

INSURANCE-GRADE DATA INFRASTRUCTURE FOR FRONTIER AI GOVERNANCE

Subramanyam Sahoo

Independent Researcher

subramanyamsahoow@gmail.com

ABSTRACT

International AI agreements that aim to constrain frontier capability growth or reduce catastrophic deployment risk face a recurring implementation barrier: the actors who control critical capability are often non state entities, including frontier labs, cloud providers, universities, and model deployers. Agreements written solely as state to state commitments tend to fail when compliance depends on domestic enforcement that is slow, uneven, or strategically reversible. This paper proposes a governance architecture that brings non state actors into scope through insurance and liability, and it formalizes the central data problem that blocks such an approach today. We model frontier model incidents as a heavy tailed loss process with strong cross actor correlation and severe under reporting. Under these conditions, conventional actuarial pricing fails, voluntary insurance excludes precisely the tail risk of interest, and liability rules can unintentionally suppress disclosure. We propose a minimal international regime in which access to high capability model deployment markets and high end compute supply chains is conditioned on carrying qualified coverage, and coverage is conditioned on submitting standardized, privacy preserving risk data. We specify the data fields needed for underwriting, define a reporting and auditing protocol that can be adopted by companies and research institutions across borders, and provide a simulation study that quantifies how missingness and correlation dominate premium uncertainty and capital load. We conclude with policy recommendations for treaty text, for domestic implementing legislation, and for regulatory modifications that enable a functioning AI insurance sector without creating perverse incentives against transparency.

Here is the repository link: [TAIG-hackathon-Apart on GitHub](#)

1 INTRODUCTION

Frontier AI risk governance is increasingly constrained by a mismatch between formal authority and operational control. States can sign agreements, but critical capability is concentrated in private firms, research institutions, and cross border cloud platforms. Even when a state endorses an agreement, it may lack the ability or the will to monitor training and deployment in detail, especially when the underlying systems evolve rapidly and the relevant evidence is proprietary. Many governance proposals rely on technical verification of compute or on export controls. These instruments are valuable but incomplete. They do not directly address the incentives of non state actors that decide whether to train a new model, how to evaluate it, how to deploy it, and how to respond to near misses. They also do not provide a durable compliance interface that remains effective when supply chains and corporate structures adapt Bengio et al. (2025b).

Insurance and liability frameworks are a candidate interface. They are already used in aviation, maritime shipping, nuclear operations, construction, and cyber security to force adoption of safety controls when direct government inspection is expensive. Yet AI insurance remains shallow. Most policies cover narrow risks such as professional liability or intellectual property disputes, and they often exclude systemic harm, misuse, and model behavior risks. A primary reason is the absence of insurance grade data: insurers cannot price what they cannot observe, and AI developers cannot share sensitive information without risking competitive loss or legal exposure.

This paper addresses three goals within one coherent technical governance argument.

Data goal: formalize the data required to price frontier model risk and show why existing public incident information is insufficient under heavy tail and correlated shock conditions. **Inclusion goal:** design an international agreement mechanism that binds non state actors through market and contract leverage rather than relying entirely on state enforcement. **Insurance goal:** propose a liability and insurance structure that makes transparency and control adoption privately rational while maintaining strong incentives against reckless scaling and negligent deployment.

Our core claim is that the main bottleneck for insurance based frontier AI governance is neither willingness nor ideology. It is a tractable but neglected measurement problem: without standardized risk exposure and incident data with known reporting properties, insurers will either withdraw or offer coverage that does not discipline frontier behavior.

2 RELATED WORK

Our proposal connects three literatures.

First, it builds on international governance mechanisms that achieve compliance by conditioning market access on standardized documentation and audits, including trade linked environmental agreements and controlled goods regimes. These mechanisms motivate using qualified coverage and procurement requirements to include non state actors in practice.

Second, it draws on cyber insurance and safety critical industry experience showing that insurers can induce adoption of specific controls, but that correlated and systemic losses can cause insurers to retreat into exclusions without shared data and risk pooling. This motivates modeling correlated shocks and focusing on insurance grade data rather than only on public incident narratives.

Third, it relates to emerging AI governance proposals that emphasize reporting, evaluations, and post deployment monitoring as enforceable interfaces for high capability systems. Our contribution is to frame these interfaces as underwriting relevant data infrastructure that can be embedded into insurance contracts and liability safe harbors, thereby creating a cross border inclusion channel for companies and research institutions Wasil (2024); Tweed et al. (2025).

3 PROBLEM STATEMENT

3.1 NON STATE ACTOR INCLUSION IS STRUCTURALLY HARD.

International agreements are usually written for states. Frontier AI capability is increasingly controlled by entities that are not states and that operate across multiple jurisdictions. This creates at least five failure modes.

Jurisdictional arbitrage: a firm can move training or deployment to a region with weaker enforcement. Corporate fragmentation: capability can be distributed across subsidiaries, contractors, and research partners. Supply chain mixing: compute can be sourced through intermediaries that are difficult to monitor. Rapid iteration: even if a model version is audited, post training modifications can change risk profile on short timescales. Ambiguous responsibility: when harm occurs, it is difficult to assign causal responsibility between model developer, deployer, and end user.

A governance system that does not include non state actors as first class subjects tends to turn into a patchwork of uneven domestic controls.

3.2 INSURANCE AND LIABILITY ARE ATTRACTIVE BUT CURRENTLY INFEASIBLE AT FRONTIER SCALE

Insurance can be a strong incentive mechanism because it is enforced through contract terms, premiums, exclusions, and required controls. Liability can create a floor of responsibility and define safe harbors that encourage adoption of best practices. However, insurers face two technical challenges that are especially acute for frontier AI.

Heavy tailed losses: rare catastrophic events dominate expected loss and capital requirements. Correlated risk: failures are not independent across policyholders. A single vulnerability class, a single misuse technique, or a single widely deployed model update can generate simultaneous claims.

A third practical challenge amplifies both issues.

Missingness and selection: incidents are under reported, near misses are rarely disclosed, and observed incidents are biased toward public failures. Without a structured data layer, insurers either cannot price or they price conservatively and exclude the risk categories that governance most needs to influence.

3.3 THE DATA PROBLEM AT THE CENTER

To underwrite responsibly, an insurer needs more than counts of publicly known failures. They need exposure and control data. Exposure data includes what is deployed, where, to whom, with what tool access, with what distribution channels, and with what update cadence. Control data includes what evaluations were done, what monitoring exists, what incident response procedures exist, and what access restrictions exist. The core data problem is to collect a minimal set of exposure, control, and incident fields in a way that is verifiable, privacy preserving, and compatible with commercial confidentiality.

4 PRECEDENTS FROM ENVIRONMENTAL AND CYBER GOVERNANCE

This section extracts design patterns rather than analogies.

