SECOND REVISION / AUGUST 2026

SYSTEMIC CATASTROPHE UNDER ARCHITECTURE AND AMBIGUITY

Dependence Architecture, Welfare Exposure, and Model Ambiguity

Component reliability does not determine systemic safety; dependence architecture does.

TAIL PRESSURE
$T = p\,\Delta$

ARCHITECTURE
$I_s(\alpha)$

ROBUST BOUNDARY
$\min\{I_s, R\} \gtrless g(\gamma - 1)$

Michael Schymura

Second revision, August 2026

RESEARCH CONTRIBUTION

Abstract

Systemic risk in a replicated technology is not determined by component reliability alone. This paper derives the probability of joint failure from a latent-factor architecture and connects the resulting large-deviation rate to welfare exposure and model ambiguity. Conditional on an architecture-wide factor, severe failures are Bernoulli; the factor itself obeys a large-deviation principle. Under an explicit uniform conditional tail condition, systemic-failure probability has exponent

$$I_s(\alpha) = \inf_z \{J_s(z) + d_\alpha(q_s(z))\}.$$

The expression identifies the least-cost combination of an adverse common state and an unusual concentration of local failures. A Gaussian–logistic primitive supplies conditions under which the uniform approximation holds.

The economic result follows by composing this architecture rate with two objects that are usually analyzed separately. If residual welfare after systemic failure contracts at rate g under fixed-person CRRA utility with curvature $\gamma > 1$, nominal tail pressure vanishes only when $I_s(\alpha) > g(\gamma - 1)$. If probability laws are instead evaluated in a binary Kullback–Leibler neighborhood whose radius contracts at rate R, the effective safety exponent becomes $\min\{I_s(\alpha), R\}$. Architecture can therefore improve nominal safety without improving robust welfare at the same rate.

The large-deviation machinery and binary relative-entropy expansions are established tools. The contribution is their economic composition when architecture is a choice variable, welfare exposure grows with scale, and evidential ambiguity may contract more slowly than nominal risk. Exact finite-system quadrature and binary-KL optimization converge to the predicted limits for the stated primitive. These computations are theorem bridges, not estimates of catastrophic AI risk.

Keywords: systemic risk; common-factor models; large deviations; distributional robustness; catastrophic risk; artificial intelligence
JEL codes: D61; D62; D81; C61; O33

Scope of the result. The paper proves an architecture–welfare–ambiguity boundary and evaluates its finite-system convergence for one explicit primitive. It does not estimate a real-world probability of catastrophic AI failure. The distinction is substantive: numerical agreement with a theorem validates an implementation, not the theorem's empirical premises.

Argument map

SECTION 1

Introduction

A small probability attached to a large loss is not yet a model of catastrophic risk. The description leaves out the joint rate at which probability falls and consequence grows, whether the probability is known, whether today's action changes both objects, and whether the event destroys the institutions that would otherwise learn from it. Those omissions are secondary for ordinary insurance. Near a welfare boundary, they decide the result.

The classical contamination experiment is simple. A normal lottery is assigned probability $1 - p$, while a catastrophic payoff c receives probability p. If Bernoulli utility is bounded below, the catastrophic term disappears as $p \downarrow 0$. If utility is unbounded below, one can construct paths along which an arbitrarily small probability drives the certainty equivalent to its lower endpoint. Earlier work identified these polar possibilities and the associated normative discomfort (Buchholz and Schymura, 2012). The mathematical statement, however, is existential. Lower unboundedness proves that a tyrannical path *exists*; it does not prove that every scientifically admissible path is tyrannical.

The distinction opens a middle regime. Once probability and severity are connected by a specified path, expected utility depends on their product in utility units. That product may vanish, converge to a finite positive value, or diverge. I call it *catastrophic tail pressure*. The concept does not rescue expected utility by assumption. It moves the real dispute to where it belongs: the joint physical, statistical, and ethical model that determines the path.

Artificial intelligence makes the distinction unusually consequential because deployment multiplies systems without necessarily multiplying independent failure modes. Applications can share a foundation model, cloud provider, identity layer, monitoring stack, training data, or theory of control. Component reliability may improve while the least unlikely route to joint failure becomes cheaper. The relevant object is therefore not a free-standing "probability of doom." It is the probability generated by a specified dependence architecture and evaluated against an explicit welfare path.

> **The paper's central result.** Architecture determines a nominal systemic-safety rate $I_s(\alpha)$. Model ambiguity can reduce it to $\min\{I_s(\alpha), R\}$. Welfare remains asymptotically protected only when that effective rate outruns exposure growth in utility units. Even then, rollback leaves a legacy floor whenever a latent trigger may already exist.

The paper makes one economic contribution. It composes an architecture-generated systemic-failure rate, a welfare-exposure rate, and an ambiguity-contraction rate in a single phase boundary. The probability theory is closely related to common-factor large-deviation models; the new economic object is the ordering of these three rates when architecture is a safety choice rather than a fixed stochastic specification.

Three supporting results make that composition explicit. First, a pathwise tail-pressure identity clarifies the relationship to Buchholz and Schymura (2012): the earlier article established susceptibility over possible probability–severity paths, whereas the present paper conditions on a specified path and then endogenizes its probability side. Second, a Gaussian–logistic architecture derives the uniform conditional approximation used by the general theorem. Third, exact finite-N calculations connect the asymptotic rate to observable system counts without

treating the calculation as empirical calibration.

The argument proceeds from the pathwise welfare statistic to ambiguity, architecture, and the combined boundary. A compact symmetric game then identifies the familiar private–social wedge when firms internalize only part of continuation loss. The final sections report the finite-system theorem bridge, delimit what it establishes, and state the evidence required to apply the framework.

SECTION 2

What the literature establishes—and what it does not

Three literatures meet in this problem but rarely share the same state space.

2.1 Catastrophic tails and well-posed welfare

The dismal-theorem debate showed that fat-tailed uncertainty can make conventional welfare calculations extraordinarily sensitive (Weitzman, 2009; Pindyck, 2011; Ackerman et al., 2010). Rare-disaster macroeconomics demonstrated, by contrast, that large risk prices can coexist with finite expected utility on calibrated domains (Barro, 2009; Gabaix, 2012). The compatibility program made the mathematical point explicit: a utility function and a tail law cannot be chosen independently (Ikefuji et al., 2015, 2020). Bounded utility is sufficient over unrestricted lottery classes, but it is not necessary on restricted families. Conversely, finite expected utility at each parameter value does not guarantee continuity as the family approaches a catastrophic boundary.

