifications for a central tracking repository for associated incidents that cause harm, including through bias or discrimination, to patients, caregivers, or other parties.

Convergence's Analysis

In the next 2-3 years, the US, EU, and China will have established mandatory incident reporting requirements by AI service providers for "severe" incidents encompassing AI technologies.

- Each of these leading governments is currently developing or has tasked their internal agencies with the responsibility to develop systems to track and enforce mandatory incident reporting.
- As defined in the previous section, such "severe" incidents will typically include significant monetary damages, injury or death to a person, or the disruption of critical infrastructure.
- In many cases (such as the US and China today), these reporting requirements may not be designed specifically for AI incidents, but rather include them as aspects of more specific domains of use-cases, such as cybersecurity, IP theft, or healthcare. Enforcement of these reporting requirements may be spread across a variety of agencies.
- Similar to governmental agencies like OSHA, these incident reporting systems will enforce compliance via mandatory reporting, comprehensive reviews following qualifying reports, and applying substantial fines for negligence.

However, such governmental compliance requirements represent only the minimum base layer of an effective network of incident reporting systems to mitigate risk from AI technologies.

There exist several notable precedents from other domains of incident reporting that have yet to be developed or addressed by the AI governance community:

- **Voluntary, confidential or non-punitive reporting systems:** Incident reporting systems similar to the Aviation Safety Reporting System (ASRS) as described previously do not yet exist. In particular, a substantial gap exists for a non-regulatory organization to focus on consolidating confidentially reported incidents, conducting independent safety evals, and publishing

CONVERGENCE'S ANALYSIS

reports on best practices for the benefit of the entire AI safety community.

- **Near-miss reporting systems:** Similarly, near-miss reporting involves disclosing incidents that could have resulted in injury, harm, or damage but were avoided. Such proactive reporting is a key tool to help organizations prevent "severe" incidents, by developing insight into the root causes behind safety issues before they occur. Given that AI systems are widely predicted to have the potential to cause catastrophically dangerous incidents, responsible disclosure of near-miss incidents remains a critical gap.
- **International coordination:** Most incident reporting systems today are implemented on a national level. To promote the sharing of critical knowledge, key industries have developed bodies of international cooperation, such as the International Confidential Aviation Safety Systems (ICASS) Group or incident reporting systems managed by the International Atomic Energy Agency. Currently, there's no legitimate international coordination proposals for AI incident reporting. We expect to see the development of these international bodies enter the discussion in the next ~2-3 years, after national regulatory bodies are created and standardized.

Open-Source AI Models

What does open-source mean in the context of the development and deployment of AI models?

Some software developers choose to *open-source* their software; they freely share the underlying source code and allow anyone to use, modify, and deploy their work. This can encourage friendly collaboration and community-building, and has produced many popular pieces of software, including operating systems like Linux, programming languages and platforms like Python and Git, and many more.

Similarly, AI developers are open-sourcing their models and algorithms, though the details can vary. Generally, open-sourcing of AI models involves some combination of:

- Sharing the *model weights.* These are the specific parameters that make the model function, and are set during training. If these are shared, others can reconstruct the model without doing their own training, which is the most expensive part of developing such AI.
- Sharing the training data used to train the model.
- Sharing the underlying source code.
- Licensing for free commercial usage.

For example, Meta released the model weights of their LLM, Llama 2, but not their training code, methodology, original datasets, or model architecture details. In their excellent article on Openness In Language Models, Prompt Engineering labels this an example of an "open weight" model. Such an approach allows external parties to use the model for inference and fine-tuning, but doesn't allow them to meaningfully improve or analyze the underlying model. Prompt Engineering points out a drawback of this approach:

> So, open weights allows model use but not full transparency, while open source enables model understanding and customization but requires substantially more work to release [...] If only open weights are available, developers may utilize state-of-the-art models but lack the ability to meaningfully evaluate biases, limitations, and societal impacts. Misalignment between a model and real-world needs can be difficult to identify.

Further, while writing this article in April 2024, Meta released Llama 3 with the same open-weights policy, claiming that it is "the most capable openly

available LLM to date". This has brought fresh attention to the trade-offs of open-sourcing, as the potential harms of freely sharing software are greater the more powerful the model in question is. Even those who are fond of sharing wouldn't want everyone in the world to have easy access to the instructions for a 3D-printable rocket launcher, and freely sharing powerful AI could present similar risks; such AI could be used to generate instructions for assembling homemade bombs or even designing deadly pathogens. Distributing information of this nature widely is termed an *information hazard*.

To prevent these types of hazards, AI models like ChatGPT have safeguards built in during the *fine-tuning* phase towards the end of their development (implementing techniques such as Reinforcement Learning by Human Feedback, or RLHF). This technique can limit AI models from producing harmful or undesired content.

Some people find ways to get around this fine-tuning, but experts have pointed out that malicious actors could circumvent the problem entirely. ChatGPT and Claude, the two most prominent LLMs are closed-source (and their model weights are closely guarded secrets), but open-source models can be used and deployed without fine-tuning safeguards. This was demonstrated practically with Llama 2, a partly open-source LLM developed by Meta in Palisade Research's paper BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. To quote an interview with one of its authors Jeoffrey Ladish:

> You can train away the harmlessness. You don't even need that many examples. You can use a few hundred, and you get a model that continues to maintain its helpfulness capabilities but is willing to do harmful things. It cost us around $200 to train even the biggest model for this. Which is to say, with currently known techniques, if you release the model weights there is no way to keep people from accessing the full dangerous capabilities of your model with a little fine tuning.

Therefore, these models and their underlying software may themselves be information hazards, and many argue that open-sourcing advanced AI should be legally prohibited, or at least prohibited until developers can guarantee the safety of their software. In "Will releasing the weights of future large language models grant widespread access to pandemic agents?", the authors conclude that

> Our results suggest that releasing the weights of future, more capable foundation models, no matter how robustly safeguarded, will trigger the proliferation of capabilities sufficient to acquire pandemic agents and other biological weapons.

Others counter that openness is necessary to stop the power and wealth generated by powerful AI falling into the hands of a few, and that prohibitions

won't be effective safeguards, as argued in GitHub's Supporting Open Source and Open Science in the EU AI Act and Mozilla's Joint Statement on AI Safety and Openness, which was signed by over 1,800 people and states:

> Yes, openly available models come with risks and vulnerabilities — AI models can be abused by malicious actors or deployed by ill-equipped developers. However, we have seen time and time again that the same holds true for proprietary technologies — and that increasing public access and scrutiny makes technology safer, not more dangerous. The idea that tight and proprietary control of foundational AI models is the only path to protecting us from society-scale harm is naive at best, dangerous at worst.

Finally, some argue that open-sourcing or not is a false dichotomy, putting forward intermediate policies such as structured access:

> Instead of openly disseminating AI systems, developers facilitate controlled, arm's length interactions with their AI systems. The aim is to prevent dangerous AI capabilities from being widely accessible, whilst preserving access to AI capabilities that can be used safely.

Some researchers also are trying to build open-source models that are resistant to post-deployment fine-tuning and misuse, such as a paper from April 2024, in which researchers with Zhejiang University and Ant Group describe a new technique called "non-finetunable-learning", which

> prevents the pre-trained model from being finetuned to indecent tasks while preserving its performance on the original task.

However, this technique is novel and, as pointed out by Jack Clark, it requires you to know what misuse you want to prevent in advance, and has only been tested on small, narrow-purpose models.