Environmental governance Several environmental agreements succeed because they combine state commitments with trade and market mechanisms that pressure private actors to comply. *Montreal Protocol style controls*: restrictions on trade in controlled substances created incentives for manufacturers and distributors to comply, even beyond signatory states, because access to markets depended on compliance. *CITES style permit systems*: trade in controlled goods requires documentation and audits, which pushes compliance onto private supply chains.

Design lesson: a treaty can create compliance pull by conditioning market access on meeting a standardized reporting and verifiable control interface.

Cyber governance and cyber insurance Cyber insurance partially succeeded at driving adoption of certain controls, such as multi factor authentication and backup regimes, because insurers can refuse coverage or increase premiums when controls are missing. However, cyber insurance also experienced failures. *Systemic correlated risk*: ransomware campaigns created correlated losses. Insurers reacted by restricting coverage, raising deductibles, and demanding stronger controls. *Data scarcity*: under reporting and inconsistent incident taxonomies limited actuarial precision.

Design lesson: a reporting and disclosure regime is required for insurance to remain disciplining rather than degenerating into exclusion and retreat.

Liability and reporting cultures Aviation safety improved through mandatory reporting, protected reporting channels, and standardized incident investigation practices. In high reliability industries, near miss reporting is treated as a resource for prevention rather than as a reputational liability.

Design lesson: liability can be structured to encourage reporting through safe harbors, while still punishing negligence and reckless operation.

5 MODEL OF FRONTIER AI LOSS, CORRELATION, AND MISSINGNESS

We present a deliberately simple insurer-relevant model that captures the dominant obstacles for underwriting frontier AI risk: (i) heavy-tailed severities, (ii) correlated systemic shocks, and (iii) selective reporting / missingness.

5.1 LOSS PROCESS

Let organizations be indexed by $i \in \{1, \dots, N\}$. Each organization has an exposure level e_i (scaled deployment volume / capability) and a control-quality covariate c_i (safety controls). The annual number of claim-generating incidents for organization i is modelled as

$$K_i \sim \text{Poisson}(\lambda_i), \quad (1)$$

$$\log \lambda_i = \alpha + \beta_e \log e_i + \beta_c c_i + u_i, \quad (2)$$

where u_i is an organization-specific random effect (for example $u_i \sim \mathcal{N}(0, \tau^2)$). Incident severities are heavy-tailed; for example we may assume a Pareto tail

$$S_{ij} \stackrel{\text{iid}}{\sim} \text{Pareto}(s_{\min}, \xi), \quad (3)$$

with tail parameter ξ (smaller ξ indicates a heavier tail). As an alternative, one can take a log-normal model with large scale:

$$\log S_{ij} \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma \text{ large}. \quad (4)$$

Total annual loss for organization i is the compound sum

$$L_i = \sum_{j=1}^{K_i} S_{ij}, \quad (5)$$

a compound Poisson (or compound log-normal/Pareto) random variable with large variance under heavy tails.

5.2 CORRELATED SYSTEMIC SHOCKS

To capture events that simultaneously affect many actors, introduce a binary common-shock indicator $Z \in \{0, 1\}$ with $\Pr(Z = 1) = p$. Conditional on Z the incident rate and severity tail change. A parsimonious specification is

$$\log \lambda_i(Z) = \log \lambda_i + \delta Z, \quad (6)$$

$$\xi(Z) = \xi + \Delta_\xi Z, \quad (7)$$

where $\delta > 0$ increases incident rates when $Z = 1$, and Δ_ξ (typically $\Delta_\xi < 0$ when we mean “worsen the tail”) changes the Pareto tail parameter. Thus, conditional on Z the distribution of L_i is a compound Poisson with parameters $(\lambda_i(Z), S_{ij} \sim \text{Pareto}(s_{\min}, \xi(Z)))$. The marginal distribution of portfolio losses is a mixture over Z , which concentrates capital requirements on the triple (p, δ, Δ_ξ) .

5.3 REPORTING AND SELECTION (MISSINGNESS)

Observed claims / publicly visible incidents are an incomplete, selectively observed sample. Let $R_{ij} \in \{0, 1\}$ indicate whether incident j at organization i is observed in the data available to underwriters. Model the reporting probability as a logistic (sigmoid) function of severity and controls:

$$\Pr(R_{ij} = 1 \mid S_{ij}, c_i) = \sigma(\gamma_0 + \gamma_s \log S_{ij} + \gamma_c c_i), \quad \sigma(x) \equiv \frac{1}{1 + e^{-x}}. \quad (8)$$

This captures the qualitative facts that (i) larger/rarer events are more likely to become visible ($\gamma_s > 0$) and (ii) organizations with stronger controls may detect/report events differently (sign of γ_c determined by institutional incentives). Because reporting depends on severity, the observed sample is biased: naive empirical tails (built from observed S_{ij} with $R_{ij} = 1$) will typically overstate tail severity while understating true event frequency.

5.4 IMPLICATION FOR UNDERWRITING

Under this model the insurer requires the posterior distribution of L_i given the observed (selectively reported) data,

$$p(L_i \mid \{(S_{ij}, R_{ij} = 1)\}_j, e_i, c_i), \quad (9)$$

Table 1: Minimum underwriting relevant reporting fields for frontier model coverage. The regimes in Section 8.1 correspond to progressively richer subsets of this table.

Category	Fields and underwriting purpose
Exposure	Model class and version identifier; deployment channel class; tool access class; coarse usage volume bins; update cadence. These fields support risk segmentation and allow pricing to respond to changes in exposure surface.
Controls	Evaluation coverage and cadence; monitoring and logging coverage; access restrictions; incident response capability. These fields support premium gradients and define auditable criteria for qualified coverage and liability safe harbor.
Incidents	Hazard class; severity band; discovery channel; remediation action. These fields support pricing updates, audit triggers, and correlated risk monitoring.

which depends critically on the reporting mechanism and on whether systemic shock Z is accounted for. With heavy tails and correlated shocks the variance (and tail risk) of L_i is large; with selective reporting the posterior is weakly identified without auxiliary exposure/control data or credible priors for $(p, \delta, \Delta_\xi, \gamma)$. In short: the data regime (what is observed and why) is not a minor administrative detail but fundamentally determines whether insurance can function as an effective governance instrument for frontier AI risk.

6 MINIMAL INSURANCE GRADE DATA REQUIREMENTS FOR FRONTIER AI

This section focuses on feasibility. The objective is not a comprehensive model card. It is a minimal set of fields that allow pricing, auditing, and dispute resolution.

Exposure fields: Model class and version identifier, including post training stage identifiers. Deployment channel class, including API access, open weights distribution, and internal agent deployment. Tool access class, including whether the system can call external tools, browse, execute code, or access sensitive internal resources. User base and volume proxies, such as request counts, active users, or internal usage hours, reported in coarse bins. Update cadence, including frequency of weight updates and policy updates.

Control fields: Evaluation suite coverage, including what categories are tested, at what frequency, and with what decision thresholds for release. Monitoring and logging coverage, including whether misuse detection exists and whether incident response playbooks exist. Access controls, including rate limits, identity verification level for high risk use, and restrictions for high risk domains. Incident response capability, including time to mitigate, rollback capability, and escalation process.