Direct responses to the catastrophic-risk problem construct catastrophe-sensitive robust preferences and non-expected-utility indicators with certainty-equivalent elasticities (Grechuk and Zabarankin, 2014; Geiger, 2024). Those papers rule out any claim that multiplying probability and severity, or naming an elasticity, is by itself a new decision theory. The present paper retains expected utility as a transparent benchmark and seeks its incremental content in the architecture-generated probability rate, ambiguity-rate erosion, and the dynamic legacy mechanism.

The present paper differs from a pure boundedness test. Its unit of analysis is a *joint path* of event probabilities and conditional welfare distributions. This retains the domain discipline while showing exactly when a finite middle response arises.

2.2 Dynamic catastrophes, ambiguity, and misspecification

Dynamic climate models introduce tipping, learning, precautionary capital, and multiple interacting catastrophes (Cai and Lontzek, 2019; Martin and Pindyck, 2015; Liski and Salanié, 2026). Their central lesson is that prevention values cannot be added state by state when survival and consumption are shared across hazards. Irreversibility creates an option value of waiting (Arrow and Fisher, 1974); endogenous information can create an option value of controlled use. Which force dominates depends on where information comes from and whether the action can be reversed.

Risk, ambiguity among structured models, and misspecification of every structured model are distinct layers (Aydogan et al., 2023; Hansen and Sargent, 2022; Cerreia-Vioglio et al., 2026).

Robust control disciplines model deviations through an explicit penalty (Hansen and Sargent, 2011; Zhao et al., 2023). It does not guarantee well-posedness. Entropic criteria require an exponential moment of negative utility, a stronger condition than ordinary integrability.

2.3 AI growth, adoption, and catastrophe

Task experiments identify sizable productivity effects in bounded workplace settings (Brynjolfsson et al., 2025); macroeconomic projections remain sensitive to task shares, bottlenecks, diffusion, and general-equilibrium adjustment (Acemoglu, 2025; Jones, 2026). A growing economics literature now places transformative growth and existential risk in one model (Jones, 2024, 2025; Growiec and Prettner, 2026; Trammell and Aschenbrenner, 2024; Trammell and Korinek, 2026). Adoption studies show why gradual deployment, learning by doing, reversibility, and private information can reverse policy rankings (Acemoglu and Lensman, 2024; Gans, 2025; Guerreiro et al., 2023).

The evidence frontier is improving but not yet an identified catastrophe map. The International AI Safety Report synthesizes rapid capability gains and substantial scientific uncertainty (Bengio et al., 2026). METR's time-horizon measures and public AISI evaluations are informative leading indicators within tested domains (METR, 2026; AI Security Institute, 2025); they are not cardinal measures of welfare loss or repeated-event estimates of extinction. Financial-market evidence can reveal changes in investors' combined beliefs about growth and tails only under a structural asset-pricing model (Andrews and Farboodi, 2025).

The gap is therefore precise. Economics has models of growth, adoption, ambiguity, and rare disasters, while evaluation science has changing proxies for capability. What is missing is a disciplined map from those proxies through deployment and dependence to an identified social-welfare tail. This paper provides the model architecture and the falsification conditions, not an invented point estimate.

2.4 Closest common-factor literature and contribution boundary

The closest mathematical comparators are multifactor portfolio-credit models. Frey and McNeil (2003) model dependent defaults through Bernoulli mixtures, while Glasserman et al. (2007) derive logarithmic aggregate-loss limits in a multifactor Gaussian credit model as portfolio size grows and losses become rare. Those papers already establish that conditional independence does not remove systemic dependence and that a common factor can determine aggregate tails. Table 1 states the boundary between those results and the present paper.

Table 1. Relationship to the closest common-factor literature

Object	Established comparator	Role in this paper
Conditional dependence	Bernoulli-mixture defaults generate dependence through latent factors (Frey and McNeil, 2003)	Severe system failures use the same conditional-independence logic; no new claim attaches to that construction
Aggregate tail	Multifactor Gaussian credit models yield logarithmic portfolio-loss limits under growing-portfolio and rare-default regimes (Glasserman et al., 2007)	A critical failed-system share is indexed by an architecture choice, with the required uniform conditional rate stated explicitly
Economic boundary	Portfolio loss is evaluated within the credit-risk objective	The architecture exponent is composed with welfare-exposure and ambiguity-contraction rates
Decision margin	Factor exposure is a portfolio characteristic or model input	Shared technical dependencies become an object of hardening and a source of a private–social wedge

The comparison is functional rather than literal: credit losses and severe system failures have different observables, institutions, and recovery processes. What carries over is the conditional-factor logic. What is added is the economic ordering of architecture, exposure, and ambiguity rates when architecture becomes a design choice.

SECTION 3

General catastrophic tail pressure

3.1 Notation and units

The general welfare results use n for an arbitrary probability–severity sequence. The architecture result uses N for the number of deployed systems. Every exponent in the core architecture–welfare–ambiguity boundary is measured per deployed system along that same sequence; none is an annual hazard rate. The later dynamic model uses calendar time t and is a separate finite-horizon extension. Comparing its time rates with the per-system rates would require an explicit scaling relation between time and system count.

Table 2. Reserved notation in the core architecture model

Symbol	Meaning	Unit or scale
N	Number of deployed systems in the asymptotic sequence	system count
s	Safety architecture index	design choice
α	Systemic-event threshold	failed-system share
Z_N	Architecture-wide factor	primitive-dependent
$q_s(z)$	Conditional severe-failure probability	probability
$I_s(\alpha)$	Nominal systemic-failure exponent	per system
R	KL-radius contraction exponent	per system
g	Residual-welfare exposure exponent	per system

3.2 Environment

Let $u : \mathbb{R}_{++} \to \mathbb{R}$ be continuous and strictly increasing. Concavity is imposed only when risk aversion is interpreted. For each n, let G_n be a normal conditional distribution and H_n a catastrophic conditional distribution on $\mathbb{R}_{++}$. Define

$$A_n = \int u(x)\, G_n(\mathrm{d}x), \qquad B_n = \int u(x)\, H_n(\mathrm{d}x), \qquad \Delta_n = A_n - B_n, \tag{1}$$

whenever the integrals are finite. The contaminated law is

$$P_n = (1 - p_n)G_n + p_n H_n, \qquad p_n \in (0, 1), \tag{2}$$

with expected utility

$$V_n = (1 - p_n)A_n + p_n B_n = A_n - T_n, \qquad T_n := p_n \Delta_n. \tag{3}$$

Definition 3.1 Catastrophic tail pressure. T_n is the probability-weighted conditional utility gap between the normal and catastrophic branches. A path is *negligible* if $T_n \to 0$, *interior* if $T_n \to \lambda \in (0, \infty)$, and *dominant* if $T_n \to \infty$.