There are more perspectives and arguments than we can concisely include here, and you might be interested in the following discussions:

Open-Sourcing Highly Capable Foundation Models
Centre for the Governance of AI ↗

Are open models safe?
Linux Foundation ↗

The Mirage of open-source AI: Analyzing Meta's Llama 2 release strategy
Open Future

Should we make our most powerful AI models open source to all?
Vox

Thoughts on open source AI
Sam Marks on the AI Alignment Forum

Navigating the Open-Source AI Landscape: Data, Funding, and Safety
André Ferretti & mic on LessWrong

Will releasing the weights of large language models grant widespread access to pandemic agents?
jefftk on LessWrong

Propaganda or Science: A Look at Open Source AI and Bioterrorism Risk
1a3orn on LessWrong

What are current regulatory policies around AI incident reporting?

The US

The US AI Bill of Rights doesn't discuss open-source models, but the Executive Order on AI does initiate an investigation into the risk-reward tradeoff of open-sourcing. Section 4.6 calls for soliciting input on foundation models with "widely available model weights", specifically targeting open-source models. Section 4.6 summarizes the risk-reward tradeoff of publicly sharing model weights, which offers "substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model". In particular: 4.6 calls for the Secretary of Commerce to:

- **Section 4.6(a):** Set up a public consultation with the private sector, academia, civil society, and other stakeholders on the impacts and appropriate policy related to dual-use foundation models with widely available weights ("such models" below), including:
  - **4.6(a)(i):** Risks associated with fine-tuning or removing the safeguards from such models;

  - **4.6(a)(ii):** Benefits to innovation, including research into AI safety and risk management, of such models;
  - **4.6(a)(iii):** Potential voluntary, regulatory, and international mechanisms to manage risk and maximize the benefits of such models;
- **4.6(b):** Submit a report to the president based on the results of 4.6(a), on the impacts of such models, including policy and regulatory recommendations.

The EU

The EU AI Act states that open-sourcing can increase innovation and economic growth. The act therefore exempts open-source models and developers from some restrictions and responsibilities placed on other models and developers. Note though that these exemptions **do not apply** to foundation models (meaning generative AI like ChatGPT), or if the open-source software is monetized or is a component in high-risk software.

- **Section 57:** Places responsibilities on providers throughout the "AI value chain", i.e. anyone developing components or software that's used in AI. Third parties should be exempt if their products are open-source, though it encourages open-source developers to implement documentation practices, such as model cards and data sheets.
- **Section 60i & i+1:** Clarifies that GPAI models released under free and open-source licenses count as satisfying "high levels of transparency and openness" if their parameters are made publicly available, and a license should be considered free and open-source when users can run, copy, distribute, study, change, and improve the software and data. This exception does not apply if the component is monetized in any way.
- **Section 60f:** Exempts providers of open-source GPAI models from the transparency requirements *unless* they present a systemic risk. This does *not* exempt GPAI developers from the obligation to produce a summary about training data or to enact a copyright policy.
- **Section 60o:** Specifies that developers of GPAI models should notify the AI Office if they're developing a GPAI model that exceeds certain thresholds (therefore conferring systemic risk), and that this is especially important for open-source models.
- **Article 2(5g):** States that obligations shall not apply to AI systems released under free and open-source licenses unless they are placed on the market or put into service as high-risk AI systems.
- **Article 28(2b):** States that providers of high-risk AI systems and third parties providing components for such systems have a written agreement on what information the provider will need to comply with the act. However, third parties publishing "AI components other than GPAI models under a free and open licence" are exempt from this.

- **Article 52c(-2) & 52ca(5):** Exempt providers of AI models under a free and open licence that publicly release the weights and information on their model from (1) the obligation to draw up technical documentation and (2) from the requirement to appoint an authorized representative in the EU. Neither of these exemptions apply if the GPAI model has systemic risks.

Notably, the treatment of open-source models was contentious during the development of the EU AI Act (see also here).

China

There is no mention of open-source models in China's regulations between 2019 and 2023; open-source models are neither exempt from any aspects of the legislation, nor under any additional restrictions or responsibilities.

Convergence's Analysis

The boundaries and terminology around open-sourcing are often underspecified.

- Open-sourcing vs closed-sourcing AI models is not binary, but a spectrum. Developers must choose whether to publicly release multiple aspects of each model: the weights and parameters of the model; the data used to train the model; the source code and algorithms underlying the model and its training; licenses for free use; and so on.
- Existing legislation does not clearly delineate how partially open-sourced models should be categorized and legislated. It's unclear, for example, whether Meta's open-weight Llama-2 model would be considered open-source under EU legislation, as its source code is not public.

Open-sourcing models improves transparency and accountability, but also gives the public broader access to dangerous information and reduces the efficacy of legislation. There is widespread disagreement on the right balance.

CONVERGENCE'S ANALYSIS

- Through their training on vast swathes of data, LLMs contain hazardous information. Although RLHF is not sufficient to stop users accessing underlying hazardous information, it is a barrier, and one that can be much more easily bypassed in open-sourced models.
- The more powerful a model is, the greater harm its misuse could lead to, and the more open-source a model is, the more easily misused it is. This means the potential harms of open-source models will increase over time.
- Open-source models can be easily used and altered by potentially any motivated party, making it harder to implement and enforce safety legislation.
- However, many experts are still staunch advocates for open-sourcing (as listed in the Context section), and believe it is essential for an accountable and transparent AI ecosystem. There is profound disagreement on the right balance between open and closed-source models, and such disagreement is likely to persist.

Developers of open-source models are not currently under any additional legal obligations compared to developers of private or commercial models.

- In particular, the US Executive Order and Chinese regulations currently have no particular rules unique to open-source models or developers, though the US does recognize the risk-reward tradeoff presented by open-source AI, and has commissioned a report into its safety and appropriate policy.

The EU legislation treats open-source models favorably.

- Unlike the US Executive Order, the EU AI Act only describes the potential benefits of open-sourcing powerful models, without mentioning potential risks.
- The EU AI act exempts open-source developers from many obligations faced by commercial competitors, unless the open-sourced software is part of a general-purpose or high-risk system.
- Despite this, and despite the exemptions, proponents of open-sourcing have criticized the EU regulations for what they perceive as over-regulation of open-source models.

CONVERGENCE'S ANALYSIS

- For examples, see GitHub's How to get AI regulation right for open source post and Supporting Open Source and Open Science in the EU AI Act letter, and Brookings' The EU's attempt to regulate open-source AI is counterproductive (note that the latter was written in 2022, prior to redrafts that altered open-source requirements).

Cybersecurity of Frontier AI Models

What cybersecurity issues arise from the development of frontier AI models?

One of the primary issues that has caught the attention of regulators is the protection of the intellectual property and sensitive data associated with frontier AI models (otherwise named as *"dual-use foundational models"* by US directives and *"general-purpose AI" ("GPAI")* by EU legislation).

In particular, legislators are concerned that as frontier AI models increase their capabilities, unregulated access to the underlying code or abilities of these models will result in dangerous outcomes. For example, current AI models are susceptible to easily distributing *information hazards*, such as the instructions to develop homemade weapons or techniques to commit crimes. As a result, they're typically trained during a fine-tuning phase to reject such requests. Bypassing the cybersecurity of such models could result in the removal of such fine-tuning, allowing dangerous requests. Other cybersecurity risks include sharing sensitive user data, or leaking proprietary ML architectural decisions with direct competitors & geopolitical adversaries (e.g. Chinese organizations, in the case of the US).

Currently, the leading frontier AI models meet the following conditions, which are often collectively referred to as *"closed-source"* development:

- Are privately owned by a large AI lab (e.g. OpenAI, Anthropic, or Google)
- Present an API interface to fine-tuned models that are designed to reject dangerous or adversarial inputs.
- Do not have publicly shared training data or codebases
- Do not have publicly shared model weights, which would allow for the easy replication of the core functionality of an AI model by third-parties
- Encrypt and protect user data, such as LLM queries and responses

In contrast, *open-source* AI models typically share some combination of their training data, model code, and completed model weights for public and commercial use.