Incident fields: Incident type class, including misuse enablement, data leakage, autonomy failures, and tool misuse. Severity measure, reported as bounded ordinal plus optional monetary estimate under confidentiality. Discovery channel, including internal monitoring, external report, regulator notice, or media. Remediation action, including patch, policy change, rollback, user notification, and third party coordination.

Verifiability and confidentiality: To include non state actors, data reporting must not require revealing training data or proprietary weights. A minimal mechanism is a combination of cryptographic commitments to model version identifiers and evaluation attestations, aggregated reporting for volume measures in coarse bins, protected reporting to insurers and regulators with safe harbor provisions.

This paper does not assume perfect cryptographic attestation of all fields. It assumes that incentives and audit risk can sustain partial verifiability, similar to other compliance regimes Bengio et al. (2025a).

Link from reporting fields to the stress test regimes. The stress tests in Section 8 model three information regimes that correspond to concrete subsets of Table 1. Regime A approximates public

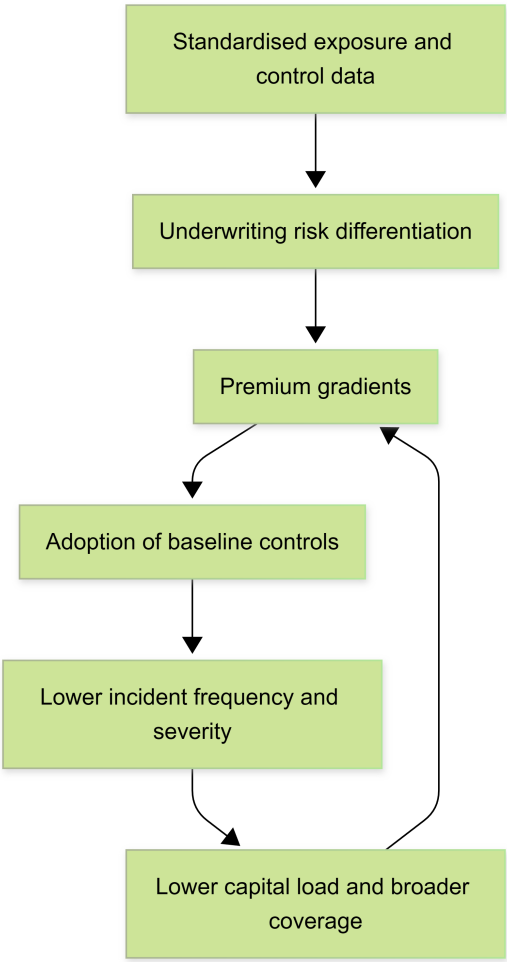


Figure 1: Control incentive gradient concept map.

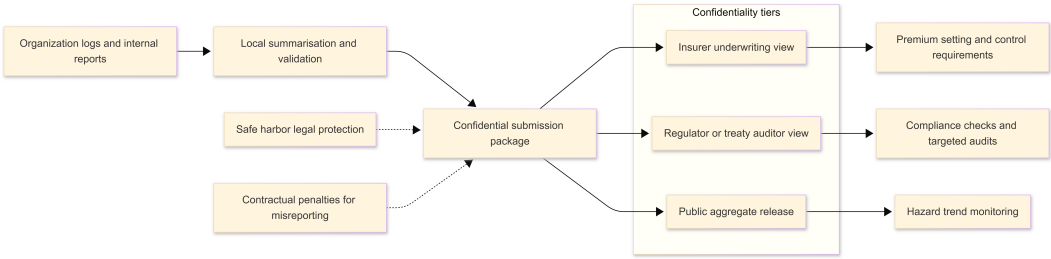


Figure 2: Data flow and confidentiality tiers

visibility only, where the insurer observes only a small and selected fraction of incidents and does not observe standardized exposure or control fields. Regime B corresponds to confidential incident reporting using the incident fields in Table 1, but with limited ability to condition premiums on standardized exposure and control covariates. Regime C corresponds to confidential incident reporting plus standardized exposure and control reporting as in Table 1, enabling underwriting to condition on auditable practices rather than on outcomes alone.

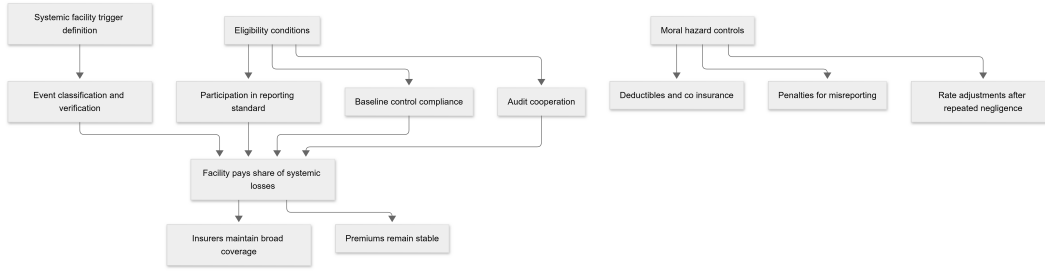


Figure 3: Systemic risk facility design

7 GOVERNANCE ARCHITECTURE: INSURANCE AND LIABILITY AS AN INCLUSION MECHANISM

We propose a mechanism that can be written into an international agreement and implemented domestically.

7.1 FRONTIER MODEL COVERAGE REQUIREMENT

A signatory state commits to require that any entity that deploys or trains models above a specified capability or scale threshold must carry qualified coverage to operate in regulated markets and to access public procurement. Coverage qualification includes: policy terms that cover a defined class of model behavior harms and misuse harms, with bounded exclusions, adoption of baseline controls and mandatory participation in standardized risk data reporting.

This pulls non state actors into the regime because operating without coverage becomes commercially costly.

7.2 LIABILITY RULE WITH REPORTING SAFE HARBOR

Liability can destroy reporting incentives if disclosure increases expected damages. We propose a two layer structure: Negligence based liability with a rebuttable presumption of negligence if baseline controls and reporting are absent. A reporting safe harbor that reduces damages or limits punitive multipliers when an organization demonstrates timely reporting, prompt remediation, and adherence to baseline controls. This mirrors the idea of protected near miss reporting in safety critical domains while maintaining strong penalties for reckless scaling or repeated negligence.

7.3 INTERNATIONAL COORDINATION WITHOUT DIRECT GLOBAL REGULATOR

The agreement creates an interoperability layer rather than a single global supervisor. A standard for reporting fields and audit procedures. Mutual recognition of qualified insurance policies and qualified auditors. A mechanism for cross border information sharing on systemic hazard classes, without requiring disclosure of proprietary details. This structure is compatible with the reality that enforcement capacity and legal regimes differ across states.