The definition includes severity distributions, not merely a scalar worst state. It also permits the normal branch to change with technology, provided its expected utility has a limit.

Theorem 3.2 General pathwise limit theorem. *Suppose $A_n \to A \in u(\mathbb{R}_{++})$, $\Delta_n \geq 0$ eventually, and $u(\mathbb{R}_{++})$ has lower endpoint $\underline{u} = -\infty$. Extend the inverse by $u_e^{-1}(-\infty) = 0$. For certainty equivalent $m_n = u_e^{-1}(V_n)$,*

$$\liminf_{n\to\infty} m_n = u_e^{-1}\left(A - \limsup_{n\to\infty} T_n\right), \tag{4}$$

$$\limsup_{n\to\infty} m_n = u_e^{-1}\left(A - \liminf_{n\to\infty} T_n\right). \tag{5}$$

Consequently, m_n converges if and only if T_n converges in $[0, \infty]$. Conditional on convergence, there are three regimes: $T_n \to 0$ implies $m_n \to u^{-1}(A)$; $T_n \to \lambda \in (0, \infty)$ implies $m_n \to u^{-1}(A - \lambda)$; and $T_n \to \infty$ implies $m_n \to 0$.

Proof. Equation (3) and $A_n \to A$ give $\liminf V_n = A - \limsup T_n$ and $\limsup V_n = A - \liminf T_n$. The extended inverse is continuous and increasing, which gives the displayed identities. The three convergent cases follow immediately. Conversely, strict monotonicity of the extended inverse implies that convergence of m_n is equivalent to convergence of V_n and hence of $T_n = A_n - V_n$. □

Corollary 3.3 Affine invariance. *For every positive affine representation $\widetilde{u} = a + bu$, $b > 0$, one has $\widetilde{\Delta}_n = b\Delta_n$ and $\widetilde{T}_n = bT_n$. The three regimes are invariant, and the certainty equivalent in payoff units is unchanged.*

Tail pressure is cardinal in utility units, but its regime is not an artifact of the chosen von Neumann–Morgenstern scale. A reported finite value of T_n therefore requires a normalization; a reported regime does not.

3.3 Regular variation and the critical boundary

Let catastrophe severity be indexed by $c \downarrow 0$. Write $p(c)$ for catastrophe probability and $\Delta(c)$ for the conditional utility gap. Suppose

$$p(c) = c^{\beta} L_p(1/c), \qquad \Delta(c) = c^{-\rho} L_\Delta(1/c), \tag{6}$$

where $\beta > 0$, $\rho \geq 0$, and the L functions are measurable, eventually positive, and slowly varying.

Theorem 3.4 Regular-variation phase boundary. *Under* (6),

$$T(c) = c^{\beta-\rho} L_p(1/c) L_\Delta(1/c). \tag{7}$$

Hence $T(c) \to 0$ if $\beta > \rho$, $T(c) \to \infty$ if $\beta < \rho$, and the product of slowly varying terms determines the asymptotic behavior when $\beta = \rho$; at that boundary it may tend to zero, a finite positive value, infinity, or fail to converge.

Proof. Multiplication gives the displayed expression. A positive power dominates every slowly varying function at infinity, while a negative power dominates in the opposite direction. At equality, no power term remains. □

These are standard consequences of Karamata theory (Bingham et al., 1987). The contribution here is not regular variation itself, but its connection to the endogenous architecture rate derived below.

The critical line is thin in parameter space but not economically irrelevant. Estimated exponents are uncertain, and lower-order terms decide policy precisely when confidence intervals overlap the boundary.

For differentiable paths define

$$\beta(c) = \frac{c p'(c)}{p(c)}, \qquad \rho(c) = -\frac{c \Delta'(c)}{\Delta(c)}. \tag{8}$$

Then $\mathrm{d}\log T / \mathrm{d}\log c = \beta(c) - \rho(c)$. A uniform positive gap implies neglect; a uniform negative gap implies dominance. A vanishing local gap is insufficient because its integral over log severity may diverge in either direction.

3.4 CRRA and logarithmic utility

Use normalized CRRA utility

$$u_\gamma(c) = \begin{cases} \dfrac{c^{1-\gamma} - 1}{1-\gamma}, & \gamma \neq 1, \\ \log c, & \gamma = 1. \end{cases} \tag{9}$$

For $\gamma > 1$, $\Delta(c) \sim c^{1-\gamma}/(\gamma - 1)$. If $c = p^b$, then

$$T(p) \sim \frac{1}{\gamma - 1} p^{1-b(\gamma-1)}. \tag{10}$$

Corollary 3.5 CRRA boundary. *For $\gamma > 1$ and $c = p^b$, tail pressure vanishes when $b < 1/(\gamma - 1)$, converges to a finite positive constant at equality, and diverges when $b > 1/(\gamma - 1)$. For $\gamma < 1$, utility is bounded below and the same polynomial paths are negligible. For log utility, $p[A - b \log p] \to 0$ for every $b > 0$, although sufficiently fast non-polynomial severity paths can remain dominant.*

The result is not a calibration of ethical curvature. It shows why the curvature parameter and the empirical probability–severity map must be reported together.

3.5 Domain continuity and the true pathology

Proposition 3.6 Uniform-integrability domain. *Let $X_n \to X$ in probability and let u be continuous. If $\{u(X_n)\}_{n \geq 1}$ is uniformly integrable, then*

$$\mathbb{E}[u(X_n)] \longrightarrow \mathbb{E}[u(X)]. \tag{11}$$

If u is strictly increasing and the limiting expectation lies in the interior of its range, certainty equivalents converge as well.

