
No One Thanks You for Disasters That Never Happened: Pricing AI Risk While Making AI Safety Investable¹

Ana Belen Barbero Castejon

Stan Threepwood

With
Apart Research

Abstract

This paper examines whether risk quantification and pricing can function as a practical mechanism of AI governance. We present a prototype framework for pricing AI risk under deep uncertainty, using a scenario-based frequency–severity decomposition with dependency-aware propagation and aggregation with catastrophe-style tail modeling. Loss distributions are evaluated using standard actuarial metrics such as Expected Loss (EL), Value at Risk (VaR), and Tail Value at Risk (TVaR), enabling premium estimation in data-scarce environments.

These findings suggest a potential pathway for AI insurance and reinsurance to function as coordination infrastructure, aligning incentives across companies, developers, insurers, and regulators while enabling scalable investment in AI safety, security and operations.

Application available here: <https://pricing-ai-risks.netlify.app>

Code available here: <https://github.com/stanthreepwood/ai-risk-pricing>

Keywords: AI risk quantification, AI insurance, AI safety incentives, algorithmic risk modeling, AI supply-chain catastrophe modeling

This work was developed during a weekend hackathon. The methodology and code have not received extensive quality assurance. Claims in this report should be approached with appropriate caution.

1. Introduction

1.1. Governance problem: When Risk has no price.

International efforts to govern advanced AI systems rely on technical thresholds, risk classifications and safety commitments to guide prioritization, enforcement and cross-border coordination (e.g., [EU AI Act](#); [NIST AI Risk Management Framework](#); [Anthropic Responsible Scaling Policy](#)). Yet these initiatives lack a shared way to quantify and compare risk across systems, use cases and actors. **In the absence of comparable risk signals, governance remains fragmented: oversight does not scale, obligations are difficult to align across jurisdictions and commitments are hard to verify or enforce.**

¹ Research conducted at the Apart Hackathon, [The Technical AI Governance Challenge](#) (2026)

More critically, this gap undermines incentives. When technical risk is not translated into economically meaningful signals, actors across the AI supply chain face prioritization challenges when investing in safety, security or governance. As a result, risk often remains an unquantified external regulatory constraint, rather than something managed through everyday decisions about product development, capital allocation and operations.

1.2. Risk pricing as governance infrastructure.

One underexplored way to address this gap is through risk pricing. In domains marked by uncertainty and asymmetric information, pricing risk, especially through insurance, has historically served as a coordination mechanism, translating diverse technical risks into standardized metrics.

Beyond loss transfer, **insurance markets aggregate dispersed information, shape behavior through coverage conditions and exclusions and create incentives for risk reduction by making governance and safety investments economically legible.** Recent work has argued that similar mechanisms could play a structural role in governing frontier AI systems, embedding safety requirements into capital allocation and operational practices rather than relying solely on prescriptive regulation ([Stetler, 2025](#)).

Applied to AI, risk pricing could complement regulatory approaches by aligning the incentives of governments, laboratories, and downstream actors across the AI supply chain. Rather than replacing regulation, pricing-based mechanisms offer a way to operationalize it, converting abstract thresholds and risk categories into signals that support prioritization, verification and cross-border coordination.

1.3. Related work

1.3.1. AI insurance and lessons from CAT and cyber risk.

The current state of AI risk pricing closely mirrors the early development of cyber insurance. Early cyber markets emerged under conditions of sparse incident data, rapidly evolving threat surfaces, and deep uncertainty, relying heavily on expert judgment, scenario analysis, and iterative model updating rather than stable actuarial regularities ([OECD, 2017](#); [Romanosky et al., 2019](#)).

Despite these limitations, cyber insurance nonetheless became a governance-relevant coordination mechanism: it aggregated risk information, shaped security practices through coverage conditions, and provided decision-relevant signals long before precise loss estimation was possible ([MacColl et al., 2021](#)). **The core lesson is that useful risk signals can precede mature statistical models, and that pricing can play a governance role even under deep uncertainty.**

1.3.2. From sparse data to systemic dependence: the AI risk modeling problem.

Quantifying AI risk remains challenging for three related reasons. First, the field is still at an early stage of empirical observation, with limited reported incident data. As a result, efforts to characterize AI risks have largely focused on the development of qualitative taxonomies, relying on heterogeneous classification schemes ([Wisakanto et al., 2025](#), [MIT AI Risk Initiative](#)).

Second, and more fundamentally, AI risk exhibits structural characteristics that strain conventional risk models. AI systems are embedded in a tightly coupled supply chain (spanning foundation model developers, application-layer providers, and downstream enterprises) where

shared architectures, datasets, and infrastructure create strong dependencies between actors. Losses are therefore highly correlated: operational failures, misuse events, regulatory breaches, or capability risks at one layer can propagate across many organizations simultaneously.

In this setting, risk is driven less by independent incident frequency than by aggregation, concentration, and tail dependence. Effective quantification thus requires modeling systemic interactions and layered exposure across the AI value chain.

Third, AI risks differ in their degree of observability and evidentiary grounding. Some risk surfaces have already partially materialized and can be informed by incident data and taxonomies; others are anticipated through regulatory and policy frameworks that define severity and liability boundaries despite limited empirical realization; and a further class remains poorly characterized, involving speculative or catastrophic failure modes.

1.4. Research contributions.

This study examines **whether risk pricing can function as a practical mechanism of AI governance**. The goal is to assess whether partial evidence and heterogeneous information (derived from observed incidents, regulatory exposure and anticipated risk surfaces) can be combined with explicit modeling of risk propagation across the AI supply chain to generate governance-relevant risk signals that support incentive alignment across actors.

A core premise in this study is that AI risks differ in their degree of observability. To scope what can and cannot be meaningfully analyzed, we adopt and adapt an awareness–understanding risk matrix (inspired by Rumsfeld matrix, 2002).

We distinguish between (i) **known–known risks**, informed by early AI incidents; (ii) **known–unknown risks**, where research, regulatory and policy frameworks define boundaries for harm despite limited empirical realization; and (iii) **unknown–unknown risks**, comprising speculative or catastrophic failure modes that resist defensible quantification.

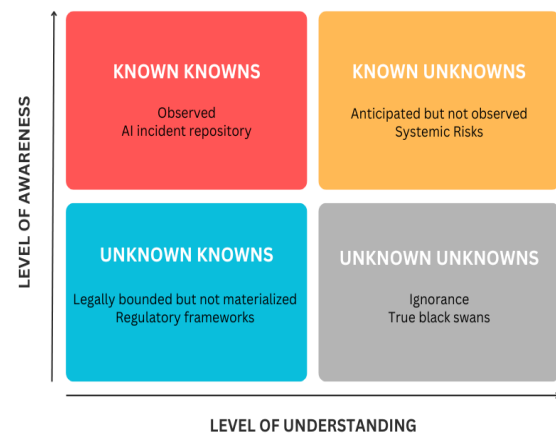


Figure 1 | Rumsfeld-inspired risk matrix for AI governance.

Building on this distinction, the paper proposes a **systemic AI risk pricing framework** that models correlation, aggregation, and tail behavior across the AI supply chain (spanning foundation model providers, application-layer systems, and downstream enterprises) rather than treating AI incidents as independent events. This framework constitutes the primary contribution of the study and provides the foundation for the empirical analysis that follows.

Against this backdrop, we formulate two research questions and associated hypotheses:

RQ1: Under plausible assumptions grounded in observed incidents and legally bounded regulatory exposure, do modeled AI risks remain within insurable/reinsurable bounds?

RQ2: Can risk pricing mechanisms translate AI safety and governance measures into economically meaningful incentives?