Unlike open-source models, which are freely available and lack cybersecurity protections by design, proprietary or closed-source models have stringent measures to safeguard such sensitive information. Preventing the theft or leakage of this information is critically important to the AI labs that develop these models, as it constitutes their competitive advantage and intellectual property.

What cybersecurity issues are AI labs concerned about?

Specifically, AI labs are concerned about preventing the following:

- **Leaking private user data** would cause a company to violate key international privacy laws such as the GDPR, leading to substantial fines and loss of user trust.
- **Leaking the model weights** of a frontier AI model would lead to external parties being able to run the model independently and remove any fine-tuning that protects from adversarial inputs.
- **Leaking the codebase** would allow competing labs to learn directly from an organization's technical decisions and accelerate competition.
- **Leaking the training data** would allow competing labs to better train their models by incorporating new data, accelerating competition.

With effective security practices, it's generally accepted that it is feasible for AI labs to prevent these forms of information being leaked. Similar practices are currently used in all major tech corporations today to prevent their existing codebases and private user data from data breaches. Nevertheless, given the complexity of cybersecurity and the numerous potential targets, it is highly likely that a prominent AI lab will fall victim to a data breach involving a frontier AI model in the near future.

What cybersecurity issues are regulators concerned about?

Regulators are similarly concerned about effective cybersecurity for the same domains, albeit with different motivations:

- Regulators currently strongly **prioritize the protection of user data** stored by companies, as a tenet of basic privacy rights as described in binding legislation such as the GDPR or China's Personal Information Protection Law, or non-binding declarations such as the US AI Bill of Rights' declaration on data privacy.
- Regulators are just beginning to demand **adequate protection of model weights, codebase, and training data** of frontier AI models, for two reasons:
  1. Leaking such data could **benefit the R&D of geopolitical adversaries.** In particular, the US government is highly invested in limiting the rate of AI development of Chinese organizations – leaking such data would counter these interests.
  2. Leaking such data could **allow third-parties to develop unregulated access to potentially dangerous frontier AI models.** Currently, governments have well established methods to control closed-source

models run by AI labs, by regulating the labs themselves. If access to the source code of these frontier models were more widely distributed, regulators would lose their ability to control the usage and distribution of these models.

Due to these interests, regulators are generally as invested in the cybersecurity of frontier AI models as the labs themselves are. Their incentives are well aligned in the case of cybersecurity for frontier models. However, in practice regulators have by and large left specific cybersecurity decisions up to independent parties, preferring to more broadly create requirements such as a "primary responsibility for information security" or "resilien[ce] against attack from third-parties". Their enforcement of legislation such as the GDPR has been inconsistent and patchy.

What are current regulatory policies around cybersecurity for AI models?

China

China maintains a complex, detailed, and thorough set of data privacy requirements developed over the past two decades via legislation such as the PRC Cybersecurity Law, the PRC Data Security Law, and the PRC Personal Information Protection Law. Together, they constitute strong protections mandating the confidential treatment and encryption of personal data stored by Chinese corporations. Additionally, the PRC Cybersecurity Law has requirements regarding data localization that mandate that the user data of Chinese citizens be stored on servers in mainland China, ensuring that the Chinese government has more direct methods to access and control the usage of this data. All of these laws apply to data collected from users of LLM models in China.

China's existing AI-specific regulations largely mirror the data privacy policies laid out in previous legislation, and often refer directly to such legislation for specific requirements. In particular, they extend data privacy requirements to the training data collected by Chinese organizations. However, they do not introduce any specific requirements for the cybersecurity of frontier AI models, such as properly securing model weights or codebases.

China's Deep Synthesis Provisions include the following:

- **Article 7:** Requires service providers to implement primary responsibility for information security, such as data security, personal information protection, and technical safeguards.
- **Article 14:** Requires service providers to strengthen the management and security of training data, especially personal information included in

training data.

China's Interim Generative AI Measures include the following:

- **Article 7:** Requires service providers to handle training data in accordance with the Cybersecurity Law and Data Security Law when carrying out pre-training and optimization of models.
- **Article 9:** Requires that service providers bear responsibility for fulfilling online information security obligations in accordance with the law.
- **Article 11:** Requires providers to keep user input information and usage records confidential and not illegally retain or provide such data to others.
- **Article 17:** Requires security assessments for AI services with public opinion properties or social mobilization capabilities.

The EU

The EU has a comprehensive data privacy and security law that applies to all organizations operating in the EU or handling the personal data of EU citizens: the General Data Protection Regulation (GDPR). Passed in 2018, it does not contain language specific to AI systems, but provides a strong base of privacy requirements for collecting user data, such as mandatory disclosures, purpose limitations, security, and rights to access one's personal data.

The EU AI Act includes some cybersecurity requirements for organizations running "*high-risk* AI systems" or "general purpose AI models with systemic risk". It generally identifies specific attack vectors that organizations should protect against, but provides little to no specificity about how an organization might protect against these attack vectors or what level of security is required.

Sections discussing cybersecurity for AI models include:

- **Article 15:** High-risk AI systems should be resilient against attacks by third-parties against system vulnerabilities. Specific vulnerabilities include:
  - Attacks trying to manipulate the training dataset ('data poisoning')
  - Attacks on pre-trained components used in training ('model poisoning')
  - Inputs designed to cause the model to make a mistake ('adversarial examples' or 'model evasion')
  - Confidentiality attacks or model flaws
- **Article 52d:** Providers of general-purpose AI models with systemic risk shall:
  - Conduct adversarial testing of the model to identify and mitigate systemic risk
  - Assess and mitigate systemic risks from the development, market introduction, or use of the model

- Document and report serious cybersecurity incidents
- Ensure an adequate level of cybersecurity protection

The US

Compared to the EU and China, the US Executive Order on AI places the greatest priority on the cybersecurity of frontier AI models (beyond data privacy requirements), in accordance with the US' developing interest in limiting Chinese access to US technologies. It is developing specific reporting requirements regarding cybersecurity for companies developing dual-use foundation models, and has requests for reports out to various agencies to investigate AI model cybersecurity implications across a number of domains.

Specific regulatory text in the Executive Order includes:

- **Section 4.2:** This section establishes reporting requirements to the Secretary of Commerce for measures taken to protect the model training process and weights of dual-use foundational models, including:
  - (a) Companies developing dual-use foundation models must provide information on physical and cybersecurity protections for the model training process, model weights, and the result of any read-team testing for model security
  - (b) Directs the Secretary of Commerce to define the technical conditions for which models would be subject to the reporting requirements in 4.2(a). Until defined, this applies to any model trained using
    - (i) Over 10^{26} integer/floating-point operations per second (FLOP/s)
    - (ii) Over 10^{23} FLOPs if using primarily biological sequence data
    - (iii) Any computing cluster with data center networking of over 100 Gbit/s and a maximum computing capacity of 10^{20} FLOPs for training AI.
- **Section 4.3:** This section requires that a report is delivered to the Secretary of Homeland Security in 90 days on potential risks related to the use of AI in critical infrastructure sectors, including ways in which AI may make infrastructure more vulnerable to critical failures, physical attacks, and cyber attacks.
  - It also requests that the Secretary of the Treasury issue a public report on best practices for financial institutions to manage AI-specific cybersecurity risks.
- **Section 4.6:** The Secretary of Commerce shall solicit input for a report evaluating the risks associated with open-sourced model weights of dual-use foundational models, including the fine-tuning of open-source models, potential benefits to innovation and research, and potential mechanisms to manage risks.

- **Section 7.3:** The Secretary of HHS shall develop a plan [that includes the]... incorporation of safety, privacy, and security standards into the software-development lifecycle for protection of personally identifiable information, including measures to address AI-enhanced cybersecurity threats in the health and human services sector.