Proof. The continuous mapping theorem gives $u(X_n) \to u(X)$ in probability. Uniform integrability and Vitali's theorem imply convergence in L^1, hence convergence of expectations. □

This standard proposition is not claimed as novel. Prospect-space topology and integrability-restricted domains already formalize the distinction (Dentcheva and Ruszczynski, 2013). For catastrophic lower tails, uniform integrability of the negative parts is the operative condition once positive parts are controlled. The result yields a disciplined middle position: lower-unbounded utility can be continuous on an admissible domain, but the domain restriction must come from science or a transparent robustness neighborhood. Selecting a convenient tail envelope after seeing the policy result is not identification.

SECTION 4

How alternative decision criteria transform the tail

The failure of one expected-utility specification does not imply that another criterion is automatically well posed. Each criterion changes the effective race between probability and welfare loss.

4.1 Multiple priors and an ambiguity floor

Suppose the evaluator considers catastrophe probabilities in a nonempty compact set $\mathcal{P}_n \subset [0, 1]$ and applies maxmin expected utility (Gilboa and Schmeidler, 1989). Let $\overline{p}_n = \sup \mathcal{P}_n$. When the catastrophic branch has lower conditional expected utility than the normal branch,

$$V_n^{\mathrm{MP}} = A_n - \overline{p}_n \Delta_n. \tag{12}$$

Proposition 4.1 Upper-envelope pressure. *Theorem 3.2 applies under maxmin preferences with effective tail pressure* $\overline{T}_n = \overline{p}_n \Delta_n$. *If* $\liminf_n \overline{p}_n > 0$ *while* $\Delta_n \to \infty$, *then maxmin value diverges downward even when a reference estimate tends to zero.*

An ambiguity floor is therefore operational. Evidence that moves a median expert forecast while leaving the upper scientifically defensible envelope unchanged does not relax the maxmin constraint. The relevant empirical question is what observation contracts the model set.

4.2 Probability weighting

For a genuinely binary reduced lottery whose two outcomes are the normal and catastrophic branch indices, rank-dependent evaluation takes the form (Quiggin, 1982)

$$V_n^w = A_n - w(p_n)\Delta_n, \tag{13}$$

where $w(0) = 0$ and w is increasing. If $w(p) \sim Kp^\alpha$ near zero and objective probability has regular-variation exponent β, the effective boundary becomes

$$\alpha\beta = \rho. \tag{14}$$

Small-probability overweighting ($\alpha < 1$) enlarges the dominant region. Underweighting shrinks it. For the original compound conditional distributions this reduction is generally invalid: their full outcome ranks, not only branch expected-utility indices, determine rank-dependent value. The label "behavioral" determines neither sign nor magnitude; the local weighting exponent does.

For Prelec weighting $w(p) = \exp\{-a[-\log p]^d\}$, a polynomial utility loss is dominated by the decision weight when $d > 1$ and dominates it when $d < 1$; coefficients decide the $d = 1$ boundary (Prelec, 1998). This illustrates a recurring point: apparently modest functional-form choices can alter asymptotic policy.

4.3 Smooth ambiguity away from and near the boundary

In the smooth-ambiguity model of Klibanoff et al. (2005), let models $m = 1, \ldots, M$ have second-order weights $\mu_m > 0$, $\sum_m \mu_m = 1$, and model-specific expected utility

$$V_{mn} = A_n - p_{mn}\Delta_n. \tag{15}$$

For a strictly increasing twice continuously differentiable ambiguity aggregator ϕ, define

$$V_n^{\mathrm{SA}} = \phi^{-1}\left(\sum_{m=1}^{M} \mu_m \phi(V_{mn})\right). \tag{16}$$

Proposition 4.2 Local smooth-ambiguity equivalence. *If* $A_n \to A$, $\max_m p_{mn}\Delta_n \to 0$, *and* $\phi'(A) > 0$, *then*

$$V_n^{\mathrm{SA}} = A_n - \sum_{m=1}^{M} \mu_m p_{mn}\Delta_n + O\left(\max_m [p_{mn}\Delta_n]^2\right). \tag{17}$$

If some model-specific pressure does not vanish, this first-order reduction need not hold; curvature of ϕ *and the full distribution across models matter.*

Proof. Apply a second-order Taylor expansion of $\phi(A_n - z)$ uniformly over the finite model set, average, and apply a first-order expansion of ϕ^{-1} around $\phi(A_n)$. The linear coefficients $\phi'(A_n)$ cancel. □

Smooth ambiguity is therefore locally equivalent to average model pressure only in the negligible neighborhood. Using that approximation to evaluate a boundary case assumes the conclusion.

4.4 Entropy robustness and exponential pressure

Consider a reference binary law with normal utility A and catastrophic utility $B = A - \Delta$. A multiplier evaluator solves

$$R_\theta(p, \Delta) = \inf_{Q \ll P} \left\{ \mathbb{E}_Q[U] + \theta D_{\mathrm{KL}}(Q \| P) \right\} = -\theta \log \mathbb{E}_P[e^{-U/\theta}], \tag{18}$$

where $\theta > 0$. Factoring out the normal utility yields

$$R_\theta(p, \Delta) = A - \theta \log \left(1 - p + p e^{\Delta/\theta} \right). \tag{19}$$

Define exponential tail pressure $Z_n = p_n e^{\Delta_n/\theta}$.

Theorem 4.3 Robust tail-pressure trichotomy. *If $p_n \to 0$, then*

$$R_{\theta,n} \to \begin{cases} A, & Z_n \to 0, \\ A - \theta \log(1 + z), & Z_n \to z \in (0, \infty), \\ -\infty, & Z_n \to \infty. \end{cases} \tag{20}$$

The adverse catastrophe probability is

$$q_n^* = \frac{p_n e^{\Delta_n/\theta}}{1 - p_n + p_n e^{\Delta_n/\theta}}. \tag{21}$$

Proof. Substitute Z_n into (19) and use $1 - p_n \to 1$. The probability formula follows from exponential tilting. □

Expected utility asks whether probability outruns a utility gap. Entropic robustness asks whether it outruns the *exponential* of that gap. The robust value is finite exactly when the corresponding exponential moment exists. Robustness is valuable precisely because it tests misspecification; it should not conceal that its neighborhood can imply a worst-case probability close to one. Reporting q^* alongside R_θ makes this diagnostic visible.