Regarding (iv) **unknown-known risks and information**, this paper can be understood as an attempt itself to surface insights and forms of knowledge that may be hidden, overlooked, or siloed across other disciplines, and to bring them into the AI governance space.

2. Methods

The proposed framework operationalizes AI risk pricing under partial observability, structural dependence and regulatory uncertainty. It combines **scenario-based risk decomposition into frequency-severity models**, **explicit dependency modeling across the AI supply chain**, and **Monte Carlo simulation** to generate an aggregated portfolio-level loss distribution and pricing-relevant risk metrics.

Consistent with the adapted Rumsfeld matrix, the methodology distinguishes between *observed risks*, *legally bounded but weakly observed risks*, and *speculative or poorly characterized risks*. In this paper, only the first two categories are quantitatively simulated. Unknown-unknown risks are explicitly excluded from the scope of this project at this stage, as they constitute a necessarily speculative domain. Nevertheless, they are considered within this framework and may be incorporated in later iterations of this research.

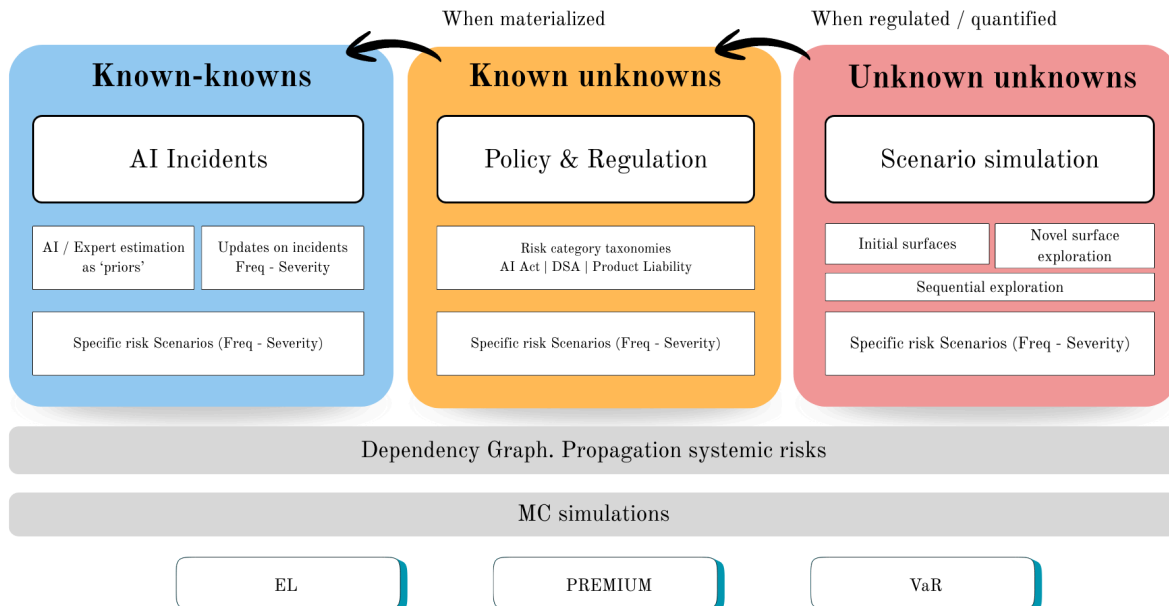


Figure 2 | Overview of the methodology and experimental framework for AI risk pricing. AI risk scenarios derived from incidents and regulatory exposure are mapped to a known-unknown risk matrix and propagated through a dependency graph of the AI supply chain. Monte Carlo simulations produce aggregate loss distribution and pricing-relevant risk metrics.

At a high level, the framework proceeds in four steps:

1. Construction of AI risk scenarios with parameterized frequency and severity distributions, with explicit consideration of mitigation measures.
2. Encoding of systemic dependencies via a supply-chain dependency graph. Foundational models, SaaS application layer and enterprise *layers*.
3. Monte Carlo simulations and portfolio-level loss aggregation.
4. Computation of risk metrics and insurance premiums.

Each component is described in detail in [Appendix B](#).

2.1. Metrics and Ratios.

To evaluate whether the proposed framework produces governance-relevant risk signals, and to test the research questions, we rely on a standard set of actuarial and risk management metrics. These metrics are chosen to capture not only expected losses, but also tail behavior, capital intensity and the economic implications of risk pricing.

1. **Expected Loss (EL):** It represents the mean annual loss across simulation runs. EL serves as the baseline measure of risk magnitude and is used to assess how different risk blocks, dependency structures and mitigation measures affect average loss exposure:

$$EL_c = \sum_{s \in \mathcal{S}} \lambda_s \cdot \mathbb{E}[L_{c,s}^{\text{mit}}] = \sum_{s \in \mathcal{S}} \lambda_s \cdot \phi_{c,s} \cdot \mathbb{E}[L_{c,s}]$$

being:

$c \in \mathcal{C}$: companies

$s \in \mathcal{S}$: AI risk scenarios

λ_s : annual frequency of scenario sss

$L_{c,s}$: random loss severity for company c under scenario s, before mitigation

$\phi_{c,s} \in [\epsilon, 1]$: residual risk factor from mitigation

2. **Value at Risk (VaR):** Value at Risk at the 95% confidence level measures the loss threshold exceeded with a fixed low probability. In other words, it indicates the size of loss we expect not to exceed in most years, except in rare, adverse scenarios. VaR captures the impact of aggregation and correlation on tail outcomes and is used to assess how dependency modeling alters risk concentration relative to independent assumptions:

$$\text{VaR}_\alpha(Z) = \inf \{z \in \mathbb{R} : \mathbb{P}(Z \leq z) \geq \alpha\}$$

being:

α = tail confidence level, in our case 95%-99%

Z : aggregated annual loss

Z : exceed loss

3. **Tail Value at Risk (TVaR):** It measures the expected loss conditional on losses exceeding VaR_{95} . TVaR is particularly relevant for systemic and correlated AI risks, as it reflects tail severity under extreme but plausible scenarios. It is usually used while evaluating insurability under tail risk constraints. For continuous loss distributions, it is defined by:

$$\text{TVaR}_\alpha(Z) = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_u(Z) du$$

being:

α = tail confidence level 95%-99%

4. **Premium:** Insurance premiums are derived from simulated loss distributions, incorporating EL, tail risk (VaR and TVaR), and risk loading assumptions. Premium levels provide a direct economic signal linking modeled AI risk to capital costs and risk transfer feasibility (RQ2, RQ3). We define it as (note the extra uncertainty typical from CAT modelling and operative expense ratio):

$$\Pi = \text{EL}_P + \beta \cdot \text{TVaR}_\alpha(Z_P) + \gamma \cdot \text{EL}_P$$

being:

α = tail confidence level, in our case 95%-99%
 $\beta \in (0, 1)$: ambiguity / model uncertainty loading
 γ : expense ratio as a EL proportion

To facilitate comparison across insurance industry standards, risk blocks, and mitigation settings, we also compute the following ratios:

- **Rate on Line (RoL):** It captures the capital intensity of risk transfer and provides a benchmark for comparing modeled AI risks with established insurance lines. It is defined by Coverage Limit / Premium. For cyber it is usually around [7%, 20%] while for catastrophic risks in the [100%, 400%] range ([Heinrich et al, 2019](#)).
- **Premium / EL Multiple:** This ratio reflects the degree of risk loading required to cover uncertainty, correlation, and tail risk. Rising multiples indicate increasing uninsurability or governance stress. Cyber, early-stage risks 3x-6x, more than that deep uncertainty and ambiguity-dominated pricing.
- **TVaR / EL Ratio:** This ratio measures tail dominance relative to average loss and serves as a proxy for systemic risk intensity. High values indicate fat-tailed behavior and aggregation risk that challenge conventional pricing and capital adequacy assumptions. The range 3x-6x corresponds to heavy-tailed cyber, catastrophe-like or systemic collapse risk.