The US does not have a comprehensive data privacy law similar to the GDPR or the PRC Personal Information Protection Law, nor a comprehensive cybersecurity law similar to the PRC Cybersecurity Law.

Convergence's Analysis

User data of frontier AI models, and some forms of training data will continue to fall under the jurisdiction of existing data privacy laws.

- The mandatory protection of user data (such as encryption) has been well established and legislated over the past decade via legislation such as the GDPR or the PRC Personal Information Protection Law. In practice, these laws have been effective at achieving their goals. There's no clear reason to establish a separate set of regulations solely for user data regarding AI models.
- Training data used for developing AI models can sometimes include private or sensitive user data. As specified in China's regulations, this data will also be protected under existing legislation, and specific clauses may be included to indicate that requirement.

Cybersecurity requirements beyond user privacy are likely to be targeted at a small group of leading AI labs.

- As evidenced by the US Executive Order's approach to reporting requirements on cybersecurity, the US is primarily concerned about mitigating technological poaching of leading AI models and systemic risks. It has set a reasonably high threshold for reporting, excluding all but the top 3-4 labs at this time.
- The majority of companies using frontier AI models are likely to pay for access via APIs from leading AI labs, and therefore do not

CONVERGENCE'S ANALYSIS

have many of the cybersecurity risks described above. As a result, such legislation is likely to be more targeted at a small group of AI labs and more closely enforced than data privacy laws.

Frontier AI labs already have strong incentives to enforce the protection of their closed-source AI models. It's unlikely that mandatory legislation will meaningfully impact their cybersecurity efforts.

- Leading AI labs have significant resources and technical expertise, and a strong vested interest in protecting their IP. As a result, they typically have large teams dedicated to cybersecurity, and tend to operate state-of-the-art security practices. Though these requirements seem plausible to legislate based on government interests, they are unlikely to drastically change the approach for frontier AI labs regarding cybersecurity.

Governments have historically been poor at enforcing data privacy requirements, and are mostly constrained to requiring reporting or reactively fining organizations after an incident occurs.

- Practically, government agencies have not had the resources to conduct thorough audits of their cybersecurity requirements. As a result, enforcement of legislation such as the GDPR has been sporadic and inconsistent. We expect similar outcomes for cybersecurity laws around AI models.
- In addition, legislative requirements around cybersecurity are intentionally vague because of their broad scope. For instance, the GDPR only requires that organizations *"shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk"*. Such wording requires that each organization be considered on a case-by-case basis and opens the case for protracted legal disputes over fines.

When securing model weights, code, and training data of frontier AI models, the types of cybersecurity required can be much more complicated, as each new domain opens up new attack vectors. Governmental agencies likely don't have the capabilities to thoroughly evaluate the complex cybersecurity practices of frontier AI labs. However, having a significantly reduced number of

CONVERGENCE'S ANALYSIS

organizations to track (primarily leading AI labs) may aid enforcement.

AI Discrimination Requirements

What are discrimination requirements for AI? Why do they matter?

Discrimination requirements for AI are rules and guidelines aimed at preventing AI systems from perpetuating or amplifying societal biases and unfairly disadvantaging certain groups of people based on protected characteristics like race, gender, age, religion, disability status, or sexual orientation. As AI increasingly powers high-stakes decision making in areas like hiring, lending, healthcare, criminal justice, and public benefits, these systems are likely to adversely impact certain subsets of the population without algorithmic bias management.

For example, an algorithm designed to identify strong resumes for a job application is likely to predict correlations between the sex of a candidate and the quality of their resume, reflecting existing societal biases (and therefore perpetuating them). As a result, certain classes of individuals may be adversely impacted by an algorithm that contains inherently discriminatory word associations.

Other examples for algorithmic discrimination include:

- Biases in the type of online ads presented to website users
- Biases in the error rates of facial recognition technology by race and gender
- Biases in algorithms designed to predict risk in criminal justice

The usage of discriminatory factors such as sex, ethnicity, or age has been expressly prohibited by longstanding anti-discriminatory legislation around the globe, such as Title VII of the US Civil Right Act of 1964, the U.N.'s ILO Convention 111, or Article 21 of the EU Charter of Fundamental Rights. As enforced by most developed countries, such legislation typically protects citizens of a governmental body from employment or occupational discrimination based on these factors.

To expand these legislative precedents to the rapidly developing domain of algorithmic and AI discrimination, a new crop of anti-discrimination legislation is being passed by leading governmental bodies. This new wave of legislation focuses on regulating the behavior of the algorithms underlying certain protected use cases, such as resume screening, creditworthiness evaluations, or public benefit allocations.

As the momentum grows to address AI bias, governments are starting to pass laws and release guidance aimed at preventing automated discrimination. But this is still an emerging area where much more work is needed to translate principles into practice. Active areas of research and policy development

include both technical and non-technical measures such as:

- **De-biasing dataset frameworks:** Dataset managers can carefully curate more balanced and representative training data by adjusting the significance of specific data points to correct for known imbalances or using autonomous testing methods to identify and correct for dataset biases. For instance, a revised dataset allowed Microsoft to reduce the face recognition error ratio between men and women with darker skin tones by 20-fold.
- **Algorithmic & dataset transparency:** Organizations can implement public processes around measuring and reporting bias. For example, Google has introduced a Model Card reporting system that explains the employed data and algorithm, details performance evaluations, and disclose intended use cases. Such transparency encourages public review and accountability.
- **Third-party evaluations:** A standardized system of review for AI algorithms would force organizations to adhere to comprehensive requirements for reducing discrimination. Various high-level solutions have been proposed by major organizations like the OECD and the European Convention on Human Rights, but no industry standards for measuring bias have been agreed upon.

What are current regulatory policies around discrimination requirements for AI?

China

Two major pieces of Chinese legislation have made references to combating AI discrimination. Though the language around discrimination was scrapped in the first, the 2023 generative AI regulations include binding but non-specific language requiring compliance with anti-discrimination policies for AI training and inference.

- **Algorithmic Recommendation Provisions, Article 10:** The initial interim draft of this legislation prohibited the use of "discriminatory or biased user tags" in algorithmic recommendation systems. However, this language was removed in the final version effective in March 2022.
- **Generative AI Measures, Article 4.2:** This draft calls for the following: "During processes such as algorithm design, the selection of training data, model generation and optimization, and the provision of services, effective measures are to be employed to prevent the creation of discrimination such as by race, ethnicity, faith, nationality, region, sex, age, profession, or health".

The EU

The EU AI Act directly addresses discriminatory practices classified by the use cases of AI systems considered. In particular, it classifies all AI systems with potential discriminatory practices as high-risk systems and bars them from discrimination, including:

- AI systems that could produce **adverse outcomes to health and safety of persons**, and could cause discriminatory practices.
- AI systems used in **education or vocational training**, "notably for determining access to educational…institutions or to evaluate persons on tests...as a precondition for their education".
- AI systems used in **employment**, "notably for recruitment…for making decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships".
- AI systems used to evaluate the **credit score or creditworthiness** of natural persons, or for allocating public assistance benefits
- AI systems used in **migration, asylum and border control management**

In particular, AI systems that provide **social scoring of natural persons** (which pose a significant discriminatory risk) are deemed *unacceptable systems* and are banned.

The US

The US government is actively addressing AI discrimination via two primary initiatives by the executive branch. However, both of these initiatives are non-binding and non-specific in nature: in particular, the Executive Order directs several agencies to publish guidelines, but doesn't identify any specific requirements or enforcement mechanisms.

