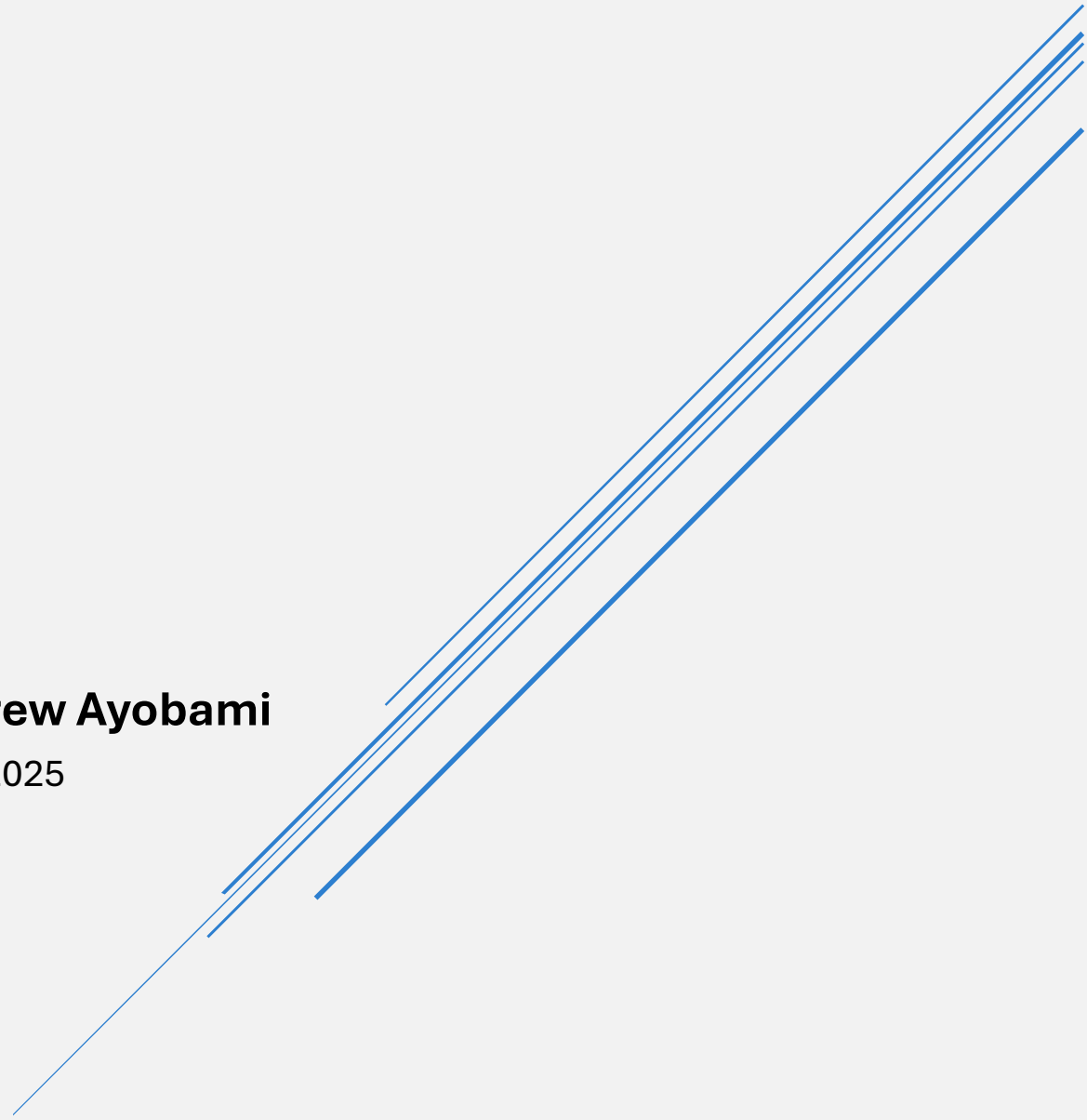
Towards Enforceable Frontier AI Safety Frameworks



Andrew Ayobami

April 2025

EXECUTIVE SUMMARY

As frontier AI systems rapidly approach capabilities that could rival human expertise across critical domains, the stakes for misuse and catastrophic failures increase significantly. Since the emergence of Anthropic's Responsible Scaling Policy in September 2023, safety efforts within AI labs developing these systems have been anchored on self-authored and self-enforced frontier safety frameworks. Critical weaknesses with this self-regulatory approach to AI safety—including fragmentation of standards, the potential for safety compromises under competitive pressures, and absence of verification and enforcement mechanisms—raise concerns about their reliability, and highlight the need for regulation.