Multiplier preferences and constrained distributional robustness are related but not identical. The latter produces a sharper result for a systemic event whose nominal probability itself has an exponential architecture rate. Let

$$p_N = e^{-IN + o(N)}, \qquad r_N = e^{-RN + o(N)}, \tag{22}$$

and define the worst catastrophe probability in a binary KL ball,

$$\overline{p}_N = \sup \left\{ q \in [0, 1] : D_{\mathrm{KL}}(\mathrm{Bern}(q) \| \mathrm{Bern}(p_N)) \le r_N \right\}. \tag{23}$$

Theorem 4.4 Ambiguity-rate erosion. *For $I, R > 0$,*

$$-\lim_{N\to\infty} \frac{1}{N} \log \overline{p}_N = \min\{I, R\}. \tag{24}$$

If $R < I$, then

$$\overline{p}_N \sim \frac{r_N}{\log(r_N/p_N)} \sim \frac{r_N}{(I-R)N}. \tag{25}$$

If $R > I$, then $\overline{p}_N - p_N \sim \sqrt{2p_N r_N}$ and $\overline{p}_N/p_N \to 1$. At $R = I$, the exponent remains I and subexponential terms determine the inflation.

Proof. When $R < I$, $p_N/r_N \to 0$ exponentially and the optimizer satisfies $\overline{p}_N \gg p_N$. Uniformly on the relevant boundary, binary relative entropy obeys

$$D_{\mathrm{KL}}(q\|p_N) = q\log(q/p_N) - q + p_N + o(q). \tag{26}$$

Solving the boundary equation gives the displayed equivalences and exponent R. When $R > I$, $r_N/p_N \to 0$ exponentially. The local expansion

$$D_{\mathrm{KL}}(q\|p_N) \sim \frac{(q-p_N)^2}{2p_N} \tag{27}$$

gives $\overline{p}_N - p_N \sim \sqrt{2p_N r_N}$ and $\overline{p}_N/p_N \to 1$.

It remains to check equality. Feasibility gives $\overline{p}_N \geq p_N$, so its decay exponent cannot exceed I. For any $\varepsilon > 0$, a candidate $q_N \geq e^{-(I-\varepsilon)N}$ satisfies

$$q_N \log(q_N/p_N) \geq e^{-(I-\varepsilon)N}[\varepsilon N + o(N)], \tag{28}$$

which is exponentially larger than $r_N = e^{-IN+o(N)}$. Such a candidate is eventually infeasible. Hence the decay exponent is at least $I - \varepsilon$ for every $\varepsilon > 0$, proving the equality case. □

The theorem makes calibration of the ambiguity radius economically visible. A fixed or subexponentially shrinking KL radius has $R = 0$ and can erase every positive nominal large-deviation exponent. A statistically shrinking radius with $R > I$ preserves the nominal rate. Modern distributionally robust optimization provides the broader ambiguity-set context (Kuhn et al., 2025); the incremental economic result is the rate comparison in (24).

SECTION 5

Dynamic capability, safeguards, and irreversible legacy

Static tail pressure identifies the welfare boundary but not the optimal path. Advanced technology changes benefits, hazards, information, and the state from which future choices are made.

5.1 State dynamics and catastrophe time

Let the pre-catastrophe state be $X_t = (K_t, S_t, L_t)$: deployed capability, safety capital, and legacy exposure. Controls are capability investment $a_t \geq 0$ and safety effort $i_t \geq 0$. Before catastrophe,

$$\mathrm{d}K_t = (a_t - \delta_K K_t)\,\mathrm{d}t + \sigma_K(K_t)\,\mathrm{d}W_t^K, \tag{29}$$

$$\mathrm{d}S_t = (i_t - \delta_S S_t)\,\mathrm{d}t + \sigma_S(S_t)\,\mathrm{d}W_t^S, \tag{30}$$

$$\mathrm{d}L_t = (\xi a_t - \mu L_t)\,\mathrm{d}t + \sigma_L(L_t)\,\mathrm{d}W_t^L. \tag{31}$$

Legacy increases when capability is released into persistent systems, copied weights, unresolved access, or other states that cannot be instantly recalled. Catastrophe arrives at stopping time τ with intensity

$$\lambda(X_t, a_t, i_t) = \lambda_0 \exp\{\alpha_K K_t - \alpha_S S_t + \alpha_L L_t + \alpha_a a_t - \alpha_i i_t\}. \tag{32}$$

This log-linear form is a transparent benchmark, not an empirical estimate.

Let pre-catastrophe flow surplus be

$$f(X, a, i) = b(K) - c_a(a) - c_i(i) - d(L), \tag{33}$$

and let $W_C(X)$ be the value at catastrophe. The planner's value is

$$V(x) = \sup_{a,i} \mathbb{E}_x \left[\int_0^\tau e^{-rt} f(X_t, a_t, i_t)\,\mathrm{d}t + e^{-r\tau} W_C(X_{\tau^-}) \right]. \tag{34}$$

Assumption 5.1 Regular control problem. *The control set is compact or costs are coercive; coefficients are locally Lipschitz with linear growth; f is continuous and concave in controls; c_a, c_i are strictly convex; and the hazard is nonnegative and locally bounded. Growth bounds on f and W_C, together with discounting and a transversality condition, make the objective finite from above uniformly over admissible policies. The dynamic-programming principle and comparison for the resulting HJB are assumed to hold.*

Under standard dynamic-programming conditions, the pre-catastrophe value is the viscosity solution of

$$rV(x) = \sup_{a,i} \left\{ f(x, a, i) + \mathcal{L}^{a,i} V(x) + \lambda(x, a, i)[W_C(x) - V(x)] \right\}, \tag{35}$$

where $\mathcal{L}^{a,i}$ is the diffusion generator absent catastrophe.

5.2 Marginal conditions and continuation-value pressure

At an interior differentiable solution, let $D(x) = V(x) - W_C(x) > 0$ denote the continuation-value gap. The first-order conditions are

$$c_a'(a^*) = V_K + \xi V_L - \alpha_a \lambda D, \tag{36}$$

$$c_i'(i^*) = V_S + \alpha_i \lambda D. \tag{37}$$

Capability effort earns the shadow value of capability and legacy, but direct hazard growth subtracts $\alpha_a \lambda D$. Safety effort earns its state value plus the avoided instantaneous tail load $\alpha_i \lambda D$.