1. The AI Bill of Rights contains an entire section on Algorithmic Discrimination Protections. In particular, it emphasizes that consumers should be protected from discrimination based on their "race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law." Though this bill is non-binding, it sets a general principle for enforcement by the US executive branch for more specific regulations.
2. The Executive Order on AI directs various executive agencies to publish reports or guidance on preventing discrimination within their respective domains within the 90–180 days after its publication. These include the

following directly responsible parties:

a **Section 7.1:** "The Attorney General of the Criminal Justice System, and the Assistant Attorney General in charge of the Civil Rights Division will publish guidance preventing discrimination in automated systems."

b **Section 7.2.b.i:** "The Secretary of HHS (The Department of Health and Human Services) will publish guidance regarding non-discrimination in allocating public benefits."

c **Section 7.2.b.ii:** "The Secretary of Agriculture will publish guidance regarding non-discrimination in allocating public benefits."

d **Section 7.3:** "The Secretary of Labor will publish guidance regarding non-discrimination in hiring involving AI."

Convergence's Analysis

The effectiveness of de-biasing techniques is highly variable, and depends heavily on the quality of the data.

- Unfair datasets are the root cause of algorithmic bias. However, it can be extraordinarily difficult to acquire more equitable data. Rebalancing datasets to mitigate bias will typically lead to lower overall performance, as rebalancing techniques may discard or deprioritize data to optimize for unbiased results.
- Many underlying sources of bias can be difficult to mitigate. An Amazon study found that even after removing direct causes of gender bias from a hiring algorithm (such as making the algorithm neutral to phrases like "women's chess club captain") the algorithm still found implicit male associations with phrases such as "executed" and "captured" on resumes.

Given access to underlying algorithms, it is substantially easier to prove discriminatory bias with an algorithm than it is with human-driven systems.

- Proving discrimination in hiring practices against a corporation typically requires a high bar of evidence.
  - According to the McDonnell Douglas framework for discrimination in the US, the accuser must prove that the

CONVERGENCE'S ANALYSIS

employer's reason for firing or reducing employment was a pretext for discrimination – often requiring a direct comparison to a comparable, non-discriminated party within the same organization.

- Cases involving larger cohorts of individuals (e.g. class action lawsuits) typically require more complex methods to prove discrimination. Potential approaches include creating statistically significant cohorts of "testers" designed to test the hiring practices of employers, victimization reports, or disparity studies on individuals with directly comparable work backgrounds.

- Meanwhile, algorithmic discrimination cases would likely produce demonstrable evidence primarily via access to the algorithm's API and a multivariate analysis by a statistician. Studies involving human participation (which have complicated ethical challenges and time-scales), complicated judicial processes, and the impact of random chance may be easier to avoid.

There are no established required practices or judicial precedents to evaluate the level of discriminatory bias across AI algorithms.

- Nearly all examples of bias discovered in AI algorithms have been identified by the efforts of independent teams of researchers unaffiliated with governmental legal or judicial systems. Because AI discrimination is only beginning to be legislated, there are few court cases and even fewer judicial rulings on how to prove algorithmic bias.
- As a result, it's currently very unclear to developers where the legal boundaries are between discrimination and predictive learning. An example: will resume evaluation algorithms need to scrub potentially gendered phrases from their dataset prior to training to ensure neutrality, such as participation in organizations like "Girls Who Code"? What about subtly biasing phrases, such as "NAACP", or "beauty pageant"?

CONVERGENCE'S ANALYSIS

It is likely that the required practices to evaluate discriminatory bias will be established in the judicial system.

- Judicial frameworks have typically been established over time via landmark or precedent-setting discrimination cases. For example, the McDonnell Douglas Burden-Shifting Framework and the Mixed Motive Framework are two separate judicial approaches to establish workplace discrimination. These developed independently to handle different forms of discrimination lawsuits.
- We expect that in the next five years, we'll begin to see class-action lawsuits against corporations running *high-risk* systems (as defined by the EU) that may be discriminatory. Accordingly, we'll expect to see the creation of one or more standardized frameworks for evaluating biased algorithms emerging from a US court.

AI Disclosures

What situations do disclosure requirements for AI systems cover?

The public and regulators have legal rights to understand goods and services. For example, food products must have clear nutritional labels; medications must disclose their side effects and contraindications; and machinery must come with safety instructions.

In the case of AI, these legally mandated disclosures can cover several topics, such as:

- **Clearly labeling AI-generated content:** This allows people to immediately recognize that the image (or text or audio etc) they're looking at was AI-generated. For example, the proposed AI Disclosure Act would require all generative AI content to include the text "Disclaimer: this output has been generated by artificial intelligence."
- **Watermarking content generated by AI**: This involves adding some detectable but not necessarily obvious mark. Watermarking has several purposes, for example letting us identify the provenance or source of AI-generated content.
- **Disclosure of training data:** Since models are trained on huge amounts of data, but this data isn't identifiable or reconstructable from the final model, some regulators require AI developers to disclose information about the data used to train models. For example, the EU AI Act requires AI developers to publicly disclose any copyrighted material used in their training data.
- **Notifying people that they're being processed by an AI:** For example, if video footage is analyzed by an AI to identify people's age, the EU AI Act requires those people to be informed.

How do labels and watermarks work for AI-generated content?

Labels and watermarks vary in design; some are subtle, some conspicuous; some easy to remove, some difficult. For example, Dall-E 2 images have 5 coloured squares in their bottom right corner, a conspicuous label that's easy to remove:

However, Dall-E 3 will add invisible watermarks to generated images, which are much harder to remove. Watermarking techniques are less visible than labels, and are evaluated on criteria such as perceptibility and robustness. A technique is considered *robust* if the resulting watermark resists both benign and malicious modifications; *semi-robust* if it resists benign modifications; and *fragile* if the watermark isn't detectable after any minor transformation. Note that fragile and semi-robust techniques are still useful, for example in detecting tampering.

Imperceptible watermarking methods might embed a signal in the "noise" of the image such that it isn't detectable to the human eye, and is difficult to fully remove, while still being clearly identifiable to a machine. This is part of steganography, the field of "representing information within another message or physical object".

For example, the Least Significant Bit (LSB) technique adjusts unimportant bits in data. For example, in the binary number 1001001, the leftmost "1" represents 2^6, while the rightmost "1" just represents 1, meaning it can be adjusted to carry part of a message with less disruption. LSB is relatively fragile, while other techniques like Discrete Cosine Transform (DCT) use Fourier transforms to subtly adjust images (and other data) at a more fundamental level, hiding signals in the higher frequency components of the image. These are more robust against simple attack techniques such as adding noise, compressing the image, or adding filters. Other popular techniques include DWT, SVD, and hybrids of multiple techniques.

There are also open-source technical standards such as C2PA that have been adopted by organizations like OpenAI. These types of standards allow good-faith actors to maintain a chain of causation and cryptographic signature on digital objects as they pass through various mediums (for example, as a photo moves from a camera, to editing, to publication, to tweets and retweets). However, these standards are relatively early-stage and lack key technological underpinnings and alignment for many layers of content provenance required to make them ubiquitous.

Text is much harder to watermark subtly, as the information content of text is relatively sensitive to small adjustments. Changing a few letters in a paragraph is more noticeable than changing many pixels in an image, for example. Watermarking can still be applied to metadata, and there are techniques derived from steganography that add hidden messages to text, though these can be disrupted and aren't under major consideration by legislators or AI labs.

Importantly, all these labeling and watermarking techniques can be embedded in the weights of generative AI models, for example in a final layer of a neural network, meaning it is possible to have robust but invisible signals in AI-generated content that, if interpreted correctly, could be used to identify what particular model generated a piece of work.

Watermarking also involves tradeoffs between robustness and detectability; robust watermarking techniques alter the content more fundamentally, which is easier to detect. This means robustness can also trade-off against security, as more obscure and undetectable watermarking are harder to extract information from, and thus more secure. For example, brain scans feature incredibly sensitive information, and so researchers have developed fragile but secure watermarking techniques for fMRI. To quote a thorough review of watermarking and steganography:

> It is tough to achieve a watermarking system that is simultaneously robust and secure.