2.2. Implementation and reproducibility.

The framework is implemented as an interactive application to support exploration, sensitivity analysis, and scenario testing. For reference:

- Project website: <https://pricing-ai-risks.netlify.app/>
- ◆ Application (also available from web): <https://pricing-ai-risks.streamlit.app/>
- Github repo: <https://github.com/stanthreepwood/ai-risk-pricing>

3. Results

3.1. Our framework.

The experiments validate the proposed framework as a coherent way to model AI risk at the system level rather than as isolated incidents. As discussed, by explicitly representing dependency, aggregation, and tail behavior across the AI supply chain, the framework produces loss distributions that differ materially from independent risk assumptions and reveal systemic amplification effects.

3.1.1 Mitigation measures in the proposed framework (RQ2).

As described in the Methodology and Expected Loss computation, the framework incorporates an explicit term for mitigation measures. In practice, each risk scenario is associated with a defined set of known mitigation measures. Entities exposed to these risks (directly or through propagation effects) may reduce their risk exposure by implementing appropriate AI safety tools and governance controls throughout the AI system lifecycle.

Within the created demo application, companies requesting coverage can assess their existing AI safety and security practices and directly observe their impact on modeled losses and

insurance premiums. The application also allows users to evaluate the marginal benefit of additional mitigation measures and enables the quantification of the value of AI ops, security and safety tools.

3.2. On the insurability of AI risks (RQ1).

Using the proposed framework, we evaluate the impact of different and incremental AI risk surfaces, as well as their interactions across the modeled ecosystem. Under the initial experiment setup and baseline assumptions, the expected loss is around \$1.3B, with a 99% VaR of \$8.6B and TVaR of \$12.7B, corresponding to a Rate on Line of 151% and a Premium/EL multiple of 6x. The resulting metrics are generally borderline, aligning with observations from early-stage cyber risk and catastrophic loss scenarios, as anticipated. Premiums that insurers would require are extremely high, driven primarily by systemic uncertainty, and in all cases, the loss tails remain notably heavy.

	Expected Loss (EL)	VaR 99%	TVaR 99%	Maximum Simulated Loss	Yearly Loss Occurrence Rate	Total Premium
Base Scenario	\$1,292.71 M	\$8,647.86 M	\$12,675.68 M	\$18,172.76 M	52.8%	\$7,953.73 M
1 + Known-Known AI Incidents	\$7,663.97 M	\$18,060.40 M	\$19,848.89 M	\$21,064.46 M	100.0%	\$19,504.41 M
2 + Known-Unknown Regulatory Risk	\$8,266.38 M	\$17,597.93 M	\$19,849.23 M	\$22,661.14 M	100.0%	\$20,257.59 M

It is worth noting that when additional materialized risks and regulatory exposures are incorporated into the model, the resulting loss distributions exhibit higher expected losses but a softening of the extreme tails (L rises to \$8.3B with a corresponding TVaR of \$19.8B). Under these conditions, the modeled risk may fall within insurability thresholds if reinsurance structures, mitigation measures, and policy exclusions are properly designed, opening a clear avenue for future work.

	Rate on Line (RoL)	Premium / EL Multiple	TVaR / EL Ratio
Base Scenario	150.98%	6.15x	9.81x
1 + Known-Known AI Incidents	370.23%	2.54x	2.59x
2 + Known-Unknown Regulatory Risk	384.53%	2.45x	2.40x

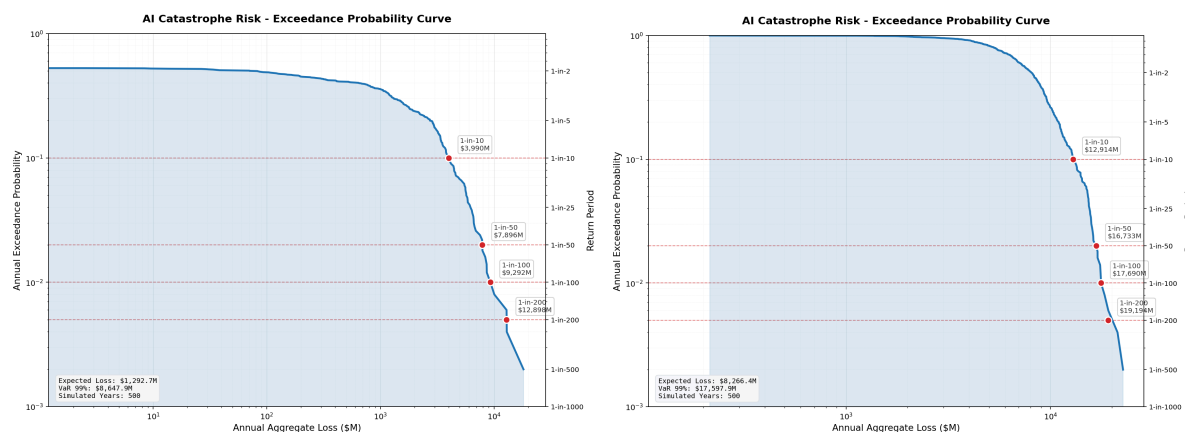


Figure 3 | AI Catastrophe Risk. Exceedance Probability Curve. Base scenario (left) vs K-K. K-U (right).

These conclusions should be interpreted as provisional and conditional on the current parameterization of the model. Nonetheless, we contend that the underlying risk decomposition, the emergence of catastrophic nonlinear dynamics, and the qualitative direction of the observed results are robust to reasonable parameter variation ([Figure 11](#)). As such, we think that the framework provides a meaningful foundation for continued development, refinement, and calibration as additional data and regulatory clarity become available, enabling insurers and reinsurers to assess both expected and catastrophic loss scenarios, and incentivizing investment in mitigation and safety measures across the AI ecosystem.

4. Conclusion and future work

These results indicate that risk pricing can serve as a practical component of technical AI governance under conditions of partial observability and systemic dependence. By translating heterogeneous AI risks into comparable economic signals, the framework helps align incentives across the AI supply chain and operationalize regulatory and safety requirements beyond static classifications. Crucially, this enables prioritization, comparability, and scalability in governance contexts where direct technical measurement and enforcement are otherwise difficult.

While the analysis is limited to known and partially known risks and depends on current modeling assumptions (see [Appendix A](#)), the findings show that governance-relevant risk signals can emerge even without mature actuarial data. Modeling aggregation, correlation, and mitigation effects allows risk pricing to complement regulatory approaches and support incentive-aligned investment in AI safety, security, and governance.

We are positive that this research direction can be industry-shaping, and acknowledge that the robust estimation of each of our framework components is a work that goes far beyond a weekend hackathon.

Future research could extend this framework along several dimensions, including:

- Validating and calibrating the model using a broader set of empirical incidents, claims data, and regulatory outcomes.
- Using an expert panel to improve the initial calibration and distribution characterization of unmaterialized risks.
- Expanding and refining regulatory taxonomies to better capture international jurisdictional differences and enforcement mechanisms.
- Broadening the concept of what we called *layers*, currently companies, to include additional social and institutional actors.
- Incorporating dynamic adversarial behavior and feedback effects into risk frequency and severity estimates.