Proposition 5.2 Static comparative statics inside the HJB. *Hold the state, value derivatives, and hazard shifters fixed. If $c_i'' > 0$ and $\lambda = \bar{\lambda} e^{-\alpha_i i}$, the unique interior solution to (37) is strictly increasing in D. If $c_a'' > 0$ and $\lambda = \bar{\lambda} e^{\alpha_a a}$, the unique interior solution to (36) is strictly decreasing in D.*

Proof. For safety, write $F(i, D) = c_i'(i) - V_S - \alpha_i \bar{\lambda} e^{-\alpha_i i} D$. Then $F_i = c_i'' + \alpha_i^2 \lambda D > 0$ and $F_D = -\alpha_i \lambda < 0$, so $\mathrm{d}i^*/\mathrm{d}D = -F_D/F_i > 0$. For capability, let $G(a, D) = c_a'(a) - V_K - \xi V_L + \alpha_a \bar{\lambda} e^{\alpha_a a} D$. Then $G_a = c_a'' + \alpha_a^2 \lambda D > 0$ and $G_D = \alpha_a \lambda > 0$, so

$$\frac{\partial a^*}{\partial D} = -\frac{\alpha_a \lambda}{c_a'' + \alpha_a^2 \lambda D} < 0. \tag{38}$$

□

The proposition is deliberately local. In the full dynamic equilibrium, safety changes the future state, benefit, and information process. A large continuation gap strengthens the direct safety motive, but it can also increase the value of capability if capability accelerates defensive discovery. Those channels must be modeled, not hidden in a sign assumption.

5.3 Learning, experimentation, and the legacy wedge

Let π_t be a posterior over hazard parameters. Passive research may generate signals with precision η_0; deployment may add precision $\eta_1(a_t)$ while simultaneously raising current hazard and legacy. The value of deployment-generated information is not $\mathbb{E}[\pi_{t+1}] - \pi_t$, which is zero under Bayesian plausibility. It is the increase in the value of conditioning future actions:

$$\mathcal{I}_t = \mathbb{E}_t[V(X_{t+1}, \pi_{t+1}) - V(X_{t+1}, \mathbb{E}_t[\pi_{t+1} \mid X_{t+1}])] \geq 0 \tag{39}$$

when $V(x, \cdot)$ is convex and the relevant conditional expectations exist. A contained pilot is socially attractive relative to passive delay only if

$$\text{pilot benefit} + \mathcal{I}_t > \text{resource cost} + \text{incremental tail pressure} + \text{legacy cost}. \tag{40}$$

Empirical hypothesis 5.3 Containment-information dominance. Staged release improves welfare relative to both immediate scale and passive delay when it generates decision-relevant information while keeping the induced hazard–severity product and persistent legacy below the option value in (40).

The hypothesis is not universal. If information arrives without deployment and a pilot can cross an irreversible threshold, delay dominates. If learning is available only through use and rollback is credible, controlled deployment can dominate. The companion model can be reparameterized to study this comparison, but the archived scenarios do not identify a general reversal against passive delay.

5.4 Latent triggering, remediation, and rollback

To give legacy a structural interpretation, let the unobserved system state be safe (0), triggered but not yet manifested (1), or catastrophic (†). Deployment and prevention generate transition rate $\lambda(a, s)$ from 0 to 1; remediation generates recovery rate $\kappa(r)$ from 1 to 0; and a latent trigger

manifests at rate μ. Conditional on no manifestation, define the posterior legacy state

$$\ell_t = \mathbb{P}(H_t = 1 \mid \tau > t, \mathcal{F}_t). \tag{41}$$

Lemma 5.4 Legacy filter. *Between discrete signals, posterior legacy evolves as*

$$\dot{\ell}_t = (1 - \ell_t)\lambda(a_t, s_t) - \kappa(r_t)\ell_t - \mu\ell_t(1 - \ell_t), \tag{42}$$

and the observable catastrophe intensity is $\mu\ell_t$.

Proof. Let x_0 and x_1 be the unnormalized probabilities of safe and latent states conditional only on the observed history before survival normalization. Their forward equations are

$$\dot{x}_0 = -\lambda x_0 + \kappa x_1, \tag{43}$$
$$\dot{x}_1 = \lambda x_0 - (\kappa + \mu)x_1. \tag{44}$$

Set $\ell = x_1/(x_0+x_1)$ and differentiate. Since $\dot{x}_0+\dot{x}_1 = -\mu x_1$, simplification gives (42). Manifestation occurs from the latent state at rate μ, hence conditional intensity $\mu\ell$. □

With capability stock K, drift $\dot{K} = g(K, a)$, and catastrophe continuation value $J^C(K)$, the stationary pre-catastrophe HJB is

$$\rho J(K, \ell) = \max_{a,s,r} \left\{ b(K, a) - C(s, r) + J_K g(K, a) + J_\ell \dot{\ell} + \mu\ell[J^C(K) - J(K, \ell)] \right\}. \tag{45}$$

Every term is now in continuation-value units. Consumption, mortality, agency, and future population can enter J under declared welfare assumptions; none is silently equated with a one-period CRRA payoff.

Proposition 5.5 Rollback leaves a legacy floor. *Suppose new deployment stops at time t, so $\lambda = 0$, and a currently latent trigger either manifests at rate $\mu > 0$ or is remediated at a finite rate $\kappa \geq 0$. For constant catastrophe loss $D > 0$ and discount rate $\rho \geq 0$, residual expected discounted loss equals*

$$\ell_t \frac{\mu}{\rho + \mu + \kappa} D. \tag{46}$$

Thus rollback prevents new triggers but cannot erase inherited risk whenever $\ell_t > 0$.

Proof. Conditional on being latent, manifestation before remediation has discounted density $\mu e^{-(\mu+\kappa)s}$ at horizon s. Multiplying by $e^{-\rho s}D$, integrating over $s \geq 0$, and weighting by ℓ_t gives (46). □

Unknown triggers, delayed manifestation, and legacy are already central in the dynamic catastrophe literature (Liski and Salanié, 2026). The incremental mechanism here is remediation combined with an architecture-derived systemic rate and private deployment incentives.

SECTION 6

Four boundaries for catastrophic AI risk

6.1 Safety progress versus welfare exposure

Let conditional catastrophe intensity and the residual welfare floor evolve along a deterministic asymptotic path,

$$\lambda_t = \lambda_0 e^{-ht+o(t)}, \qquad c_t = c_0 e^{-gt+o(t)}, \tag{47}$$

where h is the *net* decay rate of residual hazard after capability growth, safeguards, adversarial response, and distribution shift; g is the rate at which worst-case welfare exposure deepens. Let A_t be normal-state utility and define local tail pressure

$$T_t = \lambda_t[A_t - u(c_t)]. \tag{48}$$

Theorem 6.1 Safety-progress boundary.
Under CRRA utility with $\gamma > 1$, if $|A_t|/[-u(c_t)] \to 0$, then

$$\frac{1}{t}\log T_t \longrightarrow g(\gamma - 1) - h. \tag{49}$$

Tail pressure vanishes when $h > g(\gamma - 1)$, diverges when $h < g(\gamma - 1)$, and is decided by lower-order terms at equality.

