
Too Big to Fail, Too Catastrophic to Insure: Making the Labs Pay for AI's Risk¹

Fernando Moreno
Effective Altruism Brazil
(Altruísmo Eficaz Brasil)

With
Apart Research

Abstract

The multiple labs developing AI models compete with one another under the pressure of an “arms race” (Scott Alexander’s Moloch problem). This creates a unilateral incentive to cut corners on safety protocols, amplifying risks. One of the arguments frequently raised against an eventual pause by the labs is that the arms race also takes place between nation-states (the “new cold war” between the US and China), making the geopolitical problem more complex given the difficulty of credibly enforcing agreements not to develop the technology.

Considering this, this article proposes a stopgap solution: shared financial liability among labs. If a lab causes damage that exceeds its ability to pay (even driving it into bankruptcy), the remaining labs in the industry are jointly and severally liable for the excess, with each firm’s share proportional to its own risk profile (measured by compute, parameter count, revenue, or a combination of several metrics).

We should be clear here that the proposal does not solve existential risk — the scenario in which humanity is extinguished or permanently loses control over its future. It is argued, more modestly, that there is a band of Global Catastrophic Risk (GCR) — events of terrible but non-terminal harm — for which financial liability instruments have real power to discipline behavior, precisely because they presuppose the existence of an “after” in which compensation still means something. Nevertheless, even this mechanism could, to some extent, help reduce existential risk, if it makes the labs more careful overall.

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

*In his classic text, *Meditations on Moloch*, Scott Alexander analyzes how agents competing with one another can produce terrible outcomes that no one wants, even when each individual agent acts rationally in pursuit of its own interests. Alexander uses the god Moloch as a metaphor for this perverse coordinating force: competitive dynamics (arms races, unregulated capitalism, tax competition, the “race to the bottom”) push all participants to sacrifice values they collectively cherish — health, beauty, free time, ethics — because whoever does not sacrifice is outcompeted by whoever does. The text argues that civilization and culture (the “gardens”) are, in part, precarious ways of escaping these traps through coordination, but that these solutions are fragile and can collapse.*

This dynamic can materialize in the labs developing frontier AI models. Seeking not to fall behind in the race, they may, more or less deliberately, ignore due safety precautions. In response, proposals arise for the nation-state to regulate these companies, in some way disciplining, controlling, or even pausing the development of AI.

However, many argue that even if this solved coordination among labs within a country, it does not solve coordination between countries, because nation-states have national-security incentives that make a unilateral pause even more politically costly.

Diplomats can always be called upon to resolve such conflicts, leading to an agreement to pause the development of AI capabilities. But any agreement to pause or limit development would depend on the ability to verify, credibly, that the signatories are honoring the commitment.

AI lacks the physical checkpoints that make treaties such as nuclear non-proliferation (despite its difficulties) enforceable. In the nuclear case, enriched uranium is a rare physical resource, with signatures detectable by inspection (such as those conducted by the IAEA in Iran while Obama’s agreement lasted), which makes it possible to identify violations even without full state transparency. AI models, on the other hand, are ultimately knowledge and code: once researchers and algorithms exist, it is extremely difficult to prevent a country or lab from continuing development in secret, since there is no “rare physical material” equivalent to enriched uranium that could serve to detect a breach of the agreement. The only thing close to that would be verifying whether the production or purchase of cutting-edge chips and large-scale data centers is taking place, which are (poorly) physically traceable, for example by satellite imagery. Moreover, cutting-edge chips today have production concentrated in a few corporations: TSMC, ASML, and Nvidia. But even this control is fragile: it does not prevent theoretical research, algorithmic optimizations that reduce the need for computing power, or distributed development at a smaller scale.