5. References

- European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Official Journal of the European Union L 1689. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689>
- European Parliament & Council of the European Union. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act)*, Official Journal of the European Union L 277/1. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Report No. NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Anthropic. (2025, May 14). *Responsible Scaling Policy (Version 2.2)*. <https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf>
- Stetler, N. (2025). *Reinsuring AI: Energy, agriculture, finance & medicine as precedents for scalable governance of frontier artificial intelligence* (arXiv:2504.02127). arXiv. <https://arxiv.org/abs/2504.02127>
- OECD. (2017). *Enhancing the role of insurance in cyber risk management*. OECD Publishing. <https://doi.org/10.1787/9789264282148-en>
- Romanosky, S., Ablon, L., Kuehn, A., & Jones, T. (2019). *Content analysis of cyber insurance policies: How do carriers price cyber risk?* *Journal of Cybersecurity*, 5(1), tyz002. <https://doi.org/10.1093/cybsec/tyz002>
- MacColl, J., Nurse, J. R. C., & Sullivan, J. (2021). *Cyber insurance and the cyber security challenge* (Occasional Paper). Royal United Services Institute. <https://static.rusi.org/247-op-cyber-insurance-v2.pdf>
- Wisakanto, A. K., Rogero, J., Casheekar, A. M., & Mallah, R. (2025). *PRA for AI: Probabilistic Risk Assessment for AI* (Version v0.9.1-alpha). Center for AI Risk Management & Alignment. Retrieved February 1, 2026, from <https://pra-for-ai.github.io/pra/index.html>
- MIT AI Risk Initiative. (n.d.). *AI Risk Repository*. Massachusetts Institute of Technology. <https://airisk.mit.edu/>
- UK Department for Science, Innovation and Technology. (2025). *International AI Safety Report 2025: International Scientific Report on the Safety of Advanced AI*. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025>
- Heinrich, T., Sabuco, J., & Farmer, J. D. (2019). *A simulation of the insurance industry: The problem of risk model homogeneity* (INET Oxford Working Paper No. 2019-12). Institute for New Economic Thinking at the Oxford Martin School, University of Oxford. https://oms-inet.files.svcdn.com/production/files/working_paper_CAT_Insurance_2019_07_12.pdf (oms-inet.files.svcdn.com)

6. Appendix

Appendix A: Limitations.

- a. This work represents an initial attempt to quantify AI risks and to integrate historical damages and regulatory sanctions into a unified risk model. As such, the framework remains exploratory and necessarily incomplete.
- b. A key limitation concerns adversarial risk dynamics. In adversarial settings, risk frequency and severity are not independent variables: attempts to probe system vulnerabilities can both reveal exploitable weaknesses and increase the likelihood and impact of subsequent attacks. These feedback loops are not yet fully captured in the current model.

Despite these limitations, we view this work as a **meaningful proof of concept**. It demonstrates the feasibility of translating AI risk into quantifiable signals and points toward a direction with the potential to support industry-wide coordination and unlock capital for the safe adoption of AI across sectors.

Appendix B: Methodology and Experimental Setup

a) Risk scenario construction

Incident-based scenarios (known-known risks)

We begin with materialized AI risks, mapped to the *known-known* quadrant of the risk matrix. These scenarios are derived from the [MIT AI Incident Repository](#), a live database that documents real-world AI-related failures across sectors and deployment contexts.

For each incident category, we estimate frequency and severity distributions in two stages:

- **Baseline parameterization:** As an initial benchmark, frequency (λ) and severity (lognormal) distribution parameters are extracted (by bootstrapping LLMs or ideally by an expert group in a more realistic scenario) to generate coarse estimates from incident descriptions. This step serves solely as a baseline for the cold-start problem.
- **Data-driven adjustment:** These estimates are then adjusted using patterns observed in the incident repository with the weighted MLE. Incidents are grouped along two dimensions: (i) **System layer** subset (foundation model, application layer, enterprise deployment, see more in [\(b\)](#)); (ii) **Incident type** (e.g., hallucination, misuse, governance failure). For example, for updating/recomputing the lognormal severity distribution, we simply have:

$$\hat{\mu}_{n+1} = \frac{n\hat{\mu}_n + \ln s_{n+1}}{n+1} \quad \hat{\sigma}_{n+1}^2 = \frac{n\hat{\sigma}_n^2 + (\ln s_{n+1} - \hat{\mu}_n)(\ln s_{n+1} - \hat{\mu}_{n+1})}{n+1}$$

Regulatory and policy-based scenarios (known-unknown risks)

We next construct scenarios corresponding to anticipated but weakly observed risks, mapped to the *known-unknown* quadrant of the matrix. These scenarios are derived from regulatory frameworks, policies, and codes of practice that define explicit boundaries on acceptable risk, liability, and severity.

In this work, we focus on two regulatory regimes, the [EU AI Act](#) and the [Digital Services Act](#) (DSA). For each, we construct a dedicated taxonomy (documented in [Appendix E](#) and [Appendix F](#)) that maps risk categories to: System layers, Affected sectors, Relative frequency and severity ranges implied by regulatory sanctions, prohibitions, and compliance obligations

Where a given risk category can originate at multiple layers of the AI system architecture, the corresponding scenario is duplicated across layers. This avoids conflation, supports clearer attribution, and enables layer-specific aggregation and dependency modeling.

The framework is modular and can be extended to additional national or sector-specific regulatory regimes.

Accordingly, the quantitative results reported in this paper reflect known and partially known AI risk surfaces only: observed incidents and legally bounded regulatory exposures.

```
@dataclass
class Scenario:
    """
    Complete specification of an AI catastrophe scenario.

    A scenario represents a specific type of AI failure event with all
    parameters needed for stochastic simulation. Scenarios are the primary
    input to the Monte Carlo engine.

    Both the frequency (how often events occur) and severity
    (how large losses are when events occur) components of the compound frequency-severity model used in catastrophe pricing.

    The 'known_mitigations' field specifies which safety controls can reduce
    this scenario's severity, enabling risk-aware pricing that rewards
    companies with appropriate safety measures.
    """

    name: str
    event_type: EventType
    trigger: str
    propagation_vector: PropagationVector
    affected_nodes: list[str]
    base_frequency: float
    severity_distribution: SeverityDistribution
    severity_materialization: float | None = None
    legal_frameworks: dict | None = None
    exclusive_to_sectors: list[str] | None = None
    tail_multiplier: float = 1.0
    capability_threshold: float = 0.7
    threshold_multiplier: float = 3.0
    known_mitigations: dict[str, float] = field(default_factory=dict)
```

Figure 4 | Scenario class definition.

b) Dependency modeling across the AI supply chain

To capture correlation, aggregation, and concentration effects, we encode systemic dependencies using a supply-chain dependency graph. The graph represents three primary actor classes (i) **Foundation model providers**; (ii) **Application-layer AI services (AI SaaS)**; and (iii) **Downstream enterprises**.

Each node is associated with layer-specific risk scenarios and exposure characteristics. Edges encode dependency relationships through which losses may propagate—for example, when a failure or regulatory breach at the foundation model layer induces correlated losses across multiple downstream applications and enterprises.

This structure allows the model to represent correlated failures across actors, loss amplification through shared infrastructure and concentration risk arising from widely reused models or services.

c) Monte Carlo simulation and loss aggregation

We simulate annual losses using a Monte Carlo approach. Each simulation iteration proceeds as follows:

1. For each risk scenario, sample an annual event count from the corresponding frequency distribution.
2. Conditional on event occurrence, sample severities from the scenario-specific severity distribution.
3. Propagate losses through the dependency graph according to encoded relationships.
4. Aggregate propagated losses to obtain total annual loss at the portfolio or system level.