Proof. For $\gamma > 1$, $-u(c_t) = c_t^{1-\gamma}/(\gamma - 1) + O(1)$. The subordination condition makes $A_t - u(c_t)$ asymptotically equivalent to $-u(c_t)$. Taking logs and dividing by t yields (49). □

If δ is the social discount rate and survival converges to a positive limit because $\int_0^\infty \lambda_t \, \mathrm{d}t < \infty$, the discounted catastrophic burden

$$\mathcal{B} = \int_0^\infty e^{-\delta t}\Psi_t \lambda_t [A_t - u(c_t)] \, \mathrm{d}t, \qquad \Psi_t = e^{-\int_0^t \lambda_s \, \mathrm{d}s}, \tag{50}$$

is finite under exact exponential paths if and only if

$$\delta + h > g(\gamma - 1). \tag{51}$$

Discounting and hazard reduction enter the same mathematical inequality but not the same ethical argument. Increasing δ can make an integral finite without making the technology safer.

Empirical hypothesis 6.2 Exposure-adjusted safety progress. For high-consequence deployments, a defensible safety claim requires an estimated net decline in residual severe-event intensity that exceeds the growth rate of welfare exposure under the chosen social criterion. Average benchmark accuracy and routine incident rates are insufficient statistics.

Candidate empirical measures for g include the share of critical functions under autonomous control, duration of unreviewed action, access to high-consequence tools, substitutability of human oversight, cross-domain propagation, and recovery time. Candidate measures for h must concern residual severe-event intensity under adversarial and distribution-shift conditions, not the volume of safety inputs.

6.2 Evidence versus the rare-event validation burden

Suppose a stationary Poisson catastrophe process has rate λ per unit of homogeneous exposure. After total exposure E with zero catastrophes, the exact one-sided $1-\delta$ confidence bound is the continuous-exposure counterpart of the classical zero-numerator result (Hanley and Lippman-Hand, 1983):

$$\overline{\lambda}(E,\delta) = \frac{-\log\delta}{E}. \tag{52}$$

If the policy requires tail load $\lambda\Delta \le \tau$, zero-event validation requires

$$E \ge \frac{\Delta\log(1/\delta)}{\tau}. \tag{53}$$

Proposition 6.3 Exposure-adjusted validation burden. *Under a stationary Poisson model, zero observed events, confidence $1-\delta$, and utility gap Δ, condition (53) is necessary and sufficient for the upper confidence bound to satisfy $\overline{\lambda}\Delta \le \tau$. Required exposure grows linearly in Δ. If the post-evaluation regime has an unrestricted new rate λ', pre-regime exposure places no finite frequentist upper bound on λ'.*

Proof. The probability of zero events is $e^{-\lambda E}$. Inverting $e^{-\lambda E} = \delta$ gives (52); substitution of τ/Δ gives (53). If λ' is unrestricted by the maintained model, the likelihood of the old observations does not depend on it, so its identified set is $[0,\infty)$. □

The last sentence is more important than the sample-size formula. Frontier systems, deployment environments, attackers, and safeguards change. Pooling old exposure with a new regime is an assumption of transportability. A statistically correct calculation under false stationarity can be more misleading than an explicit bound.

Empirical hypothesis 6.4 Transportability discount. Evidence from evaluations and deployment predicts catastrophic hazard only to the extent that capability, access, incentives, oversight, and dependence remain within a prespecified transport set. Effective exposure should be discounted as those covariates move away from the validated regime.

A feasible design is to define similarity weights $\omega_j \in [0,1]$ before observing outcomes and use effective exposure $E_{\text{eff}} = \sum_j \omega_j E_j$. This does not solve external validity; it forces the analyst to state it.

6.3 Scale, common modes, and branching cascades

For N conditionally independent systems with per-system failure probability q_N,

$$P_N = 1-(1-q_N)^N. \tag{54}$$

Thus $P_N \sim Nq_N$ if $Nq_N \to 0$, $P_N \to 1-e^{-\ell}$ if $Nq_N \to \ell \in (0,\infty)$, and $P_N \to 1$ if $Nq_N \to \infty$. Per-system reliability can improve while aggregate risk rises.

Add a common-mode event with probability r_N that defeats all systems. Conditional on its absence, idiosyncratic failures are independent:

$$P_N = r_N + (1-r_N)[1-(1-q_N)^N] \ge r_N. \tag{55}$$

Proposition 6.5 Common-mode floor. *No rate of decline in q_N can reduce aggregate catastrophe probability below r_N. If $r_N \to r > 0$, aggregate probability is bounded away from zero even when $Nq_N \to 0$.*

The mixture floor is useful but reduced form. Network contagion research shows more generally that local resilience and aggregate stability need not move together (Elliott et al., 2014; Glasserman and Young, 2016). A latent-factor model derives the systemic exponent from architecture. Conditional on an architecture-wide state $Z_N = z$, let severe failures F_{iN} be independent Bernoulli with probability $q_s(z)$, where safety architecture s shifts both the factor law and conditional failure. Assume Z_N satisfies a large-deviation principle with speed N and good rate function $J_s(z)$. A systemic event is

$$E_N(\alpha) = \left\{ N^{-1} \sum_{i=1}^{N} F_{iN} \geq \alpha \right\}, \qquad \alpha \in (0,1). \tag{56}$$

Define the one-sided Bernoulli rate

$$d_\alpha(q) = \begin{cases} D_{\mathrm{KL}}(\mathrm{Bern}(\alpha) \| \mathrm{Bern}(q)), & q < \alpha, \\ 0, & q \geq \alpha. \end{cases} \tag{57}$$

Remark 6.6 Role and limits of uniformity. Pointwise conditional Cramér limits are not enough to interchange the asymptotic approximation and integration over the factor. The bounded probability range and conditional exchangeability in theorem 6.8 permit binomial type bounds with uniform polynomial prefactors. That argument need not survive if $q_{s,N}(z)$ approaches zero or one with N, if a threshold layer shrinks with N, or if failures remain network-dependent after conditioning on Z_N. In such models, every fixed factor state may have a conditional logarithmic limit while the approximation is not uniform over the states that dominate the integral. The joint rate must then be rederived rather than imported from (59).