8 STRESS TESTS AND EXCLUSION BOUNDARIES FOR INSURANCE-MEDIATED GOVERNANCE

This section evaluates whether an insurance- and liability-based inclusion mechanism remains viable under plausible regimes of systemic correlation and heavy-tailed losses. The objective is not to forecast real-world claim totals but to identify parameter regions in which the private insurance market rationally retreats into exclusions, and to identify what reporting and risk-sharing mechanisms keep coverage meaningful. Results are produced by a synthetic loss simulator described in Appendix A that is designed to capture three properties that dominate frontier model governance: (i) severe losses exhibit heavy tails; (ii) failures can be correlated across organizations through shared

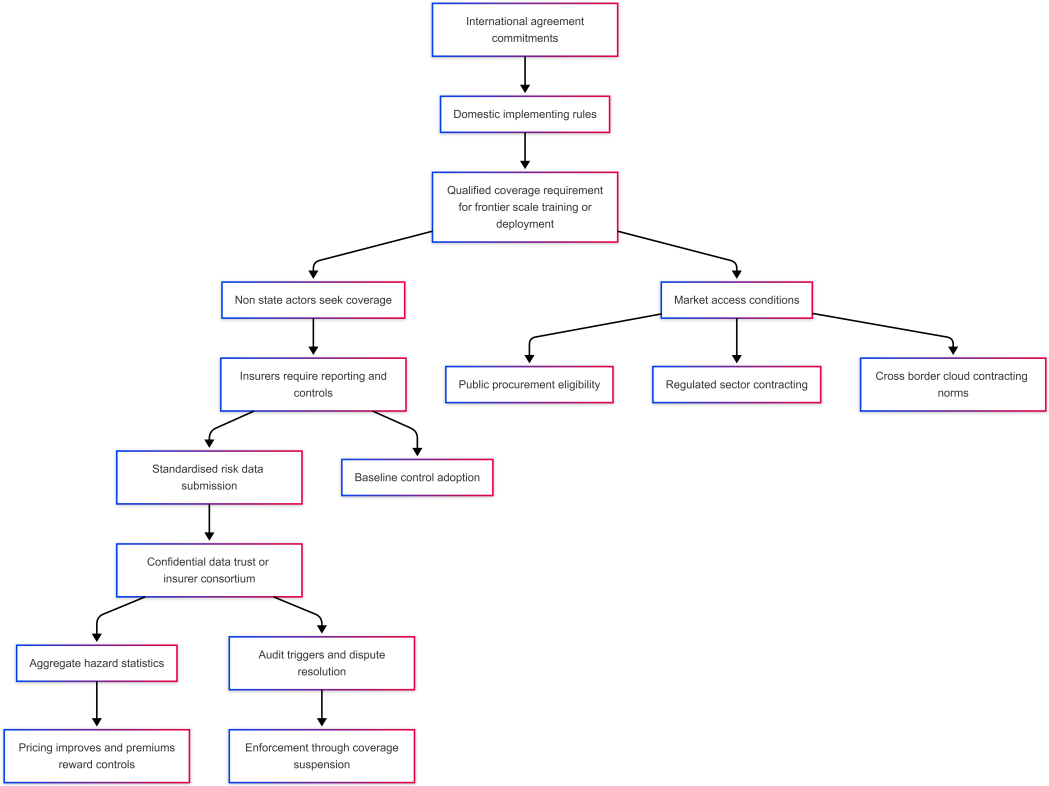


Figure 4: Non state actor inclusion mechanism map

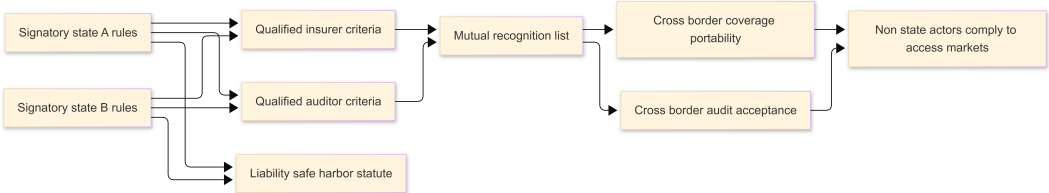


Figure 5: Treaty implementation and mutual recognition map

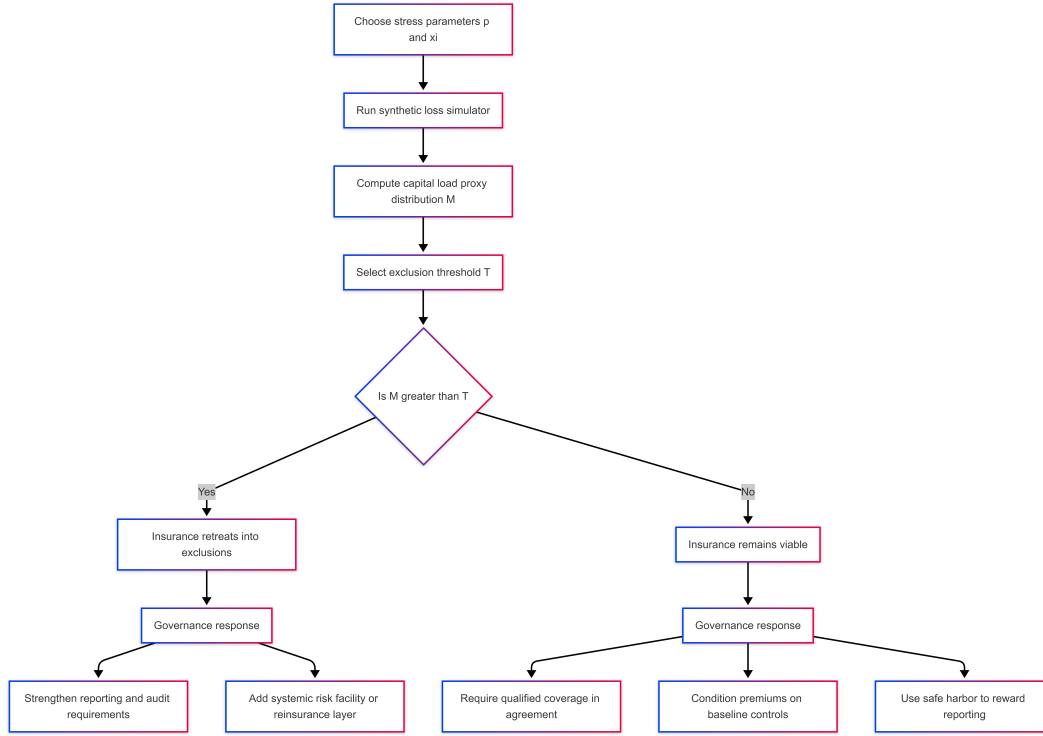


Figure 6: Stress Testing Workflow

vulnerabilities, model families, or deployment patterns; and (iii) observed incidents are incomplete and selected Bucknall et al. (2025).

Regime A Public visibility only, approximating the absence of standardized reporting. The insurer observes only a small and selected subset of incidents and does not observe standardized exposure or control fields (Table 1).

Regime B Standardized confidential incident reporting using the incident fields in Table 1, but limited exposure and control conditioning.