Formally as pseudo-code:

```
FUNCTION RunSimulationIteration
  FOR each scenario
    eventCount ← SampleEventCount(scenario)
    IF eventCount > 0 THEN
      severities ← SampleSeverities(eventCount, scenario)
    END IF
  END FOR

  losses ← PropagateLossesThroughDependencyGraph(severities)

  totalAnnualLoss ← Sum(losses)

  RETURN totalAnnualLoss
END FUNCTION
```

Repeating this process yields an empirical distribution of annual losses that reflects both individual risk behavior and systemic interactions.

d) Risk metrics and premium computation

From the simulated loss distribution, we compute standard insurance-relevant risk metrics, including:

- **Expected Loss (EL).**
- **Value at Risk (VaR)** at specified confidence levels.
- **Tail Value at Risk (TVaR).**

Premiums are then computed by combining expected loss with additional loadings reflecting tail risk, ambiguity, and expenses, producing pricing signals that can be directly linked to modeled mitigation measures.

e) Code, Reproducibility and Resources

All code, data processing pipelines, and simulation components used for scenario construction, dependency modeling, Monte Carlo simulation and risk metric computation are publicly

available on GitHub [\[link\]](#). The repository includes documentation and configuration files required to reproduce the experiments and regenerate all reported results.

The framework is designed to be modular and extensible. With minimal modification, it can be adapted to incorporate additional risk taxonomies, regulatory regimes, or supply-chain structures, enabling future research on AI risk propagation and pricing under evolving governance conditions.

An interactive [dashboard](#) is also provided to facilitate inspection, stress testing, and comparative analysis of modeled risk scenarios, supporting transparency and reuse.

Appendix C: Company’s mitigation measures

Safety Tools & Controls

Select the AI safety tools deployed by the company. These reduce risk exposure and lower premiums.

Configure Safety Tools

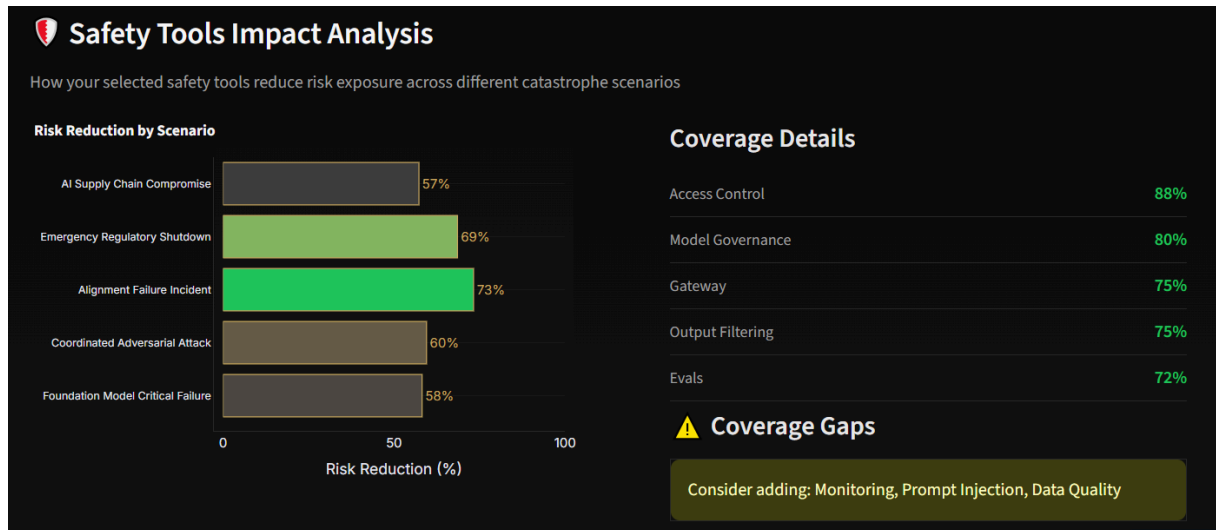
Prompt Injection Defense	AI Gateway
None	Kong AI Gateway
Runtime Monitoring	Output Filtering
None	Perspective API
Model Evaluations	Data Quality
Giskard	None
Access Control	Model Governance
Okta	MLflow

Safety Tools Deployed 5 / 8 **Computed Safety Score** 43.3%

Calculation Mode

☒ Quick Estimate ☐ Full Simulation

Calculate Premium



Appendix D: Supplementary Figures and Model Outputs

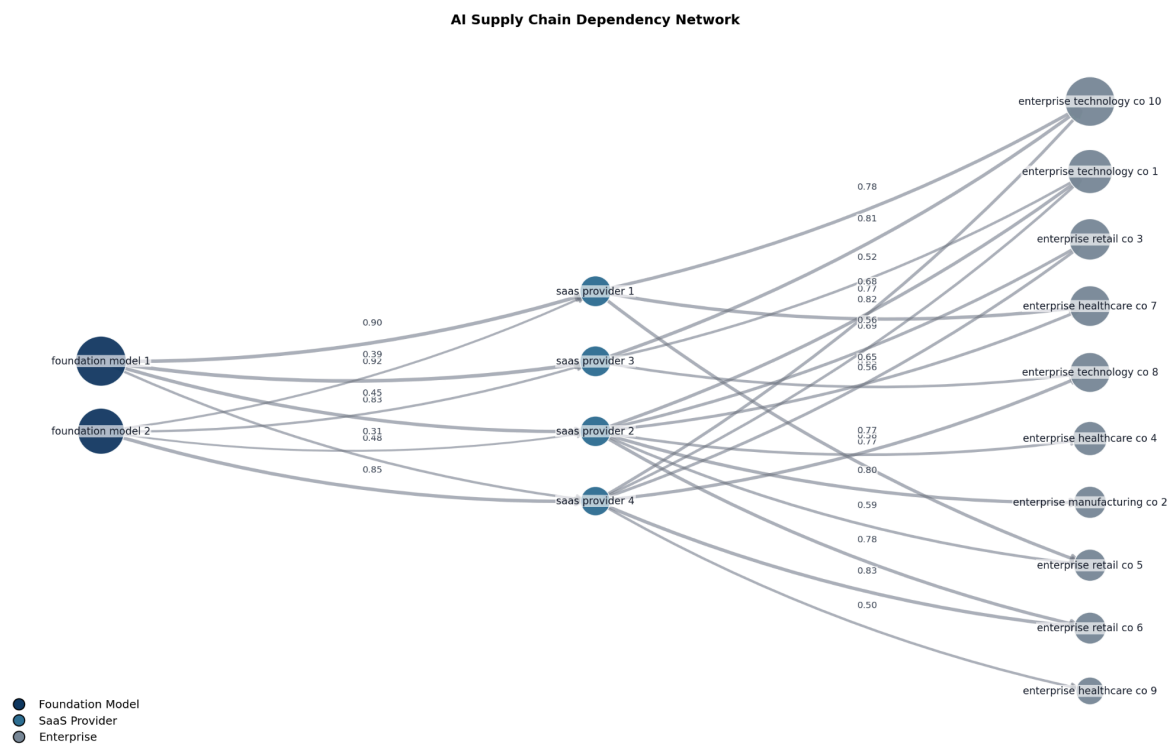


Figure 5 | AI Supply Chain Dependency Network.