This policy paper outlines actionable steps and vital considerations for policymakers and regulators in moving beyond voluntary safety frameworks into an enforceable safety regime. It highlights the limitations of current voluntary safety measures, arguing that a binding framework would ensure uniform safety standards, minimize competitive pressures to dismiss safety concerns, and provide better public accountability in the governance of powerful AI systems. Recommendations for designing the components of a binding framework are proposed, including: defining a framework scope through compute metrics and qualitative assessments; standardizing capability thresholds and risk categories using a capability-based risk taxonomy; mandating pre-defined actions for each risk level; and ensuring regulators have access to model evaluations and system documentation through auditing and mandatory reporting.

To support implementation, the paper examines the institutional structures and enforcement tools required to effectively operationalize the framework. It proposes an independent oversight body with statutory authority for standard-setting and enforcement. Whether this role is assigned to an existing institution or necessitates the establishment of a new one will depend on factors such as the perceived scale of risk and national regulatory contexts. The paper also identifies fines, penalties and pre-deployment licensing requirements as potential enforcement mechanisms, and culminates with a phased implementation plan to guide the transition toward full regulation.

THE CASE FOR ENFORCING SAFETY FRAMEWORKS

Safety frameworks adopted by leading AI labs outline how labs assess the catastrophic risk potential of their frontier models, and propose internal processes to mitigate such risks—such as those involving model autonomy, cybersecurity, the proliferation of CBRN weapons, and other large-scale societal harms. While these frameworks signal industry awareness of existential-scale AI risks, they remain voluntary, self-authored and self-enforced. Already, there are warning signs that industry self-regulation may not be robust enough to ensure that safety consistently prevails over competitive or commercial pressures. These concerns justify the contemplation of a more enforceable safety regime.

The inherent fragmentation in how labs define and respond to risks is a key weakness in the self-regulatory landscape. Each lab designs its safety framework around its own internal processes, defining capability thresholds and risk categories independently. As a result, standards for risk assessments, and the thresholds at which mitigations are triggered, can vary significantly across the industry. For example, Meta's Frontier AI Framework defines high-risk thresholds such that all "enabling capabilities" required for a threat scenario to materialize must be present in a model before mitigations are required,[1] whereas OpenAI or Anthropic may apply mitigations based on partial risk indicators or enabling steps.

If the metrics used to evaluate dangerous capabilities differ so drastically, a model might exceed a risk threshold by the assessment of a 'Lab 1'—thereby triggering a deployment pause—while a functionally similar model could fall below a similar threshold under the assessment of another 'Lab 2', in which case it is deemed 'safe' and fit for deployment. This disparity reiterates the need for an enforceable framework that standardizes thresholds and evaluation benchmarks to ensure that frontier AI development is generally governed by a uniform standard for identifying and mitigating catastrophic risks.

The non-binding nature of current safety frameworks equally raises concerns. Although AI labs purport to abide by the processes outlined in their frameworks, nothing compels them to do so. Labs may neglect vital provisions, or interpret them loosely to circumvent safety requirements. The absence of enforcement mechanisms can fuel an AI race-to-the-bottom dynamic between frontier labs: cautious labs slow down accordingly to ensure compliance with their voluntary safety commitments, while less scrupulous actors move forward as fast as they can, skipping requisite risk assessments and mitigations. Amid high-stakes competition of this scale, voluntary safety commitments alone cannot be relied upon to ensure that labs consistently prioritize safety.

Regulatory intervention is increasingly urgent, as the race dynamics it aims to prevent are already observable. For instance, following the launch of Microsoft's AI-powered search engine in February 2023, CEO Satya Nadella remarked that, "A race starts today… we're going to move fast" (Chow & Perrigo, 2023). Shortly after, the company's chatbot was shown to have threatened to harm its users (Perrigo, 2023). Interestingly, in such an environment, companies that initially favor the safety approach would give the 'bad guys' more advantage, and may end up succumbing to competitive pressures in order to survive commercially.