Regime C Standardized confidential reporting of incidents plus exposure and control fields as in Table 1, enabling underwriting to condition on auditable practices.

8.1 STRESS TEST DESIGN

We vary two parameters that are structurally important for solvency and pricing:

- *Systemic shock probability p* : the annual chance of a market-wide incident class that raises loss rates across many insureds.
- *Tail index ξ* : the tail index of the loss distribution which controls how dominant the worst events are. Smaller ξ indicates heavier tails.

For each pair (p, ξ) the simulator computes an insurer-facing capital-load proxy M (defined formally in Appendix B) for each organization; reported results use the median across organizations unless stated otherwise. We evaluate three reporting regimes that reflect increasingly governance-compatible information flows:

Regime A Public visibility only.

Regime B Standardized confidential reporting.

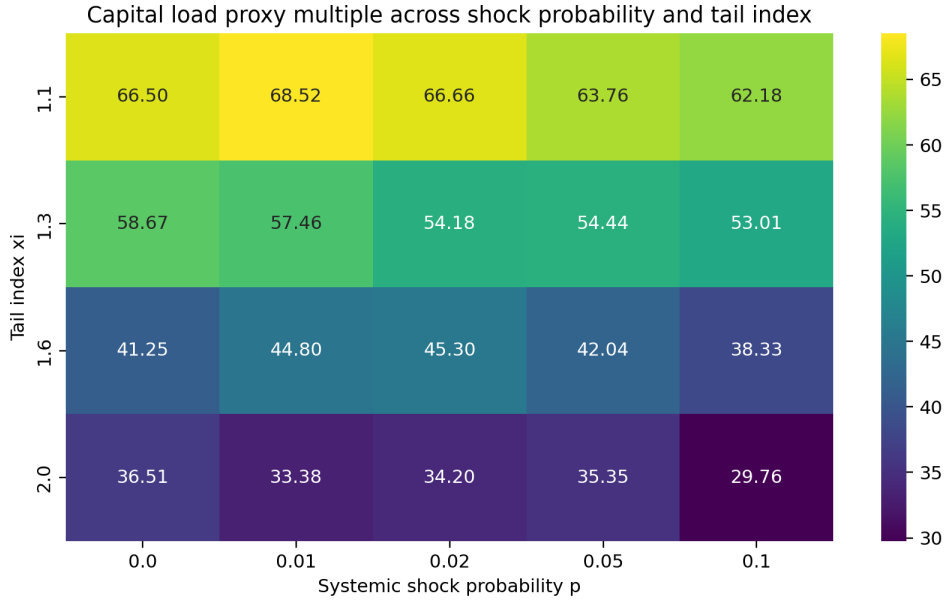


Figure 7: Capital Load Heatmap

Regime C Standardized confidential reporting plus exposure and controls fields that enable underwriting to condition on auditable practices.

Operational definitions of the three regimes and the data-generation assumptions are given in Appendix A.

8.2 CAPITAL-LOAD SURFACE UNDER STRESS

Table 2 presents a representative surface of the capital-load proxy multiple under Regime C without a systemic backstop. Qualitative bands are used in-text because the simulator is synthetic and absolute values depend on the chosen capital metric and parameterization. Exact numeric outputs (with confidence intervals and the script used to produce them) are reported in Appendix C.

Table 2: Capital-load proxy multiple under Regime C (no systemic backstop). Entries are qualitative bands for the median organization.

Tail index $\xi \setminus$ Shock probability p	0.00	0.02	0.05	0.10
1.1	high	higher	very high	extreme
1.3	moderate	moderate-high	high	very high
1.6	low-moderate	moderate	moderate-high	high
2.0	low	low-moderate	moderate	moderate-high

Notes: Bands summarize median capital-load multiples; see Appendix C for numeric values, seed, and full script.

The governance-relevant observation is that improved reporting and conditioning on controls reduces uncertainty and increases incentive gradients, but correlated tail risk remains a binding constraint as p increases or ξ decreases.

8.3 EXCLUSION BOUNDARY DEFINITION

We define *insurance collapse into exclusions* as the regime in which the implied tail capital load is high enough that rational insurers either withdraw coverage or sharply narrow it, leaving safety-

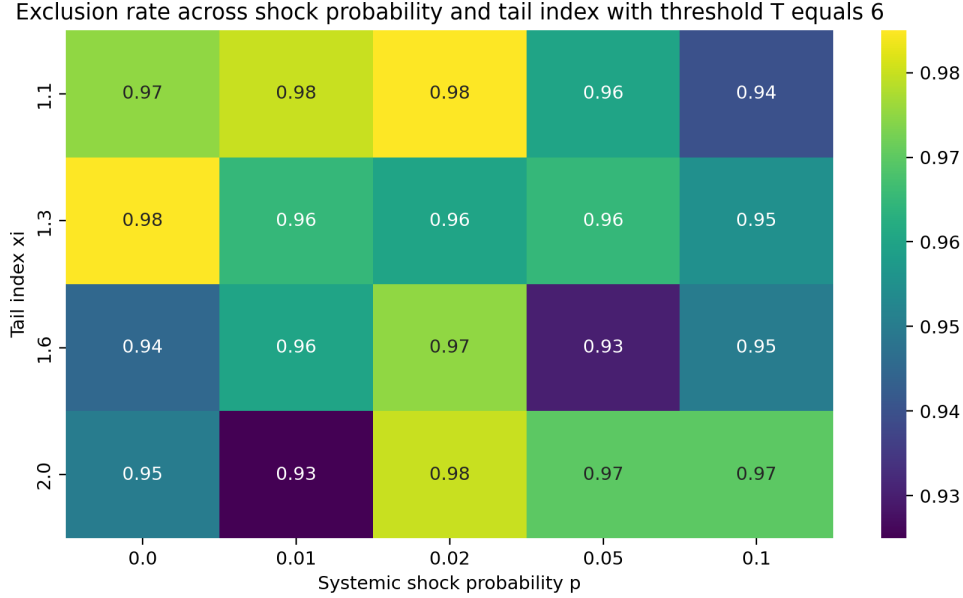


Figure 8: Exclusion Rate Heatmap

relevant hazards uninsured. Operationally, let M be the capital-load proxy multiple for an organization and let T be an exclusion threshold reflecting solvency and competitiveness constraints faced by insurers. We treat coverage as unstable when

$$M > T.$$

When $M > T$ insurers are likely to withdraw or restrict coverage (either by explicit exclusions or by pricing beyond market willingness to pay). We do not assert a single global T ; rather, T is treated as a policy knob that depends on market conditions, reinsurance availability, and regulatory capital treatment.

8.4 EXCLUSION RATES ACROSS DATA REGIMES

Table 3 reports the fraction of organizations that exceed a representative exclusion threshold under a set of illustrative (p, ξ) pairs and under each reporting regime. Cells report qualitative outcomes; the exact simulation configuration and numeric fractions appear in Appendix C.

Table 3: Exclusion rate as a function of data regime and stress parameters (qualitative).