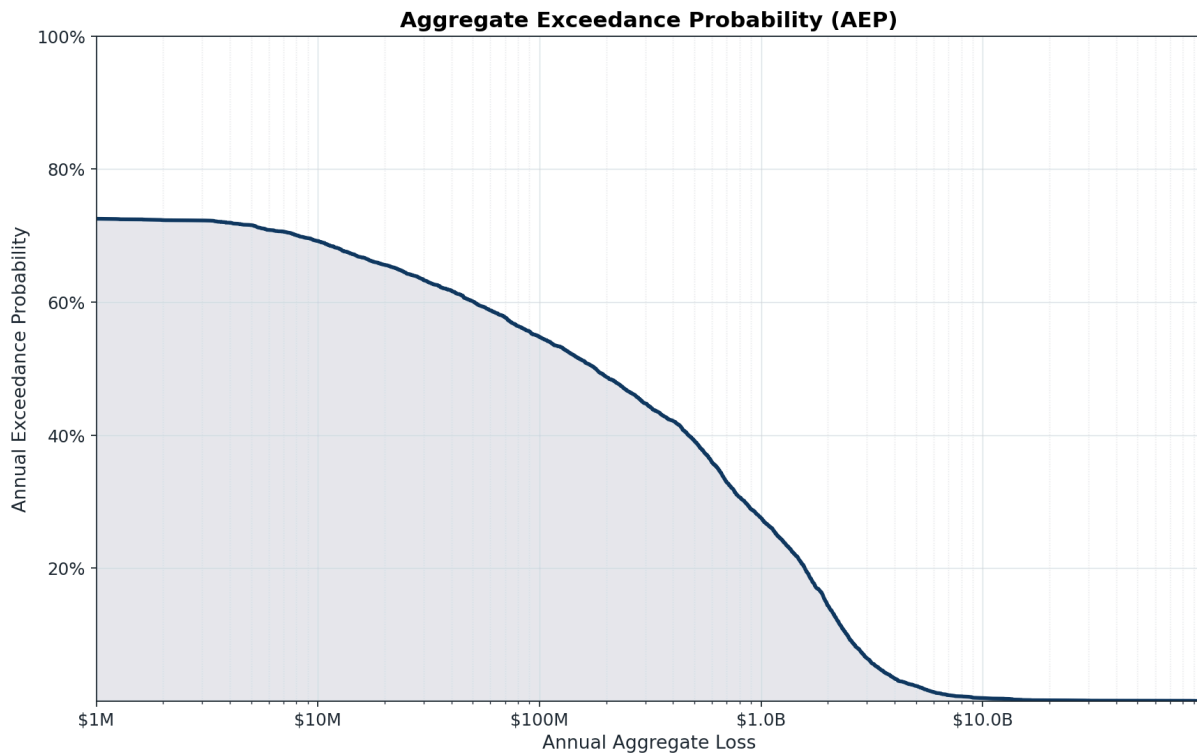


Figure 6 | Aggregate Exceedance Probability.

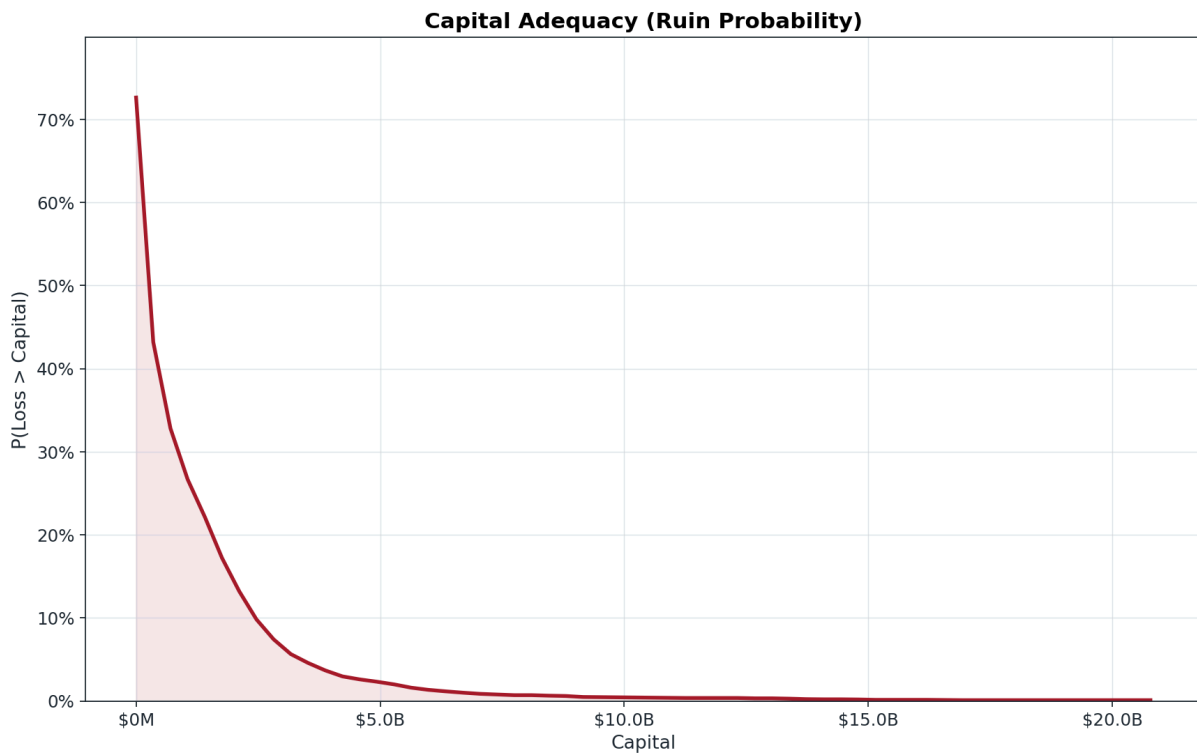


Figure 7 | Capital Adequacy. Ruin Probability.

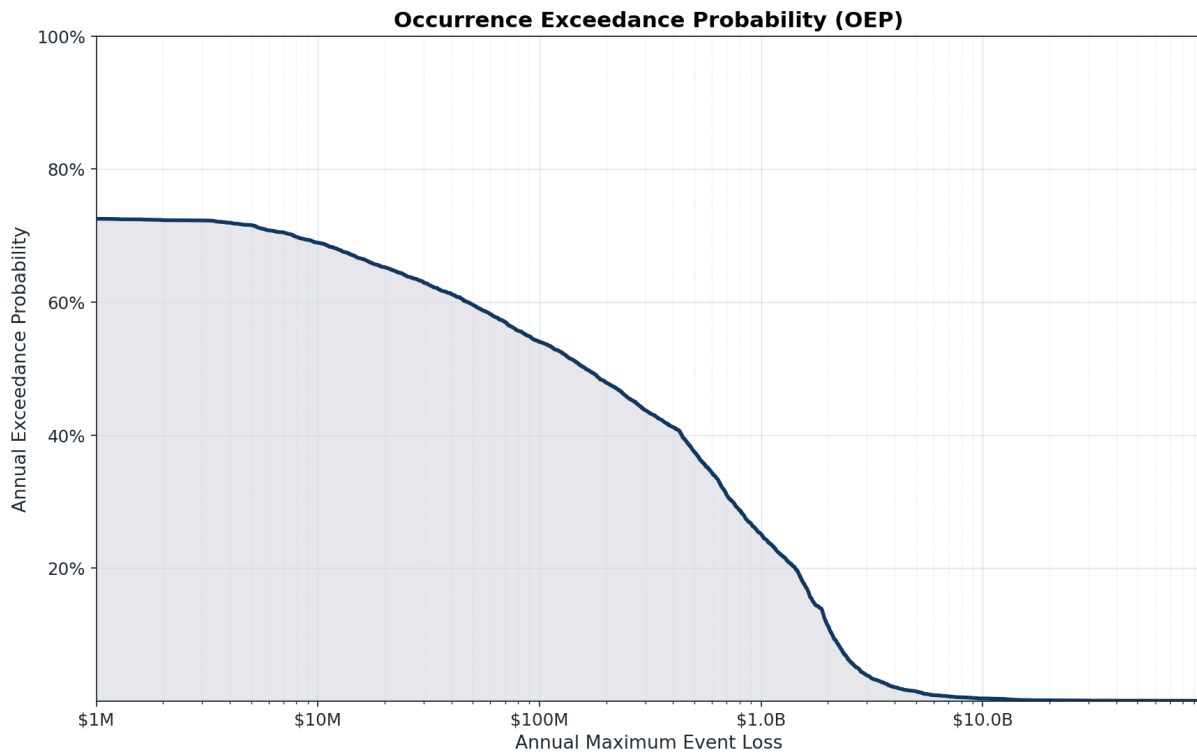


Figure 8 | Occurrence Exceedance Probability.