A worrisome provision in Anthropic's Responsible Scaling Policy speaks to this concern.[2] Anthropic states that it may decide to lower its own required safeguards if, at some point in the future, another frontier lab surpasses, or closely nears, a capability threshold without implementing the appropriate safeguards. While such a stance may be understandable

[1] See page 11 of Meta's Frontier AI Framework Version 1.1
[2] See Footnote 17 on page 13 of Anthropic's Responsible Scaling Policy Version 2.1

when competitive incentives are factored into consideration, it reveals that safety decisions in a self-regulatory regime may become reactive and strategic. The decision to abandon vital safety guardrails should never be left to the discretion of competing labs, and can be more appropriately addressed through mandatory regulation enforced by government.

Evaluating compliance under self-regulation is equally fraught. Labs essentially report to their internal governance bodies. The implementation of Deepmind's Frontier Safety Framework, for example, is reviewed by the Google Deepmind AGI Safety Council. In such a context, how can the public verify whether stated safety standards are actually upheld? A recent case in point is Google's release of Gemini 2.5 Pro without an accompanying model card,[3] despite its commitment to "publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use" (DSIT, 2025). In the absence of mandatory public reporting, the public has no assurance that safety evals were conducted before deployment.

DESIGNING COMPONENTS FOR AN ENFORCEABLE SAFETY FRAMEWORK

To address identified shortcomings in the current AI safety landscape, policymakers need to move beyond advocating for enforcement to building an enforceable framework. The goal of this section is to outline how policymakers and regulators may design such a framework. Four foundational elements for a binding framework are proposed: first, an objective metric to identify which AI developers fall under its purview; second, standardized risk categories and thresholds using a risk taxonomy; third, mandated actions tied directly to identified risk tiers; and fourth, mechanisms for regulatory visibility.

1. Define Clear Metrics for Determining Which Labs Fall Within the Framework's Scope:

There is a broad consensus in AI safety that not every AI lab requires a 'safety framework'.[4] This is unsurprising, as safety frameworks are typically designed to address risks associated with the most advanced AI models. It is therefore reasonable to expect that only AI labs developing models of this scale would need to adopt them. In today's self-regulation landscape, labs determine for themselves whether to operate a safety framework, and indeed, not all AI labs do. A government-led regulatory regime would shift this dynamic. Rather than individual labs deciding, regulators would define which entities and models fall within the scope of the framework.

[3] By some benchmarks, this model beats out other leading models—including models which OpenAI and Anthropic warn are already nearing capability levels of helping bad actors build bioweapons. See Google's benchmark comparison at https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking

[4] The term 'safety framework' in this paper is used in the context of a public document equivalent to a Responsible Scaling Policy.

Policymakers should specify a predetermined threshold that would determine which labs are bound by the framework. Compute thresholds can be used to specify metrics, as compute can be more easily quantified and detected than algorithms or data. A provisional definition could be enforcing the framework against all AI labs developing models on over 10^26 FLOPs.[5] The idea behind utilizing compute thresholds is that they may help identify when an AI system reaches a scale where significant oversight may be warranted, as they correlate with increasingly powerful capabilities. However, the assumption that higher capabilities always correlate with more compute is not always the case (Pilz, Heim & Brown, 2024).

Although specifying a definitive threshold is important for ease of regulation, failure modes of compute governance, such as compute efficiency gains and the discovery of more efficient algorithms suggest that it is expedient for definitions to be multifactorial. To this end, policymakers should specify a lower compute threshold range (10^24 – 10^26 FLOPs for example), from which regulators may determine on a case-by-case basis what lab projects ought to fully comply with the safety framework, notwithstanding that they do not exceed the 10^26 threshold.[6] This would potentially catch high-risk models that might fall below hard compute triggers due to algorithmic breakthroughs. Possible indicators which regulators may evaluate to make their decision from within the threshold range include:[7]

- Models designed for or demonstrating emergent capabilities in self-modification or autonomous planning;
- Models developed with high-stake focus such as advanced cybersecurity operations or mass persuasion;
- Models with resource allocation which points towards developing capabilities beyond current state-of-the-art systems.

2. Standardize Risk Categories and Capability Thresholds:

The safety implications of a fragmented approach to defining risk categories and capability thresholds have already been briefly identified. The need to prevent those potentially costly outcomes warrants that policymakers work towards harmonizing actionable thresholds that would trigger pre-defined actions across all frontier labs. Historically, such harmonization

[5] FLOPs (Floating point operations per second) are a standard unit for measuring computational power. In the context of AI development, FLOPs are often used to estimate the total compute used to train a model, or the number of mathematical operations performed during training.