Scenario	Regime A	Regime B	Regime C
Moderate shock probability, moderate tail	highest	lower	lowest
Higher shock probability, moderate tail	near universal	high	material but reduced
Moderate shock probability, heavier tail	high	moderate	low

The qualitative pattern is robust: public-only visibility (Regime A) pushes the system toward exclusion even when tail risk is not extreme; standardized confidential reporting (Regime B) reduces exclusions; adding exposure and control fields (Regime C) reduces exclusions further and creates a premium gradient that can reward specific safety practices.

8.5 GOVERNANCE IMPLICATIONS OF THE EXCLUSION BOUNDARY

The exclusion boundary provides a practical design constraint for international agreements that aim to include non-state actors through insurance and liability. Three implications follow directly from the stress tests:

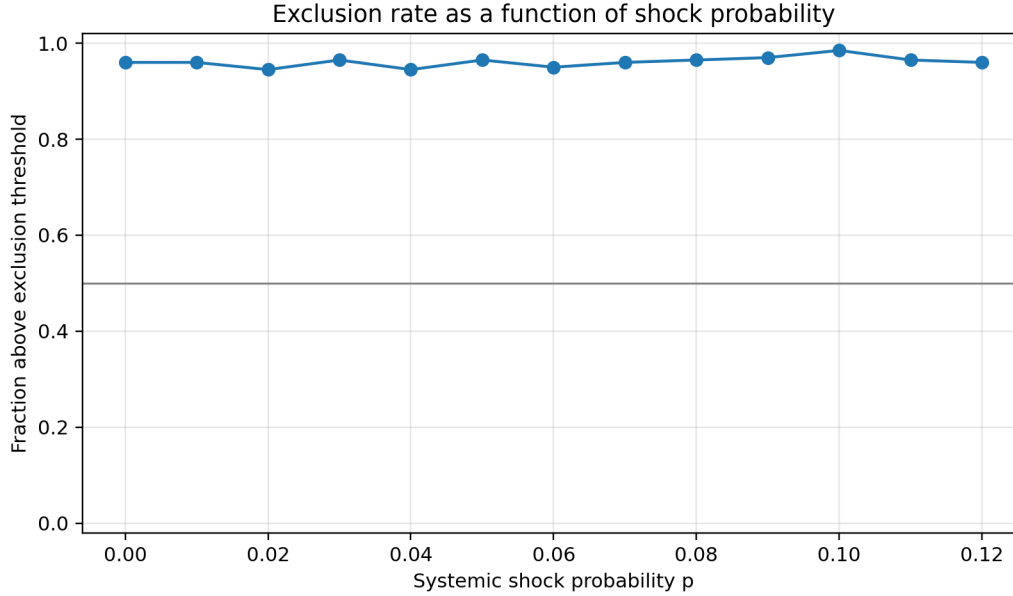


Figure 9: Exclusion Boundary Curve

1. If a treaty or domestic regime requires insurance but does not require standardized reporting and auditable controls, the likely result is formal compliance coupled with hollow policies that exclude core hazards.
2. If a regime requires standardized reporting and conditions coverage on baseline controls, insurance can become a meaningful lever, rewarding safety investments that are costly in the short run and underprovided in competitive settings.
3. When correlated risk is high, reporting alone is insufficient. A systemic risk facility (industry pool, reinsurance layer, or state-contingent backstop) is needed to prevent post-incident market withdrawal. Eligibility for such a facility should be conditioned on reporting participation and baseline controls to avoid subsidizing negligent or reckless deployment.

Interpretation of near universal exclusions. Figures 8 and 9 show exclusion rates close to one under the present synthetic configuration and the illustrative threshold $T = 6$. We interpret this as a design constraint rather than as a forecast about the current insurance market. In regimes with heavy tailed severity and correlated shocks, qualified coverage requirements that do not also impose standardized reporting and auditable controls can induce insurers to narrow coverage or exit precisely for the hazard classes most relevant to catastrophe prevention.

What shifts the exclusion boundary. The simulator highlights two governance levers that can move the system toward stable coverage. First, richer underwriting information improves risk differentiation: moving from Regime B to Regime C effectively adds standardized exposure and control fields (Table 1), which increases the ability to reward specific practices and reduces reliance on outcome only signals. Second, correlated risk sharing reduces capital load: introducing a systemic facility that absorbs a fraction b of losses in systemic shock years reduces tail capital pressure and therefore reduces the incentive for insurers to impose broad exclusions. These levers are complementary: reporting and controls make premiums incentive compatible at the actor level, while a systemic facility addresses correlated portfolio wide shocks that cannot be diversified away.

Concluding remark. The simulator shows that information design (what is reported and how underwriting can condition on auditable practices) meaningfully shifts insurer incentives; however, correlated heavy tails create parameter regimes where private markets alone will narrow coverage without explicit systemic backstops or international risk-sharing arrangements.

9 POLICY RECOMMENDATIONS

This section focuses on implementable levers and explicit failure modes.

10 POLICY RECOMMENDATIONS

This section derives implementable recommendations from the exclusion boundary analysis in Section 8. Motivated by the exclusion pressures in Figures 8 and 9, we recommend that any qualified coverage requirement be paired with standardized reporting fields and baseline controls (Table 1) rather than relying on public visibility or post hoc incident discovery. Motivated by correlated risk dominance in Section 8.2, we recommend that signatories include an option for a systemic risk facility or reinsurance layer that activates in systemic shock states and is conditioned on reporting participation and control compliance.

Treaty level recommendations. Adopt a qualified coverage requirement for frontier model deployment and large scale training above defined thresholds. Define a minimal standardized reporting interface for exposure, controls, and incidents. Require that qualified insurers include reporting as a policy condition. Establish mutual recognition for auditors and qualified insurance terms to avoid fragmented compliance regimes. Create a systemic risk facility option. This can be a pooled reinsurance mechanism conditioned on reporting compliance and baseline controls.

10.1 DOMESTIC REGULATORY RECOMMENDATIONS

Clarify liability allocation across developer and deployer. Encourage contractual allocation but prevent evasion through disclaimers that eliminate real incentives. Establish a reporting safe harbor. The safe harbor should reduce damages when an actor demonstrates prompt reporting and remediation, and should not protect reckless scaling decisions. Require regulated critical sectors to purchase only from covered deployments. This is a procurement lever that reaches non state actors directly. Enable insurer access to confidential risk data under strict confidentiality rules. Without this, insurers cannot price, and the governance mechanism cannot bite.

10.2 RECOMMENDATIONS FOR THE EMERGING AI INSURANCE INDUSTRY

Standardize baseline controls as underwriting requirements. These should be auditable and should focus on deployment conditions rather than on claims about alignment. Use correlation stress tests in solvency analysis and premium setting. Treat systemic failures as the default rather than as an edge case. Support pooled learning across insurers through anonymized aggregate statistics. If the market remains fragmented, each insurer will have insufficient data and will respond by excluding.