Our main contributions are:

- *A stopgap policy proposal that shifts the AI safety coordination problem from the hard-to-verify regulatory-diplomatic terrain to that of civil liability, sidestepping the international enforcement problem.*

- *A concrete three-layer shared-liability design for frontier labs, modeled on the 1957 Price-Anderson Act for the nuclear industry, with each firm’s share proportional to its own risk profile (compute, parameter count, revenue, or a composite index).*
- *An argument for how the mechanism simultaneously attacks three failures identified in the literature — the perverse race dynamic, the insolvency problem, and the regulatory pacing problem.*

2. Related Work

The framing of the problem draws directly on Scott Alexander’s Meditations on Moloch, which models how competition between rational agents drives a collective “race to the bottom” on shared values. Applied to frontier AI, this captures the unilateral incentive each lab faces to cut corners on safety so as not to fall behind.

On the policy side, the closest analog and direct inspiration is the Price-Anderson Act, the 1957 US law governing the civil liability of the nuclear industry. Price-Anderson channels liability through layered coverage — mandatory primary insurance per operator, a mutualized secondary pool funded by retrospective contributions from all operators, and an ultimate State backstop. This proposal reproduces that layered logic for AI labs.

It also relates to the broader literature on the limits of nuclear non-proliferation treaties as an enforcement model. That literature stresses the role of rare, physically traceable inputs (enriched uranium, detectable by IAEA inspection) in making violations observable. The contribution here is to argue that AI lacks an equivalent physical checkpoint — since models are ultimately knowledge and code — and that civil liability is therefore a more robust disciplining mechanism than verification-dependent treaties.

3. Methods

To escape this trap without relying on hardly verifiable international enforcement, this article shifts the problem from the regulatory-diplomatic terrain to that of civil liability. The proposal structures the financial liability of frontier labs in layers, in a design that is not unprecedented: it reproduces the logic of the Price-Anderson Act, the 1957 US law that governs the civil liability of the nuclear industry.

The mechanism operates in three layers.

In the first layer — mandatory own liability — each lab is required to purchase insurance or maintain reserves (self-insurance) up to a defined individual cap, a band it covers on its own before any mutualization.

In the second layer — the mutualized pool — when the damage caused by a lab exceeds its first layer (including driving it into bankruptcy), the other members of the industry are called upon retrospectively, each bearing a share proportional to its own risk profile (which can be measured by compute, parameter count, revenue, or a composite index, for example). This can be either a pre-paid fund immobilized in cash, similar to credit guarantee funds designed to prevent bank runs, or

an event-triggered contribution commitment, in which each operator commits to paying a given amount per reactor should another cause an accident that exceeds its primary coverage. In the third layer the State will inevitably foot the bill: above the aggregate capacity of the entire industry, the decision falls to the State. It is in this layer that the title of this article meets its limit: we can imagine a band of damage too large for any private aggregation. However, given recent valuation estimates of the most advanced AI labs, surpassing the incredible mark of one trillion dollars each, one can think that this layer would really be called upon only in very significant global catastrophes.

4. Results

The mechanism directly attacks three failures identified in the literature: the perverse race dynamic (the first-mover advantage incentivizes cutting corners on safety), the “insolvency problem” (no single firm has the capital to cover catastrophic damage, generating under-deterrence), and the regulatory “pacing problem” (traditional regulation is slower than the development of the technology).

The perverse race dynamic

Shared liability inverts the reward structure: cutting corners ceases to be an externality dumped onto a diffuse society that does not react, and becomes a cost imposed on identifiable and powerful competitors, who will have to pay for someone else’s recklessness — and who, symmetrically, can impose on you the cost of yours. A direct financial interest is created in the safety of what is produced by rivals. The labs come to have an incentive to monitor one another, push for sector-wide standards, and audit each other — the equivalent of what, among individuals, reputation does to discipline behavior even in the absence of personal trust. Moreover, here the race to the bottom could invert into a race to the top: effectively demonstrating superior safety could reduce the premium to be paid.