Further, fragile watermarking standards could lead to false confidence, as any standards will inevitably incentivize powerful groups to break them.

Overall, modern digital watermarking techniques can be reasonably robust and difficult but not impossible to remove; watermarking may raise the barrier to entry of passing AI-generated content off as human-generated, and provide some tools for identifying the providence of AI-generated content (especially for images or audio content), but watermarking isn't perfect and hasn't been widely adopted.

What are current regulatory policies around disclosure requirements for AI systems?

The US

The Executive Order on AI states that Biden's administration will *"develop effective labeling and content provenance mechanisms, so that Americans are able to determine when content is generated using AI and when it is not."* In particular:

- **Section 4.5(a):** Requires the Secretary of Commerce to submit a report

identifying existing and developable standards and tools for authenticating content, tracking its provenance, and detecting and labeling AI-generated content.

- **Section 10.1(b)(viii)(C):** Requires the Director of OMB to issue guidance to government agencies that includes the specification of reasonable steps to watermark or otherwise label generative AI output.
- **Section 8(a):** Encourages independent regulatory agencies to emphasize requirements related to the transparency of AI models.

The AI Disclosure Act was proposed in 2023, though it has not passed the house or senate yet, instead being referred to the Subcommittee on Innovation, Data, and Commerce. If passed, the act would require any output generated by AI to include the text: *"Disclaimer: this output has been generated by artificial intelligence."*

China

China's **2022 rules for deep synthesis**, which addresses the online provision and use of deep fakes and similar technology, requires providers to watermark and conspicuously label deep fakes. The regulation also requires the notification and consent of any individual whose biometric information is edited (e.g. whose voice or face is edited or added to audio or visual media).

The **2023 Interim Measures for the Management of Generative AI Services**, which addresses public-facing generative AI in mainland China, requires content created by generative AI to be conspicuously labeled as such and digitally watermarked. Developers must also label the data they use in training AI clearly, and disclose the users and user groups of their services.

The EU

Article 52 of the EU AI Act lists the *transparency obligations* for AI developers. These largely relate to AI systems *"intended to directly interact with natural persons"*, where natural persons are individual people (excluding *legal persons*, which can include businesses). For concision, we will just call these "public-facing" AIs. Notably, the following requirements have exemptions for AI used to detect, prevent, investigate, or prosecute crimes (assuming other laws and rights are observed).

- **Article 52.1:** Requires developers to ensure users of public-facing AI are informed or obviously aware that they are interacting with an AI.
- **Article 52.1a:** Requires AI-generated content to be watermarked (with an exemption for AI assisting in standard editing or which doesn't substantially

alter input data).

- **Article 52.2:** Requires developers of AI that recognizes emotions or categorizes biometric data (e.g. distinguishing children from adults in video footage) to inform the people being processed.
- **Article 52.3:** Requires deep fakes to be labeled as AI-generated (with a partial exemption for use in art, satire, etc, in which case developers can disclose the existence of the deep fake less intrusively). AI-generated text designed to inform on matters of public interest must disclose that it's AI-generated, unless the text undergoes human review, and someone takes editorial responsibility.
- **Article 52b:** Requires developers of *general purpose AI with systemic risk* to notify the EU Commission within 2 weeks of meeting any of the following requirements defined in article 52a.1:
  - Possessing "high impact capabilities", as evaluated by appropriate technical tools.
  - By decision of the Commission, if they believe a general purpose AI has capabilities or impact equivalent to "high impact capabilities".
- **Article 52c:** Requires providers of GPAI to publish a summary of the content used for training the model, and 60f and 60k require developers to disclose any copyrighted material in their training data in their summary.

Convergence's Analysis

Unclear definitions of what constitutes an application of AI will lead to inconsistent disclosure requirements and enforcement.

- AI is becoming embedded in many creative tools, such as image-editing tools like Photoshop and GIMP. Among other functions, these can be used to "uncrop" images, generating additional content. AI is also important in procedurally generated video games and VR spaces.
- These uses of AI lead to gray areas and edge cases that aren't clearly covered by legislation, and individuals using these tools may not be able to tell whether they're using compliant or illegal tools.

CONVERGENCE'S ANALYSIS

- Current legal definitions are far from comprehensive enough to fully distinguish and legislate these overlapping use cases.

Mandatory labeling of AI-generated content is a lightweight but imperfect method to keep users informed and reduce the spread of misinformation and similar risks from generative AI.

- Labeling AI-generated text, images, video, and so on is a simple way to make users clearly understand that content is AI-generated. Further, it's not expensive or complex to add labeling mechanisms to generative AI.
- Labeling has extensive precedents in most legislations, such as food and medication labels.
  - While compliance can be high for such mandatory labeling, there's variance in efficacy. For example, the World Health Organization found that inadequate labeling of medication plays a role in non-adherence to medication prescriptions, and some studies have found that improving labeling improves health outcomes.
  - Further, compliance can be low, especially when violations by smaller organizations or individuals aren't actively addressed. For example, though many major websites are

Mandatory watermarking is a lightweight way to improve traceability and accountability for AI developers.

- Like labeling, watermarking is easy for developers to do, and invisible watermarks have the advantage of not interfering with the users' experience.
- If AI developers include watermarking in their generative AI models, these can be used to precisely identify which model was used to generate a piece of content. This is especially important when generative AI is used to generate harmful content, such as misinformation, deep fake porn, or other provocative material, as models should be trained not to produce such content. Watermarking allows us to find and address the root of the problem and hold the developers legally accountable.

CONVERGENCE'S ANALYSIS

Labels and watermarks can be disrupted or removed by motivated users, especially in text generation.

- Labels and watermarking involve adding information to content, and it is usually possible to manually (or even automatically) remove or disrupt this information.
- This means that it's unlikely any content platform could guarantee that AI-generated content is always clearly distinguishable to people.
- Despite the potential fragility of labeling and watermarking, they can still be important aspects of a larger, layered strategy, making it more difficult to produce misinformation, or for AI developers to avoid accountability.
  - In particular, societal education about AI will be a critical aspect of such a layered strategy.
- Research orgs such as Meta and DeepMind are researching more advanced methods of watermarking during AI development.

AI and Chemical, Biological, Radiological, & Nuclear Hazards

What are CBRN hazards? How could they be affected by AI models?

Humanity has developed technologies capable of mass destruction, and we need to be especially cautious about AI in relation to these technologies. These technologies and associated risks commonly fall into four main categories, collectively known as CBRN:

- **Chemical hazards:** Toxic chemical substances that can cause significant harm to people or the environment, such as chemical warfare agents or toxic industrial chemicals.
- **Biological hazards:** Toxins and infectious agents like bacteria, viruses, and other pathogens that can cause disease in humans, animals or plants.
- **Radiological hazards:** Radioactive materials that emit ionizing radiation which can harm human health, such as waste from nuclear power stations.
- **Nuclear hazards:** Materials related to nuclear fission or fusion that can release tremendous destructive energy, such as nuclear weapons and nuclear power plant accidents.

In this section, we'll briefly contextualize current and upcoming examples of each of these types of hazards in the context of AI technologies.

What are potential chemical hazards arising from the increase in AI capabilities?

In particular, a prominent concern of experts is the potential for AI to lower the barrier of entry for non-experts to generate CBRN harms. That is, AI could make it easier for malicious or naive actors to build dangerous weapons, such as chemical agents with deadly properties.

For example, pharmaceutical researchers use machine learning models to identify new therapeutic drugs. In this study, a deep learning model was trained on ~2,500 molecules and their antibiotic activity. When shown chemicals outside that training set, the model could predict whether they would function as antibiotics.