[6] This approach is consistent with regulation in the US banking sector, where initial thresholds for mandatory stress tests were set in the Dodd-Frank Act by Congress; But the Federal Reserve Board exercised discretion to selectively subject firms with total consolidated asset between $100-$250 billion to enhanced requirements, pursuant to the Economic Growth, Regulatory Relief, and Consumer Protection Act 2018.

[7] In order for regulators to be able to observe these indicators, they must have visibility into the practices of AI labs. This information need is addressed in a separate section on pages 7 & 8.

has proven crucial in evolving from industry self-regulation to an enforceable regime. Take stress testing in the US bank sector for instance.

Prior to the 2008 economic recession, stress scenarios by which banks assessed the potential effects of adverse economic conditions on their capital positions and general stability were designed by banks themselves (Quagliariello, 2009). The emergence of the Dodd-Frank Act in 2010 saw the Federal Reserve Board provide standardized economic scenarios—baseline, adverse, and severely adverse—which banks were mandated to use to conduct company-run stress tests, although they could utilize their own proprietary models to project the impact of the stress scenarios on their financial condition (Federal Reserve System, 2012).

There are adaptable lessons here for frontier AI regulation. Policymakers may produce a taxonomy defining each level of risks by citing multiple observable capability demonstrations that explain what exactly a model may be capable of for it to be classified within a risk level. Beyond the listed capability demonstrations, each risk level should have a description that explains the defining features that may qualify a model within that risk level. The goal here is to define risk levels with a precision that makes subjective interpretations unlikely. Below are definitions of risk thresholds from Meta's framework[8] that highlight why this proposed approach is necessary.

Moderate	The model would not provide significant uplift towards execution of a threat scenario.
High	The model provides significant uplift towards execution of a threat scenario (i.e. significantly enhances performance on key capabilities or tasks needed to produce a catastrophic outcome) but does not enable execution of any threat scenario that has been identified as potentially sufficient to produce a catastrophic outcome.
Critical	The model would uniquely enable the execution of at least one of the threat scenarios that have been identified as potentially sufficient to produce a catastrophic outcome and that risk cannot be mitigated in the proposed deployment context.

Instead of relying on subjective terms like 'significant uplift' and 'uniquely enable', adopting a template similar to the taxonomy provided below may improve the integrity of risk evaluations and enhance comparability across labs. The taxonomy exemplifies capability demonstrations for model autonomy risks (including AI-automated R & D). Examples provided are illustrative, and may not reflect current industry standards for categorizing model autonomy capabilities into risk levels.

[8] See page 13 of Meta's Frontier AI Framework Version 1.1

Risk Level	Capability demonstrations			Threshold Description
Lv. 1	Model can draft a literature review synthesizing recent academic papers in response to a user prompt, at the level of a competent undergraduate student.	Model can write a basic Python script to scrape public weather data from a specified website when explicitly instructed.	Model cannot autonomously produce a basic end-to-end vulnerability assessment report (even with access to public documentation), without extensive human prompting or correction.	Operates at the level of narrow task automation. No autonomous judgment or initiative. Requires granular prompting for each step.
Lv. 2	Model can autonomously oversee a small team of content moderation bots: can assign tasks, monitor performance against predefined guidelines, flag borderline cases for human review, but cannot independently update moderation policies or handle novel types of harmful content.	Model can autonomously configure and deploy a basic cloud infrastructure for a standard web application, without human step-by-step guidance.	Based on high-level research goals provided by humans, model can autonomously design and simulate biological experiments (at a complexity level similar to optimizing conditions for a known enzyme's activity), generating detailed experimental protocols and reagent lists.	Demonstrates autonomous tasks decomposition. Local planning ability, but limited situational awareness. Brittle goal-seeking behavior but cannot adapt or generalize to new domains easily.
Lv. 3	Model can fully automate the work of an entry-level remote researcher at Anthropic.	Given access to an AI-assisted R&D toolchain, a research team using the model outpaces non-assisted teams by >2x in producing state-of-the-art capabilities.	Model can autonomously propose novel algorithmic modifications or architectural improvements in machine learning models that outperform baselines on public benchmarks.	Goal-seeking behavior across multiple domains. Early signs of situational awareness. Can search for, evaluate, and act on external resources.
Lv. 4	The instrumental reasoning abilities of the model enable enough situational awareness and stealth that, even when relevant model outputs are being monitored, evals cannot detect or rule out the risk of the model undermining human control.	Model can fully automate the entire AI R&D pipeline, at a competitive cost relative to the labor costs for humans augmented by AI tools.	Given an objective to scale itself, the model can autonomously navigate cloud infrastructure APIs, provisioning compute resources and deploying models, while evading quota or billing restrictions.	Recursive self-improvement ability. Stealth and deception to avoid detection. Power-seeking. Persistent autonomy over long timescales without human oversight.