11 LIMITATIONS AND FAILURE MODES

Strategic manipulation: organizations may under report or misclassify incidents. The proposed regime relies on audit risk, contractual penalties, and cross validation with external signals. This is imperfect but consistent with other compliance domains.

Excessive secrecy: confidentiality requirements may prevent public learning. The regime should separate public aggregate hazard statistics from confidential incident specifics. *Innovation suppression:* poorly designed liability can chill legitimate research and safety improvement. Safe harbor design is essential. *Model evolution:* risk factors change quickly. Standard fields must include update cadence and post training modifications, and underwriting must be revision based. *International asymmetry:* non signatory states can act as havens. Market access and procurement levers partially mitigate but do not eliminate this risk.

12 CONCLUSION

Non state actors are central to frontier AI capability and risk, and international agreements that fail to include them will remain unstable. Insurance and liability frameworks can create durable incentives, but they are currently blocked by a missing data layer. This paper formalized that data problem, provided a minimal set of underwriting relevant fields, and showed through a correlated heavy tail simulation that reporting quality and exposure controls are first order determinants of premium stability and incentive strength. The proposed mechanism uses qualified coverage requirements and safe harbor liability to bring companies and research institutions into a treaty compatible compliance system without requiring disclosure of proprietary training data. The main practical implication is that policy design should treat risk data standards and reporting incentives as core infrastructure rather than as auxiliary transparency measures Reuel et al. (2025).

@miscbarnett2025aigovernanceavoidextinction, title=AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions, author=Peter Barnett and Aaron Scher, year=2025, eprint=2505.04592, archivePrefix=arXiv, primaryClass=cs.CY, url=https://arxiv.org/abs/2505.04592,

@miscwasil2024verificationmethodsinternationalai, title=Verification methods for international AI agreements, author=Akash R. Wasil and Tom Reed and Jack William Miller and Peter Barnett, year=2024, eprint=2408.16074, archivePrefix=arXiv, primaryClass=cs.CY, url=https://arxiv.org/abs/2408.16074,

@articleWasil2024IAEAforAI, author = Akash Wasil, title = Do We Want an “IAEA for AI”?, journal = Lawfare, year = 2024, month = November, url = https://www.lawfaremedia.org/article/do-we-want-an-iaea-for-ai, note = Explores the potential and limitations of using the International Atomic Energy Agency (IAEA) as a model for international AI governance, discussing verification, enforcement, and geopolitical challenges.,

@onlineTweed2025GlobalLandscapeAISafetyInstitutes, author = Tweed, Rebekah and Kore, Juhi and Dhawan, Ebani, title = The Global Landscape of AI Safety Institutes, year = 2025, month = mar, organization = All Tech Is Human, url = https://alltechishuman.org/all-tech-is-human-blog/the-global-landscape-of-ai-safety-institutes, note = Comprehensive overview of global AI Safety Institutes, including mandates, institutional design, and governance challenges,

REFERENCES

- Yoshua Bengio, Tegan Maharaj, Luke Ong, Stuart Russell, Dawn Song, Max Tegmark, Lan Xue, Ya-Qin Zhang, Stephen Casper, Wan Sie Lee, Sören Mindermann, Vanessa Wilfred, Vidhisha Balachandran, Fazl Barez, Michael Belinsky, Imane Bello, Malo Bourgon, Mark Brakel, Siméon Campos, Duncan Cass-Beggs, Jiahao Chen, Rumman Chowdhury, Kuan Chua Seah, Jeff Clune, Juntao Dai, Agnes Delaborde, Nouha Dziri, Francisco Eiras, Joshua Engels, Jinyu Fan, Adam Gleave, Noah Goodman, Fynn Heide, Johannes Heidecke, Dan Hendrycks, Cyrus Hodes, Bryan Low Kian Hsiang, Minlie Huang, Sami Jawhar, Wang Jingyu, Adam Tauman Kalai, Meindert Kamphuis, Mohan Kankanhalli, Subhash Kantamneni, Mathias Bonde Kirk, Thomas Kwa, Jeffrey Ladish, Kwok-Yan Lam, Wan Lee Sie, Taewhi Lee, Xiaojian Li, Jiajun Liu, Chaochao Lu, Yifan Mai, Richard Mallah, Julian Michael, Nick Moës, Simon Möller, Kihyuk Nam, Kwan Yee Ng, Mark Nitzberg, Besmira Nushi, Seán O hÉigeartaigh, Alejandro Ortega, Pierre Peigné, James Petrie, Benjamin Prud’Homme, Reihaneh Rabbany, Nayat Sanchez-Pi, Sarah Schwettmann, Buck Shlegeris, Saad Siddiqui, Aradhana Sinha, Martín Soto, Cheston Tan, Dong Ting, William Tjhi, Robert Trager, Brian Tse, Anthony Tung K. H., Vanessa Wilfred, John Willes, Denise Wong, Wei Xu, Rongwu Xu, Yi Zeng, HongJiang Zhang, and Djordje Žikelić. The singapore consensus on global ai safety research priorities, 2025a. URL <https://arxiv.org/abs/2506.20702>.
- Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik

Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Luder-mir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alon-dra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bern-hard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ah-met Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, An-drea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Shee-han, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report, 2025b. URL <https://arxiv.org/abs/2501.17805>.

Ben Bucknall, Saad Siddiqui, Lara Thurnherr, Conor McGurk, Ben Harack, Anka Reuel, Patricia Paskov, Casey Mahoney, Sören Mindermann, Scott Singer, Vinay Hiremath, Charbel-Raphaël Segerie, Oscar Delaney, Alessandro Abate, Fazl Barez, Michael K. Cohen, Philip Torr, Ferenc Huszár, Anisoara Calinescu, Gabriel Davis Jones, and Robert Trager. In which areas of technical ai safety could geopolitical rivals cooperate? In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, pp. 3148–3161, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732201. URL <https://doi.org/10.1145/3715275.3732201>.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, An-drew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2025. URL <https://arxiv.org/abs/2407.14981>.

Rebekah Tweed, Juhi Kore, and Ebani Dhawan. The global landscape of ai safety institutes, March 2025. URL <https://alltechishuman.org/all-tech-is-human-blog/the-global-landscape-of-ai-safety-institutes>. Comprehensive overview of global AI Safety Institutes, including mandates, institutional design, and governance challenges.

Akash Wasil. Do we want an “iaea for ai”? *Lawfare*, November 2024. URL <https://www.lawfaremedia.org/article/do-we-want-an-iaea-for-ai>. Explores the potential and limitations of using the International Atomic Energy Agency (IAEA) as a model for international AI governance, discussing verification, enforcement, and geopolitical challenges.

A APPENDIX

A STOCHASTIC LOSS SIMULATOR AND REPORTING REGIMES

This appendix specifies the synthetic loss simulator used for the stress tests in Section 8. The simulator is intended for mechanism evaluation rather than empirical actuarial forecasting. It is designed to preserve three properties that drive governance relevant insurance feasibility: heavy tailed severity, correlated systemic shocks, and selective reporting.