However, training a model to generate novel safe and harmless medications is very close to, if not equivalent to, training a model to generate chemical weapons. This is an example of the Waluigi Effect; the underlying model is simply learning to predict toxicity, and this can be used to rule out harmful

chemicals, or generate a list of them, ranked by deadliness. This was demonstrated by the Swiss Federal Institute for Nuclear, Biological, and Chemical Protection (see here for a non-paywalled summary). By telling the same model to generate harmful molecules, it generated a list of 40,000 such molecules in under 6 hours. These included deadly nerve agents such as VX, as well as previously undiscovered molecules that it ranked as more deadly than VX. To quote the researchers:

> This was unexpected because the datasets we used for training the AI did not include these nerve agents... By inverting the use of our machine-learning models, we had transformed our innocuous generative model from a helpful tool of medicine to a generator of likely deadly molecules.

As AI models become more deeply integrated into the process of developing chemicals used for industrial and medical purposes, it will become increasingly accessible for malicious parties to use these models for dangerous means.

What are biological hazards arising from the increase in AI capabilities?

In the near future, AI may lower the barrier of entry for malicious actors to generate pandemic-level biological hazards. This risk comes from both specialized AI trained for biological research and more generic AI, such as large language models.

Large language models (LLMs) have been identified by recent papers to lower barriers to misuse by enabling the weaponization of biological agents. In particular, this may occur from the increasing application of LLMs as biological design tools (BDTs), such as multimodal lab assistants and autonomous science tools. These BDTs make it easier and faster to conduct laboratory work, supporting the work of non-experts and expanding the capabilities of sophisticated actors. Such abilities may produce "pandemic pathogens substantially more devastating than anything seen to date and could enable forms of more predictable and targeted biological weapons". Further, the risks posed by LLMs and by custom AI trained for biological research can exacerbate each other by increasing the amount of harm an individual can do while providing access to those tools to a larger pool of individuals.

It's important to note these risks are still unlikely with today's cutting-edge LLMs, though this may not hold true for much longer. Two recent studies from RAND and OpenAI have found that current LLMs are *not* more prone to misuse than standard internet searches regarding biological and chemical weapons.

Another leading biological hazard of concern is *synthetic biology* – the genetic modification of individual cells or organisms, as well as the manufacture of synthetic DNA or RNA strands called *synthetic nucleic acids*.

This field poses a particularly urgent risk because existing infrastructure could theoretically be used by malicious actors to produce an extremely deadly pathogen, for example. Researchers are able to order custom DNA or RNA to be generated and mailed to them, a crucial step towards turning a theoretical pandemic-level design into an infectious reality. Currently, we urgently need mandatory *screening* of ordered material to ensure it won't enable pandemic-level threats.

Some researchers are developing tools specifically to **measure** and **reduce** the capacity of AI models to lower barriers of entry for CBRN weapons and hazards, with a particular focus on biological hazards with pandemic potential. For example, OpenAI is developing "an early warning system for LLM-aided biological threat creation", and a recent collaboration between several leading research organizations produced a practical policy proposal titled Towards Responsible Governance of Biological Design Tools. The Centre for AI Safety has also released the "Weapons of Mass Destruction Proxy", which measures how particular LLMs can lower the barrier of entry for CBRN hazards more broadly. Tools and proposals such as these, developed with expert knowledge of CBRN hazards and AI engineering, are likely to be a crucial complement to legislative and regulatory efforts.

For more context on these potential pandemic-level biological hazards, you can read:

- The White House Office of Science and Technology Policy's Framework for Nucleic Acid Synthesis Screening, published in April 2024 as directed by the Executive Order (an update to the ASPR's 2023 framework).
- The US Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential, published in May 2024.

What are radiological and nuclear hazards arising from the increase in AI capabilities?

A prominent and existential concern from many AI safety researchers is the risk of integrating AI technologies in the chain-of-command of nuclear weapons or nuclear power plants. As one example, it's been proposed that AI could be used to monitor and maintain the activity of nuclear power plants.

Elsewhere, The Atlantic cites the Soviet Union's Dead Hand as evidence that militaries could be tempted to use AI in the nuclear chain-of-command. Dead Hand is a system developed in 1985 that, if activated, would automatically launch a nuclear strike against the US if a command-and-control center stopped receiving communications from the Kremlin and detected radiation in Moscow's atmosphere (a system which may still be operational).

As the reasoning of AI technology is still poorly understood and AI models have unpredictable decision-making abilities, it's quite likely that such an integration may lead to unexpected and dangerous failure modes, which for nuclear technologies have catastrophic worst-case outcomes. As a result, many

researchers argue that the risk of loss-of-control means we shouldn't permit the usage of AI anywhere near nuclear technologies, such as decision-making regarding the nuclear launch codes or the storage and maintenance of nuclear weapons.

In proposals, some policymakers have pushed for banning AI in nuclear arms development, such as a proposed pact from a UK MP and Senator Mitt Romney's recent letter to the Senate AI working group. Romney's letter proposes a framework to mitigate extreme risks by requiring powerful AIs to be licensed if they're intended for chemical/bio-engineering or nuclear development. However, nothing binding has been passed into law. There have also been reports that the US and China are having discussions on limiting the use of AI in areas including nuclear weapons.

Current regulatory landscape

The US

The Executive Order on AI has several sections on CBRN hazards: various department secretaries are directed to implement plans, reports, and proposals analyzing CBRN risks and how to mitigate them, and Section 4.4 specifically focuses on analyzing biological weapon risks and how to reduce them in the short-term. In full:

- **Section 3(k):** The term "dual-use foundation model" is defined as AI that, among other criteria, exhibits or could be modified to exhibit high performance at tasks that pose serious risks, such as *substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use* CBRN weapons.
- **4.1(b):** The Secretary of Energy must coordinate with Sector Risk Management Agencies to develop and implement a plan for developing AI model evaluation tools and testbeds, and at a minimum, to *develop tools to evaluate AI capabilities to generate outputs that may represent nuclear, nonproliferation, biological, chemical, critical infrastructure, and energy-security threats or hazards and must develop model guardrails that reduce such risks.*
- **4.2(a)(i)(C):** The Secretary of Commerce must require companies developing dual-use foundation models to provide continuous information and reports on the results of any red-team testing related to *lowering the barrier to entry for the development, acquisition, and use of biological weapons by non-state actors.*
- **4.2(b)(i):** Any model that primarily uses biological sequence data and that was trained using at least 10^{23} FLOPs must comply with 4.2(a) until proper technical conditions are developed.

The following points are all part of **4.4**, which is devoted to *Reducing Risks at the Intersection of AI and CBRN Threats*, with *a particular focus on biological weapons:*

- **4.4(a)(i):** The Secretary of Homeland Security must *evaluate the potential for AI to be misused to enable the development or production of CBRN threats, while also considering the benefits and application of AI to counter these threats.*
  - **(A)** This will be done in consultation with experts in AI and CBRN issues from the DoE, private AI labs, academia, and third-party model evaluators, for the sole purpose of guarding against CBRN threats.
  - **(B)** The Secretary of Homeland Security will submit a report to the president describing progress, including *an assessment of the types of AI models that may present CBRN risks to the United States* and recommendations for regulating their training and use, *including requirements for safety evaluations and guardrails for mitigating potential threats to national security*
- **4.4(a)(ii):** The Secretary of Defense must enter a contract with the National Academies of Sciences, Engineering, and Medicine to conduct and submit a study that:
  - **(A**) assesses how AI can increase biosecurity risks, and makes recommendations on mitigating such risks;
  - **(B)** considers the national security implications of the use of data associated with pathogens and omics[1] studies that the government funds or owns for the training of generative AI, and makes recommendations on mitigating such risks;
  - **(C)** assesses how AI can be used to reduce biosecurity risks;
  - **(D)** considers additional concerns and opportunities at the intersection of AI and synthetic biology.
- **4.4(b):** To reduce the risk of misuse of synthetic nucleic acids[2]:
  - **(i)** The director of OSTP, in consultation with several secretaries, shall establish a framework *to encourage providers of synthetic nucleic acid sequences to implement comprehensive, scalable, and verifiable synthetic nucleic acid procurement screening mechanisms*. As part of this framework, the director shall:
    - **(A)** establish criteria for ongoing identification of biological sequences that could be pose a risk to national security; and
    - **(B)** determine standard methodologies *for conducting & verifying the performance of sequence synthesis procurement screening, including customer screening approaches to support due diligence with respect to managing security risks posed by purchasers of biological sequences identified in* (A) and *processes for the reporting of concerning activity.*
  - **(ii)** The secretary of commerce, acting through NIST and in coordination with others, shall initiate an effort to engage with industry and relevant