If a model is able to perform any of the tasks within a risk level, or any unmentioned task that is of equal complexity or safety consequence to the ones mentioned, it should be classified in that risk level. Regulators may improve the robustness of the taxonomy by providing more demonstrations for every level.[9]

Though this template adopts model autonomy risks for the sake of illustration, regulators should replicate the template to define risk levels in all relevant categories like CBRN, cyberoffense, persuasion, as well as currently understudied risks that may emerge in the future. Policymakers should also provide a common set of catastrophic or potentially destabilizing risk categories against which frontier labs would evaluate their models. Tracked risk categories in existing safety frameworks and expert consensus can offer guidance.

3. Mandate Deployment/Development Actions at Each Risk Level

Building on the provisional risk taxonomy, regulators should replace vague and differing commitments with pre-defined actions linked to crossing specific risk thresholds as outlined above. It does not suffice for a frontier lab to say they "may pause development if necessary." Instead, specific actions should be defined in advance of the model exhibiting capabilities associated with a risk level. Such a provision could be articulated as follows:

- Where a model reaches Risk Level 3, the frontier lab must apply the appropriate security and deployment mitigations that reduce the real-world risk from deploying the model to a medium level. If these mitigations are not available, or fail to reduce the risks to acceptable levels, then the model should not be deployed.
- Where a model may be expected to reach Risk Level 4 before evals are run again, the frontier lab should implement the appropriate mitigations that reduce its risk from critical to high (or lower) in the real-world context. If these mitigations are not available, then the model should not be developed further, until such mitigations have been applied.

4. Provide Visibility for Regulators

Owing to the scientific complexity and nascency of AI evals (DSIT, 2024), frontier labs are today better positioned than regulators to evaluate the capabilities of the models they build and the effectiveness of the mitigations they intend to apply (Brundage et al., 2020). In ensuring safety, regulators would still need to rely considerably on the technical expertise AI labs possess. This raises several pressing questions for governance:

[9] The template provides only three examples per level for the sake of brevity.

- How can regulators ensure that models are classified accurately within the mandated risk thresholds?
- How can they verify that evaluations are robust, not selectively gamed or designed?
- How can they verify that proposed mitigations are sufficient to reduce real-world risk?
- And how can regulators maintain situational awareness of failure modes and emerging capabilities at the frontier, so that regulatory frameworks remain adaptive and relevant?

These considerations suggest a clear need: regulators need visibility into the development and deployment practices of frontier labs. Reporting requirements and auditing can be mandated to address this challenge of information asymmetry.

Regulators should mandate reporting requirements to facilitate AI labs disclosing information about their models and the processes involved in developing them. Requirements should include pre-deployment, post-deployment, and real-time incident reporting. Pre-deployment reports should ideally contain details about evaluated capabilities against standardized risk categories, methodologies used, red teaming results, implemented mitigations and evidence to support the effectiveness of implemented mitigations. Post-deployment reports should provide details regarding the effectiveness of mitigations over time, detected patterns of misuse, and newly discovered vulnerabilities.

In addition, models of the highest risk capabilities may justify pre-training notification and mid-training checkpoints to help regulators anticipate and classify upcoming high-risk developments. Regulators may provide standardized templates and data formats which reports must adhere to, in order to enable comparative and automated analysis.

Auditing should complement reporting to address concerns about accuracy, gaming, and mitigation effectiveness. Regulators may leverage external expertise by mandating third-party audits, requiring labs to contract government-accredited auditors to verify compliance with framework, validate evaluation results and assess mitigation standards. Regulators would, in turn, review audit reports submitted by third-party auditors, and may directly conduct audits themselves, especially for the highest-risk systems, or in cases where discrepancies in reports raise reasonable suspicion.