The insolvency problem

The economic theory of deterrence through civil liability assumes that the party causing the damage is able to pay for it. When the potential damage exceeds the firm’s assets, liability is in fact truncated at the value of bankruptcy: the firm internalizes the social cost only up to the limit of its own assets and treats all of the excess — precisely the catastrophic band — as someone else’s problem. The result is that the more severe the scenario, the less it weighs on the firm’s decision, because above the point of insolvency all outcomes are equivalent (it has already gone bankrupt either way). Shared financial liability among labs corrects this by expanding the assets that back each lab’s liability: the payment cap becomes the industry aggregate, causing a much larger fraction of the real social cost to return to each agent’s calculation.

The regulatory pacing problem

Traditional command-and-control regulation is slow by design: public consultation and litigation are measured in years, while AI capabilities advance in months or less. Many specific rules are

born outdated. Civil liability has two properties that make it structurally more agile. It is technologically neutral: it does not specify what is safe, it merely makes the firm pay if it causes damage. And it is self-updating: the price of the share reacts continuously to new risk information, without requiring a regulator to rewrite rules at every capability jump. The price mechanism does the work that bureaucracy cannot do at the insane pace of current development.

5. Discussion and Limitations

By relocating the AI coordination problem from international treaty-making to civil liability, the proposal turns each lab into an interested monitor of its rivals' safety and makes recklessness a cost borne by identifiable competitors rather than a diffuse externality. For AI safety more broadly, the implication is that financial instruments can discipline behavior in the Global Catastrophic Risk band without requiring the credible, verifiable international enforcement that AI's code-and-knowledge nature makes so difficult.

Limitations

Proving which model, from which lab, caused diffuse and emergent damage may be unfeasible under the usual standards of causal nexus.

Even if the measure does not pause AI, it may lead to a slowdown of its development. The country or countries that adopted it could fall behind in the arms race compared to countries that do not adopt a similar measure. A solution to this could be the requirement that all countries wishing to have access to the most modern chips adopt the same measure or other similar bargain solutions. The proposal does not solve existential risk. It is argued, more modestly, that there is a band of Global Catastrophic Risk for which financial liability instruments have real power to discipline behavior, precisely because they presuppose the existence of an "after" in which compensation still means something. Nevertheless, even this mechanism may to some extent help reduce existential risk, if it makes the labs more careful in general.

Future Work

Natural next steps include calibrating the individual liability caps and the risk-weighting formula (compute, parameter count, revenue, or a composite index); designing the causal-attribution standards needed to make claims tractable; and analyzing the international bargain — for example, conditioning access to cutting-edge chips on adoption of an equivalent liability regime — that would prevent unilateral adopters from falling behind in the race.

6. Conclusion

The arms race among frontier AI labs, compounded by an arms race among nation-states, makes verification-dependent pauses and treaties fragile, because AI — unlike enriched uranium — leaves no rare physical signature to inspect. As a stopgap, this article proposes shared financial liability among labs, structured in three layers on the model of the Price-Anderson Act: mandatory

own coverage, a mutualized industry pool with shares proportional to each firm’s risk profile, and an ultimate State backstop.

The mechanism does not pretend to solve existential risk. Within the band of Global Catastrophic Risk, however, it attacks the perverse race dynamic, the insolvency problem, and the regulatory pacing problem at once — turning rivals into mutual monitors, expanding the assets that back each lab’s liability, and letting a price signal update faster than any regulator could. Even short of eliminating existential risk, by making labs more careful in general it may help reduce it.

Code and Data

This is a policy analysis project and does not involve code, datasets, or experimental artifacts.

Author Contributions (optional)

Fernando Moreno conceived the proposal and wrote the manuscript.

References

- *Alexander, Scott. “Meditations on Moloch.” Slate Star Codex.*
<https://www.slatestarcodexabridged.com/Meditations-On-Moloch> (en); Portuguese translation:
<https://altruismoeficazbrasil.substack.com/p/meditacoes-sobre-moloch> (pt-br)
- *Price-Anderson Nuclear Industries Indemnity Act, Pub. L. 85–256 (1957), United States.*

Appendix (optional)

Not applicable.

LLM Usage Statement

Parts of this text were developed with the assistance of Claude (Anthropic). The content, arguments, and conclusions are the responsibility of the author. All claims have been reviewed and verified by the author.