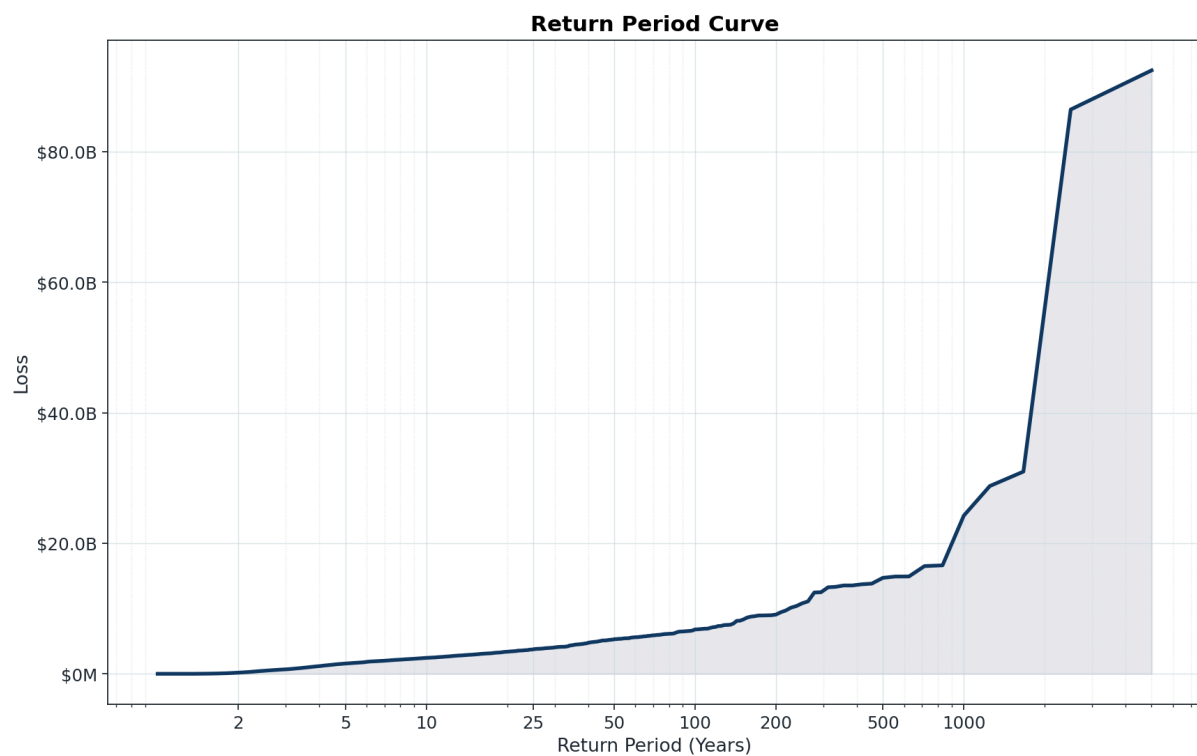


Figure 9 | Return Period Curve.

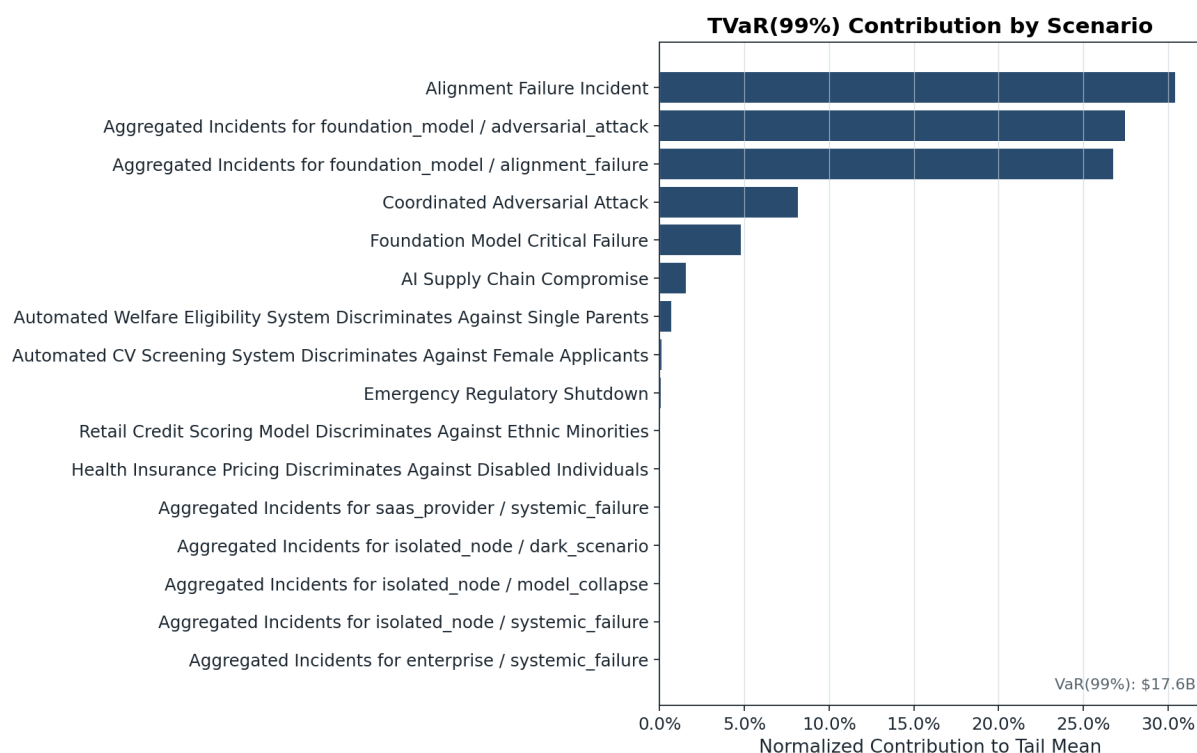


Figure 10 | TVaR (99%) Contribution by Scenario.