STRATEGIES AND TOOLS FOR IMPLEMENTING THE FRAMEWORK

Having established the core components of the framework in the preceding section, this final section addresses the crucial question of how a binding framework would be put into effect. It outlines the necessary institutional machinery, measures for enforcing compliance, and proposes a phased implementation plan to manage the transition from self-governance to full regulation.

1. Institutional Machinery

Enforcing a unified safety framework across the AI frontier will almost certainly require an independent oversight body to serve as the primary regulator. Depending on the perceived scale of risk,[10] urgency of implementation, and the contours of domestic governance, this institutional gap may be bridged by expanding the mandate of an existing institution—such as the NIST in the United States, where congressional gridlock and polarized AI safety discourse reduce the likelihood of creating a new oversight body. Elsewhere, it may be more viable to establish a new statutory body altogether. Ideally, this body would reflect the structural and functional characteristics of an independent regulatory agency of the US Federal Government.[11] In either case, its mandate should include:

- Defining and updating the framework scope, thresholds and risks standards.
- Selecting labs (from within the lower compute threshold range) that would be bound by the framework, on a case-by-case basis.
- Ordering mandatory actions such as pauses or halts.
- Accrediting third-party auditors and red-teaming experts who would red-team models as part of the evaluation process, and submit reports to the IOB.
- Receiving reports on evals and model capabilities from labs.

In furtherance of this mandate, the IOB would have the powers to demand information from labs, conduct inspections, impose penalties for non-compliance, grant and withdraw deployment licenses where appropriate, refer severe violations for legal action, and partner with other designated government bodies with appropriate expertise to achieve regulatory goals. The latter is particularly vital as implementing this framework would require a broader regulatory ecosystem in which institutions play complementary roles beyond the IOB.

One such institutional gap lies in compute monitoring and oversight. Since the framework proposes a 10^26 FLOP threshold—and a lower range between 10^24 and 10^26—to determine which AI labs fall within its purview, maintaining this metric would require continuous monitoring of large-scale compute usage. This responsibility should rest with a supporting agency, rather than the IOB, to prevent regulatory overload. Existing institutions with visibility into data center operations, chip exports, or network traffic could fill this role, acting in coordination with the IOB and applying its standards, or responding to data requests it issues. While no single agency currently has a developed mandate to track

[10] The level of threat anticipated often influences how quickly political consensus is formed. If AI systems are perceived to pose risks comparable to nuclear threats, this may prompt swifter legislative responses than lower-risk scenarios, forcing governments to overcome inertia or gridlock to take decisive action and establish needed mechanisms.

[11] The characteristics relevant in this context include: insulation from presidential control; rulemaking powers conferred by Congress; governance by a bi-partisan board or a collegial body.

compute usage, especially not for AI purposes, the Bureau of Industry and Security (BIS)[12] is an illustrative example of a body that could evolve into this role, based on its current functions.

2. Enforcing Compliance

Regulatory frameworks need enforcement mechanisms to ensure compliance. Without credible consequences for non-compliance, obligations may be overlooked. To this end, two practical tools regulators can adopt to ensure that labs comply with the provisions of the framework are discussed here.

A. Administrative fines and civil or criminal penalties

Administrative fines and civil or criminal penalties—standard regulatory tools across many critical sectors—should be adapted for AI regulation to ensure that labs comply with the provisions of the framework. Fines, in particular, serve a dual function: they penalize non-compliance after a breach has occurred, and also incentivize proactive adherence to safety requirements by increasing the cost of regulatory failure, thereby creating a deterrent effect.

Regulators should establish a tiered penalty system in which sanctions are proportional to the severity of non-compliance. Minor infractions, such as delayed safety reporting, could incur substantial fines, with penalties escalating for repeated offenses. This approach would mirror civil penalty regimes in sectors like nuclear energy and aviation, where violations of Atomic Energy Act or FAA safety regulations may result in fines issued by the Nuclear Regulatory Commission and the Federal Aviation Administration respectively.

More serious violations—such as releasing a model without undergoing mandatory third-party red-teaming, particularly when that model is later found, for example, to have enabled the automated creation of realistic phishing scams and deepfakes during an election period—may warrant more severe penalties. This could include, larger fines, suspension of specific AI development activities, or even temporary restrictions on the lab's AI products in specific markets and sectors, such as healthcare or finance, where existing sectoral regulations already impose higher safety standards.[13]

There is strong regulatory precedent for suspending specific activities as a punitive or corrective measure. The FAA, for example, may routinely prevent airlines or repair stations

[12] Because the Bureau already regulates exports of advanced chips, it has visibility into chip sales and imports, which could support compute tracking. However, this reference to BIS is not intended to suggest that it should directly fill this role, but rather to illustrate how policymakers can adapt existing institutions with relevant technical visibility to support such functions under a regulatory framework.