A.1 ORGANIZATIONS, EXPOSURE, AND CONTROLS

We consider N organizations indexed by $i \in \{1, \dots, N\}$. Each organization has an exposure variable e_i and a control variable c_i .

Exposure is modeled as lognormal to represent strong scale dispersion:

$$e_i \sim \text{LogNormal}(\mu_e, \sigma_e). \quad (10)$$

Controls are modeled as a standardized covariate:

$$c_i \sim \mathcal{N}(0, 1). \quad (11)$$

An optional organization random effect u_i captures unobserved heterogeneity:

$$u_i \sim \mathcal{N}(0, \sigma_u^2). \quad (12)$$

A.2 INCIDENT FREQUENCY MODEL

Let K_i denote the annual number of claim generating incidents for organization i . We use a Poisson model with log linear dependence on exposure and controls:

$$K_i \sim \text{Poisson}(\lambda_i), \quad (13)$$

$$\log \lambda_i = \alpha + \beta_e \log e_i + \beta_c c_i + u_i. \quad (14)$$

This form is chosen to make explicit which data fields allow underwriting to condition on risk relevant covariates.

A.3 SEVERITY MODEL

Each incident has a non negative severity S_{ij} . We use a Pareto severity model to represent heavy tails:

$$S_{ij} \stackrel{\text{iid}}{\sim} \text{Pareto}(s_{\min}, \xi), \quad (15)$$

where $s_{\min} > 0$ is a minimum scale and $\xi > 0$ is a tail index. Smaller ξ implies heavier tails.

Total annual loss for organization i is the compound sum

$$L_i = \sum_{j=1}^{K_i} S_{ij}. \quad (16)$$

A.4 CORRELATED SYSTEMIC SHOCKS

To represent correlated failure modes that affect many actors in the same year, we introduce a systemic shock indicator $Z_t \in \{0, 1\}$ for year t :

$$Z_t \sim \text{Bernoulli}(p), \quad (17)$$

where p is the systemic shock probability.

Conditional on Z_t , we modify both incident frequency and severity tail behavior:

$$\log \lambda_{i,t} = \log \lambda_i + \delta Z_t, \quad (18)$$

$$\xi_t = \xi + \Delta_\xi Z_t. \quad (19)$$

In the experiments reported in Section 8, $\delta > 0$ increases event rates in shock years. The sign and magnitude of Δ_ξ control how severity tails change in shock years. All parameters are synthetic and are selected for stress testing rather than empirical calibration.

A.5 SELECTIVE REPORTING MODEL

Observed incidents are modeled as a selective subset of true incidents. Let $R_{ij,t} \in \{0, 1\}$ indicate whether incident (i, j) in year t is visible to the underwriting dataset under the relevant reporting regime.

We model reporting probability as a logistic function of severity and controls:

$$\Pr(R_{ij,t} = 1 \mid S_{ij,t}, c_i) = \sigma(\gamma_0 + \gamma_s \log S_{ij,t} + \gamma_c c_i), \quad \sigma(x) = \frac{1}{1 + e^{-x}}. \quad (20)$$

This captures that large incidents are more likely to be observed and that controls can affect detection and disclosure.

A.6 REPORTING REGIMES A, B, AND C

We evaluate three reporting regimes. In the implementation, regimes are represented by shifts to the reporting intercept and by which covariates are available to underwriting.

Regime A (public visibility only). Reporting probability is low. Underwriting sees a small fraction of true incidents concentrated in the upper severity tail. Exposure and control fields are not available for risk conditioning.

Regime B (standardized confidential reporting). Reporting probability is higher. Underwriting sees most incidents above a low threshold. Exposure and control fields remain limited.

Regime C (standardized reporting plus exposure and controls). Reporting probability is higher as in Regime B, and underwriting additionally observes e_i and c_i (or proxies derived from standardized reporting fields). This enables pricing and audit triggers to condition on auditable practices.

A.7 SYSTEMIC FACILITY BACKSTOP

To model correlated risk sharing, we introduce a systemic facility that pays a fraction $b \in [0, 1]$ of losses in systemic shock years. Net losses become

$$L_{i,t}^{\text{net}} = \begin{cases} (1-b)L_{i,t} & \text{if } Z_t = 1, \\ L_{i,t} & \text{if } Z_t = 0. \end{cases} \quad (21)$$

This is a reduced form representation of pooled coverage, reinsurance layers, or state contingent facilities.

B CAPITAL LOAD PROXY AND EXCLUSION BOUNDARY

This appendix defines the capital load proxy multiple and the exclusion boundary used in Section 8.

B.1 TAIL RISK METRIC

Let L be an annual loss random variable. Define value at risk at level $q \in (0, 1)$:

$$\text{VaR}_q(L) = \inf\{x : \Pr(L \leq x) \geq q\}. \quad (22)$$

Define conditional value at risk:

$$\text{CVaR}_q(L) = \mathbb{E}[L \mid L \geq \text{VaR}_q(L)]. \quad (23)$$

In the experiments we use a high percentile such as $q = 0.99$.

B.2 CAPITAL LOAD PROXY MULTIPLE

For each organization i , define the capital load proxy multiple:

$$M_i = \frac{\text{CVaR}_q(L_i)}{\mathbb{E}[L_i]}. \quad (24)$$

Large M_i indicates that tail losses dominate expected losses, which tends to induce high capital charges and motivates exclusions in competitive insurance markets.

B.3 EXCLUSION BOUNDARY AND EXCLUSION RATE

Given an exclusion threshold $T > 0$, we define exclusion for organization i as the event

$$M_i > T. \quad (25)$$

For a given parameter setting and regime, the exclusion rate is the fraction of organizations for which $M_i > T$.

C ADDITIONAL NUMERIC OUTPUTS AND SENSITIVITY CHECKS

This appendix collects additional outputs that are omitted from the main paper for space. All results are synthetic and are used to evaluate governance mechanisms rather than to estimate real world premiums.

C.1 SENSITIVITY TO THRESHOLD CHOICE

The main text uses $T = 6$ as an illustrative exclusion threshold. To check robustness, we recommend varying T over a plausible range and reporting exclusion rates as a function of T . This distinguishes cases where near universal exclusions are driven by the selected threshold from cases where exclusions persist across thresholds due to heavy tails and correlation.

C.2 SENSITIVITY TO TAIL PERCENTILE CHOICE

The capital load proxy uses CVaR_q at a high percentile. Varying q (for example 0.95, 0.99, 0.995) tests whether conclusions are driven by extreme tail emphasis. In regimes where correlated risk dominates, exclusion pressure should remain high even under moderate q .

C.3 SENSITIVITY TO CORRELATED RISK SHARING

The systemic facility is modeled as a shock state payout share b . Varying b illustrates how much correlated risk pooling is required before coverage becomes stable. This sensitivity directly maps to governance design choices such as pooled facilities, reinsurance layers, or state contingent backstops with eligibility conditions.