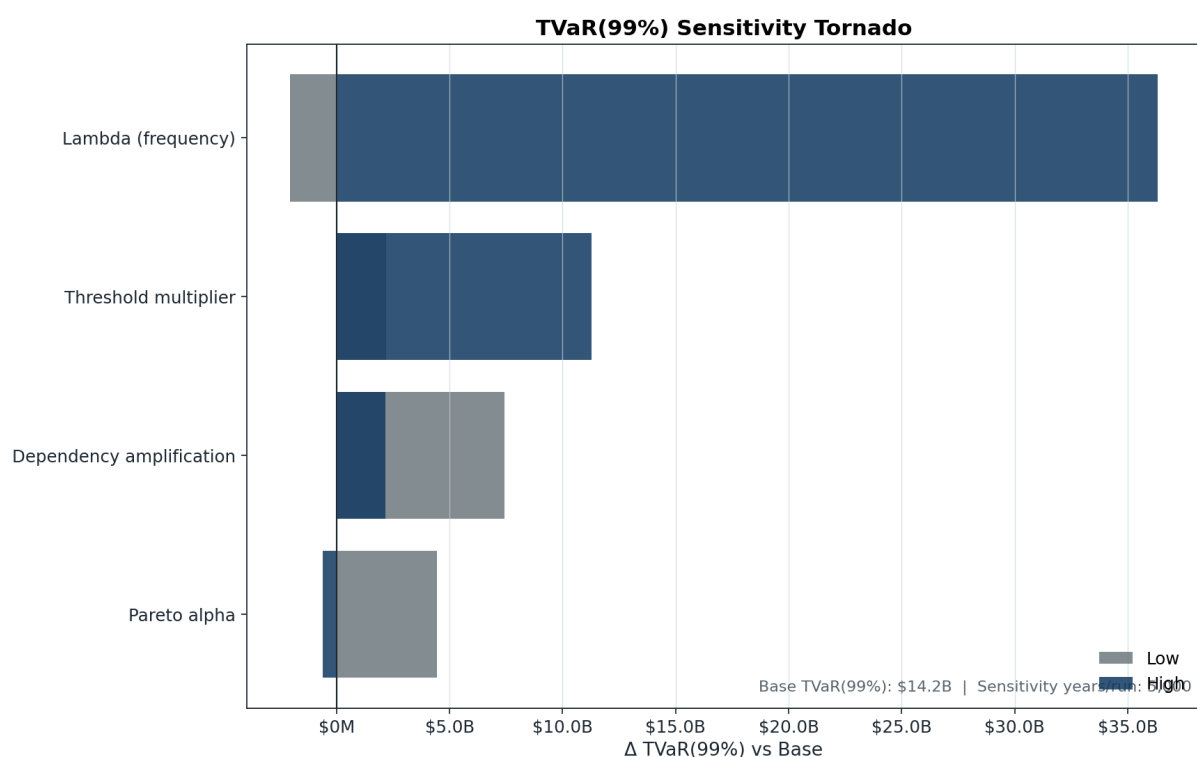


Figure 11 | TVaR (99%) Sensitivity Tornado.

Appendix E: Taxonomy EU AI Act.

ID	Name	Category	Layers	Sectors	Frequency	Severity
AIA-01a	Discriminatory or biased outcomes	Fundamental rights	foundation_model; saas_provider; enterprise	Employment; Finance; Insurance; Healthcare; Public sector	Medium	High
AIA-01b	Discriminatory or biased outcomes	Fundamental rights	saas_provider; enterprise	Employment; Finance; Insurance; Healthcare; Public sector	High	Very High
AIA-01c	Discriminatory or biased outcomes	Fundamental rights	enterprise	Employment; Finance; Insurance; Healthcare; Education; Housing	Very High	Very High
AIA-02	Discriminatory or biased outcomes	Fundamental rights	enterprise	Any regulated sector	High	High
AIA-03	Lack of transparency about AI interaction	Transparency	saas_provider; enterprise	Consumer services; Media; Customer support; Education	Very High	Medium-High
AIA-04	Undisclosed AI-generated or synthetic content	Transparency	saas_provider; enterprise	Media; Marketing; Political communication	Medium-High	High
AIA-05	Internal black-box usage of AI systems	Transparency	enterprise	Finance; Healthcare; Public sector; Human resources	High	High
AIA-06	Unsafe general-purpose model capabilities	Systemic risk	foundation_model	Cross-sectoral	Medium	Very High
AIA-07	Foreseeable downstream misuse of general-purpose models	Systemic risk	foundation_model	Cross-sectoral	High	High
AIA-08	Capability overreach through agentic or chained systems	Systemic risk	saas_provider; enterprise	Automation; Finance; Legal; Cybersecurity	Medium	High
AIA-09	Hallucinations in decision-support systems	Robustness and safety	saas_provider; enterprise	Healthcare; Legal; Finance; Public services	Very High	High
AIA-10	Model drift and performance degradation	Robustness and safety	enterprise	Any high-risk deployment	High	Medium-High
AIA-11	Adversarial vulnerabilities	Robustness and safety	foundation_model; saas_provider; enterprise	Security-sensitive sectors	Medium	High
AIA-11	Unlawful or low-quality training data	Data governance	foundation_model	Cross-sectoral	Medium	High
AIA-12	Contextual misuse of data at deployment	Data governance	saas_provider; enterprise	Human resources; Marketing; Finance	High	Medium-High
AIA-13	Shadow AI and uncontrolled internal usage	Data governance	enterprise	All sectors	Very High	Medium-High
AIA-14	Automation bias and over-reliance on AI	Human oversight	saas_provider; enterprise	Healthcare; Finance; Law enforcement	High	High
AIA-14	Unclear accountability across AI value chain	Accountability	enterprise	Complex supply chains	Medium-High	High
AIA-15	Social scoring of individuals	Prohibited practice	saas_provider; enterprise	Public sector; Finance	Low	Extreme
AIA-16	Real-time remote biometric identification outside legal exceptions	Prohibited practice	saas_provider; enterprise	Law enforcement	Low	Extreme

Appendix F: Taxonomy Digital Services Act.

ID	Name	Category	Layers	Sectors	Frequency	Severity
DSA-01a	Exposure of minors to harmful or age-inappropriate content	Protection of minors	saas_provider; enterprise	Social media; Video platforms; Gaming; Content sharing	Very High	Very High
DSA-01b	Exposure of minors to harmful or age-inappropriate content	Protection of minors	enterprise	Social media; Streaming; News aggregation	High	Very High
DSA-02	Addictive design patterns affecting minors	Protection of minors	enterprise	Social media; Gaming; Short-form video	Very High	High
DSA-03	Targeted advertising based on profiling of minors	Protection of minors	enterprise	Advertising; E-commerce; Social media	Medium-High	High
DSA-04	Inadequate age verification or age assurance mechanisms	Protection of minors	enterprise	Social media; Adult content; Gaming	High	High
DSA-05	Lack of transparency in recommender systems	Transparency	enterprise	Social media; Marketplaces; Content platforms	Very High	Medium-High
DSA-06	Dark patterns undermining user choice	User autonomy	enterprise	E-commerce; Social media; Subscription services	Very High	Medium-High
DSA-07	Insufficient ad transparency	Transparency	enterprise	Advertising; Social media; Marketplaces	High	Medium
DSA-08	Delayed or ineffective removal of illegal content	Illegal content	enterprise	Hosting services; Social media; Marketplaces	High	High
DSA-09	Abuse of notice-and-action mechanisms	Procedural fairness	enterprise	Social media; Marketplaces	Medium-High	Medium
DSA-10	Systemic amplification of harmful content	Systemic risk	enterprise	Social media; Video platforms	High	Very High
DSA-11	Algorithmic dissemination of disinformation at scale	Systemic risk	enterprise	Social media; News aggregation	Medium-High	Very High
DSA-12	Societal harm to minors at population scale	Systemic risk	enterprise	Social media; Video platforms	Medium	Extreme
DSA-13	Insufficient internal governance for DSA compliance	Accountability	enterprise	All online intermediaries	High	Medium-High
DSA-14	Lack of traceability and audit readiness	Accountability	enterprise	VLOPs; Large platforms	Medium	High