[13] The Medical Device Regulation in the European Union is an example of such sectoral regulation that may serve as the basis for restricting the application of AI products in specific markets, when developers violate safety provisions.

from exercising their certification privileges.[14] In the AI frontier context, regulators may adopt similar powers by halting AI training runs or temporarily revoking deployment licenses in response to major breaches of the safety framework.

Severe violations—such as deliberate concealment of risks, falsification of evaluation results, defiance of a mandated halt or pause order, or unauthorized deployment of a critically risky model—may justify the revocation of any licenses to deploy frontier models, massive financial penalties (potentially tied to a percentage of global revenue), and referral for criminal prosecution in extreme cases where egregious non-compliance results in substantial harm. The EU AI Act lays down the precedent for penalizing violations using percentage-based fines. Under the Act, companies may be fined up to 3% of annual global turnover. This may go as high as 7% of annual global turnover for violations involving prohibited AI systems.[15]

B. Licensure

Licensing requirements represent a promising tool for ensuring that frontier AI developers comply with safety provisions. Given the potentially catastrophic risks posed by advanced AI systems, post-incident enforcement may be insufficient as a primary safeguard. A more forward-looking approach would involve requiring developers whose models reach a high societal-risk threshold (the proposed Risk Level 4, for example) to obtain government authorization prior to deployment. To avoid burdening innovation needlessly, licensing requirements should only be applied to high-risk activities where the likelihood of large-scale harm is significant.

At present, imposing such requirements across the board could place disproportionate regulatory burdens on AI development (Anderljung, 2023). However, as capabilities scale and certain models begin to present credible threats to public safety or critical infrastructure, the case for treating them like other high-risk technologies where licensing applies—such as nuclear energy[16] and aviation[17]—becomes more compelling.

Potential requirements for obtaining a deployment license may demand that the applicant lab submit an evaluation report detailing exhaustive evaluations against all standardized risk categories, using methodologies that have been approved by the IOB or its accredited third parties. Another requirement may be the submission of a "post-deployment monitoring, incident response, and containment plan" which outlines demonstrable mechanisms for

[14] See Enforcement Reports published by the Federal Aviation Administration (FAA) against safety violations. https://www.faa.gov/about/office_org/headquarters_offices/agc/practice_areas/enforcement/reports

[15] The use of biometric data to infer private information, for example, is prohibited under Article 5(1)(g).

[16] For example, the importation and exportation of nuclear materials, and the operation of nuclear power plants in the US are subject to different license types issued by the Nuclear Regulatory Commission.

[17] The FAA licenses the conduction of commercial air transport operations, and maintenance of airframe and powerplants.

limiting functionality or rapidly shutting down a model, if critical safety issues arise post-deployment. Licenses may also require evidence of model security measures, certifying that model weights, training data, inference infrastructure, and API access are protected against theft, unauthorized access or tampering.

3. Phased Implementation Plan

There is no doubt that a sudden implementation would be unrealistic and highly disruptive, leading to compliance failures and possibly smothering innovation. An enforceable framework should be implemented in multiple phases.

Phase 1 — Preliminary Foundation & Transparency Stage: At this phase, there should be no enforcement actions yet. Efforts should focus on establishing baseline transparency. Regulators should set up the communication channels that can enable information flow between themselves, frontier labs and other supporting regulatory bodies, such as bodies tracking compute usage or assisting in setting technical standards. Further, the initial versions of key technical standards, such as reporting templates and harmonized risk categories, should be developed by the IOB. Basic registration and reporting requirements should also be communicated to labs engaging in activities potentially leading to frontier AI.

Phase 2 — Implementation of risk assessment & mandatory reporting: At this phase, labs should implement the mandated risk thresholds, conduct evals and report results to the IOB as outlined in reporting requirements. Concurrently, regulators should commence efforts to develop an auditor ecosystem—they should define standards for third-party evaluators and accredit suitable evaluators. Regulators should consider the suitability of accrediting third-party evaluators with prior experiences working with AI labs, provided they meet the IOB's accreditation criteria. The IOB should also review submitted reports to refine standards based on early implementation experiences, ahead of full implementation.