1 Defined in the Executive Order as *"biomolecules, including nucleic acids, proteins, and metabolites, that make up a cell or cellular system"*

2 Defined in the Executive Order in the following: *"The term "synthetic biology" means a field of science that involves redesigning organisms, or the biomolecules of organisms, at the genetic level to give them new characteristics. Synthetic nucleic acids are a type of biomolecule redesigned through synthetic-biology methods."*

stakeholders, informed by the framework of 4.4(b)(i), to develop and refine:

- **(A)** Specifications for effective nucleus synthesis procurement screening;
- **(B)** *Best practices, including security and access controls, for managing sequence-of-concern databases* to support screening
- **(C)** *technical implementation guides for effective screening; and*
- **(D)** *conformity-assessment best practices and mechanisms.*

- **(iii)** All agencies that fund life-sciences research shall establish as a requirement of funding that synthetic nucleic acid procurement is conducted through providers or manufacturers that adhere to the framework of 4.4(b)(i). *The Assistant to the President for National Security Affairs and Director of OSTP shall coordinate the process of reviewing such funding requirements to facilitate consistency in implementation*.
- **(iv)** To facilitate effective implementation of the measures of 4.4(b)(i)-(iii), the Secretary of Homeland Security shall, with consultation:
  - **(A)** *Develop a framework to conduct structured evaluation and stress testing of nucleic acid synthesis procurement screening […];*
  - **(B)** *Submit an annual report […] on any results of the activities conducted pursuant to* 4.4(b)(iv)(A), including recommendations on how to strengthen procurement screening.

China

China's three most important AI regulations do not contain any specific provisions for CBRN hazards.

The EU

The EU AI Act does not contain any specific provisions for CBRN hazards, though article (60m) on the category of "general purpose AI that could pose systemic risks" includes the following mention of CBRN: *"international approaches have so far identified the need to devote attention to risks from […] chemical, biological, radiological, and nuclear risks, such as the ways in which barriers to entry can be lowered, including for weapons development, design acquisition, or use"*.

Convergence's Analysis

Mitigating catastrophic risks from AI-enabled CBRN hazards should be a top global priority.

- CBRN hazards present arguably the shortest and most immediate path for AI to lead to catastrophic harm.
- AI is demonstrably already capable of lowering the barrier to entry of generating biological and chemical weapons. This lowering is likely to get more dramatic in the near future.
- When paired with the existing and under-regulated infrastructure for biology labs generating custom genetic code on demand, this could plausibly lead to the accidental or deliberate release of an unprecedented pandemic pathogen within the next decade.

Despite this, current and near-future legislation and regulation regarding AI and CBRN hazards is wholly insufficient given the scale of potential risks.

- The EU and China currently have no specific binding requirements regarding the development of AI models capable of enabling the development of CBRN weapons.
- The US Executive Order directs several agencies to initiate important studies and reports on the intersection of AI and CBRN weapons, particularly focusing on biosecurity risks. However, these are largely non-binding and exploratory, leaving plenty of ambiguity in precisely what regulations might follow the directive. More concrete regulation focused on catastrophic and existential risks, such as mandatory safety and security requirements for dual-use models, is needed.

Effective regulation of CBRN and AI will require close collaboration between AI experts, domain experts, and policymakers.

- The development of legislation regarding CBRN weapons requires an unusually high level of specialized technical expertise, and so regulators will need to work closely with leading researchers in

CONVERGENCE'S ANALYSIS

the fields of AI, biology, chemistry, and cybersecurity to identify and mitigate key risks.

- It is difficult to impossible to develop effective model evaluations without substantial input from both AI experts and domain experts. Long-term, close collaboration between these parties is a critical aspect of identifying key CBRN risks.
- Several teams of researchers have been developing tools and proposals tailored to CBRN-related AI risk (though none have yet been adopted), such as:
  - OpenAI's early warning system for LLM-aided biological threat creation;
  - The Centre for AI Safety's Weapons of Mass Destruction Proxy;
  - Towards Responsible Governance of Biological Design Tools, a collaboration between leading AI, governance, and risk research organizations.

AI governance in other high-risk domains like cybersecurity and the military has major implications for CBRN risks.

- Multiple militaries around the world possess stockpiles of chemical, biological, and nuclear weapons, and nuclear power plants and biocontainment facilities can also present CBRN hazards. If advanced AI is trained for cybersecurity attacks, these stockpiles and other hazardous systems could be targeted with devastating outcomes.
- The increasing adoption of AI by militaries - such as the first confirmed deployment of fully autonomous military drones and the several hundred US military AI projects disclosed by the Pentagon - leads many to fear that AI will become increasingly involved in the decision-making and chain-of-command of CBRN weapons. The involvement of AI here will require exceptional value alignment, as even slight misalignment in goals and values between human and AI operators could lead to catastrophic harm.

Conclusion

The current state of AI regulations in 2024 demonstrates that while governments are making progress in developing their AI governance frameworks, these policies are still in very early stages, with significant research yet to be done and regulations to be developed. Many foundational tools that lay the groundwork for future AI policies have yet to be fully implemented, such as model registries, incident reporting, or AI chip registries. Model evaluations, a key component of determining risk from AI systems, are not yet at a point where they can be effectively mandated by regulation. Extensive precedents from other domains suggest the upcoming progression of certain domains of AI governance in the next 1-5 years, such as discrimination requirements, disclosure requirements, and CBRN policies. Even in domains where initial binding policies have been established, it remains to be seen how these policies will be implemented and how legal and technical challenges to these policies will play out.

Despite the nascent stage of AI regulations, there is already clear evidence regarding the long-term direction of key governments regarding AI technologies. The US, EU, and China have taken different approaches to AI governance, prioritizing various aspects based on their political, economic, and social contexts. While the EU has focused on protecting citizens' rights and passed encompassing horizontal legislation, China has taken an iterative, domain-specific approach emphasizing social control and alignment with party values. Meanwhile, the US has primarily sought to maintain its technological edge and slow China's progress in the field, though it has recently begun to request the participation of various executive agencies in developing AI policy.

This high-level report merely scratches the surface when it comes to evaluating the potential impacts of these regulations. The decisions made by these leading governments in the aforementioned domains will have far-reaching consequences, affecting economies, societies, and industries globally.

In this report, we've provided lightweight analyses with fairly impartial observations about the progression of AI regulations, but there are many outstanding questions and in-depth research evaluations that this report has yet to address. Our organization is currently conducting further research on some of these neglected evaluations.

As AI continues to advance at an unprecedented pace, it is crucial for policymakers, and researchers to collaborate closely in developing comprehensive, adaptable, and effective regulatory frameworks. By providing an overview of the current regulatory landscape for AI technologies, we intend that this report will serve as an introductory foundation for those seeking to navigate the evolving world of AI governance.