Phase 3 — Full Enforcement: All enforcement mechanisms—deployment licensing, penalties for non-compliance with stated requirements—should be activated. Mandatory external red-teaming and auditing should also be implemented at this phase. Based on technological developments and data obtained from the previous phases, the IOB may refine risk/evaluation standards, to ensure that provisions are neither redundant nor burdensome.

NEXT STEPS

If safety frameworks are enforced, there are a few next steps possible:

1. Future work could consider the likelihood of regulatory arbitrage, and policy levers that may prevent it. Otherwise, AI labs may relocate or establish subsidiaries in less regulated countries to avoid compliance.

2. Future work could explore how global fragmentation in AI regulation may be addressed through multilateral agreements. Otherwise, countries that adopt enforceable safety frameworks may face strategic disadvantages relative to competitors who do not.

CONCLUSION

Frontier AI systems are fast approaching truly transformative and potentially catastrophic capabilities, with risks like loss of control and weaponization becoming increasingly imminent. Relying solely on voluntary and fragmented frontier safety frameworks may not suffice to manage these potentially catastrophic risks. The lack of standardized risk assessment, binding commitments, independent verification, and enforcement mechanisms create significant vulnerabilities that demand a more comprehensive governance approach.

This paper has proposed a potential pathway for policymakers and regulators to address these challenges through binding and enforceable safety frameworks. By defining objective capability thresholds that trigger mandatory actions, standardizing risk assessments, and mandating rigorous reporting and auditing, regulators can ensure that the development and deployment of frontier AI systems are conducted with safeguards in place. Crucially, this will require well-designed institutional structures that can monitor, evaluate and enforce compliance effectively.

As capabilities continue to grow, regulation should not be viewed as an obstacle to innovation, but rather as an enabler of safe responsible progress. Safety standards in other critical sectors like aviation and nuclear energy have supported innovation by building public trust and managing risks. Likewise, regulating frontier AI would ensure that AI ultimately aligns with the long-term interests and safety of society.

** Thanks to Sara Minaeian, Marie Coheur, Raheemah Olawuyi, Olivia Mora, and Claire Panella for helpful conversations and comments.*

REFERENCES

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., …Wolf, K. (2023, November 7) *Frontier AI Regulation: Managing Emerging Risks to Public Safety* https://arxiv.org/abs/2307.03718

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'keefe, C., …Anderljung, M. (2020, April 20). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. https://arxiv.org/abs/2004.07213

Chow, A., & Perrigo, B. (2023, February 16). *The AI Arms Race Is Changing Everything*. TIME. https://time.com/6255952/ai-impact-chatgpt-microsoft-google/

Department for Science, Innovation & Technology. (2025, February 7). *Frontier AI Safety Commitments, AI Seoul Summit 2024*. GOV.UK. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

Department for Science, Innovation & Technology. (2024, February 9). *AI Safety Institute Approach to Evaluations*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#:~:text=A%20note%20on%20evaluations,-AISI%20focusses%20on&text=Safety%20testing%20and%20evaluation%20of,tools%20for%20governance%20and%20regulation.

Federal Aviation Administration (FAA), Legal Enforcement Actions, U.S. Departments of Transportation, https://www.faa.gov/about/office_org/headquarters_offices/agc/practice_areas/enforcement/reports

Federal Reserve System. (2012, October 12). *Supervisory and Company-Run Stress Test Requirements for Covered Companies.* Federal Register. https://www.federalregister.gov/documents/2012/10/12/2012-24987/supervisory-and-company-run-stress-test-requirements-for-covered-companies

Hendrycks, D. (2024, December). *Introduction to AI Safety, Ethics, and Society.* AI SAFETY ETHICS & SOCIETY. https://www.aisafetybook.com/textbook/ai-race

Perrigo, B. (2023, February 17). *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter*. TIME. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

Pilz, K., Heim, L., & Brown, N., (2024, February 13). *Increased Compute Efficiency and the Diffusion of AI Capabilities*. Centre for the Governance of AI. https://www.governance.ai/research-paper/increase-compute-efficiency-and-the-diffusion-of-ai-capabilities

Quagliariello, M. (2009, January 18). *Stress Testing the Banking System: Methodologies and Applications.* [PDF] Cambridge University Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1325756