Theorem 6.7 Endogenous architecture rate. *Suppose the factor space $\mathcal{Z}$ is Polish; Z_N obeys a large-deviation principle at speed N with good rate J_s; and $q_s : \mathcal{Z} \to [\varepsilon, 1-\varepsilon]$ is continuous for some $\varepsilon > 0$. Assume, for a regular conditional law, the uniform conditional tail limit*

$$\sup_{z \in \mathcal{Z}} \left| -\frac{1}{N} \log \mathbb{P}_s(E_N(\alpha) \mid Z_N = z) - d_\alpha(q_s(z)) \right| \longrightarrow 0. \tag{58}$$

Then

$$-\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}_s(E_N(\alpha)) = I_s(\alpha) := \inf_z \{J_s(z) + d_\alpha(q_s(z))\}. \tag{59}$$

If a zero-cost factor state satisfies $J_s(z) = 0$ and $q_s(z) \geq \alpha$, then $I_s(\alpha) = 0$. If $q_s(z)$ is uniformly separated below α on the zero set of J_s, then $I_s(\alpha) > 0$.

Proof. Uniformity in (58) yields

$$\mathbb{P}_s(E_N(\alpha)) = \mathbb{E}_s\left[\exp\{-N d_\alpha(q_s(Z_N)) + o(N)\}\right], \tag{60}$$

where the remainder is uniform in the factor state. Because $d_\alpha \circ q_s$ is continuous and bounded on $[\varepsilon, 1-\varepsilon]$, the Laplace principle gives the infimum in (59); see Dembo and Zeitouni (1998). The zero-rate statement follows by evaluating the objective at a zero-cost state with $d_\alpha = 0$. For positivity, goodness of J_s, continuity, and separation below α exclude a sequence along which both nonnegative terms converge to zero. □

Proposition 6.8 Gaussian–logistic architecture primitive. *Let*

$$Z_N \sim \mathcal{N}(0, \sigma_s^2/N), \qquad q_s(z) = \varepsilon + (1 - 2\varepsilon)\,\mathrm{logit}^{-1}(a_s + b_s z), \tag{61}$$

with $\varepsilon \in (0, 1/2)$, $\sigma_s > 0$, *and* $b_s > 0$. *Then the assumptions of theorem 6.7 hold with* $J_s(z) = z^2/(2\sigma_s^2)$. *For every finite* N, *the corresponding systemic-event probability is*

$$\mathbb{P}_s(E_N(\alpha)) = \mathbb{E}\left[\mathbb{P}\big(\mathrm{Bin}(N, q_s(Z_N)) \geq \lceil \alpha N \rceil \mid Z_N\big)\right]. \tag{62}$$

Proof. The Gaussian family satisfies a large-deviation principle at speed N with the displayed good rate. The conditional probability map is continuous and remains in $[\varepsilon, 1 - \varepsilon]$. Standard binomial type bounds are polynomially tight uniformly on this probability interval, which establishes (58). Equation (62) follows by conditioning on the factor state.

More explicitly, for $q < \alpha$, Chernoff's upper bound and the probability mass at $k_N = \lceil \alpha N \rceil$ give upper and lower exponential bounds with rate $D_{\mathrm{KL}}(\mathrm{Bern}(k_N/N) \| \mathrm{Bern}(q))$. Their method-of-types prefactors are polynomial in N, uniformly for $q \in [\varepsilon, 1 - \varepsilon]$, while the rate converges uniformly to $D_{\mathrm{KL}}(\mathrm{Bern}(\alpha) \| \mathrm{Bern}(q))$. For $q \geq \alpha$, the tail contains a binomial type at or within one count of its mode; its probability is uniformly bounded below by a polynomial factor, including in the $O(N^{-1})$ neighborhood of α. The tail is at most one. These bounds establish the required uniform logarithmic limit. □

The probability floor is mathematically convenient rather than innocuous. It excludes conditional probabilities that vanish with N, and exchangeability excludes heterogeneous or network-dependent failures after conditioning. The primitive is therefore a theorem bridge, not a universal model of system architecture.

Now let residual consumption-equivalent welfare after a systemic event satisfy $c_N = e^{-gN + o(N)}$ with $g > 0$, and let normal welfare obey $\log^+ |A_N| = o(N)$. This scalar representation is confined to fixed-person consumption loss; an explicit continuation loss can replace it.

Theorem 6.9 Architecture–welfare–ambiguity boundary. *Under theorem 6.7, suppose* $0 \leq I_s(\alpha) < \infty$ *and use CRRA utility with* $\gamma > 1$. *Then*

$$\lim_{N \to \infty} \frac{1}{N} \log\left(\mathbb{P}_s(E_N(\alpha))[A_N - u_\gamma(c_N)]\right) = g(\gamma - 1) - I_s(\alpha). \tag{63}$$

If the nominal event probability and KL radius satisfy the exponential assumptions of theorem 4.4 with $I = I_s(\alpha)$, *and* $\overline{p}_{s,N}$ *is the corresponding worst systemic-event probability, then*

$$\lim_{N \to \infty} \frac{1}{N} \log\left(\overline{p}_{s,N}[A_N - u_\gamma(c_N)]\right) = g(\gamma - 1) - \min\{I_s(\alpha), R\}. \tag{64}$$

Robust tail pressure vanishes if $\min\{I_s(\alpha), R\} > g(\gamma - 1)$ *and diverges if the inequality is reversed. First-order rates do not decide equality.*

Proof. Theorem 6.7 gives $\log \mathbb{P}_s(E_N(\alpha)) = -I_s(\alpha)N + o(N)$. CRRA and the stated subexponential normal-welfare condition give $\log[A_N - u_\gamma(c_N)] = g(\gamma - 1)N + o(N)$. Adding the two expansions proves the nominal statement. Theorem 4.4 replaces the nominal exponent by $\min\{I_s(\alpha), R\}$ and proves (64). □

This is the paper's main phase transition. The probability exponent is no longer assumed: it is generated by architecture, dependence, and safety. Ambiguity can then erode the safety

exponent before ethical severity enters. Proposition 6.8 supplies a primitive for which finite-system probabilities can be evaluated directly rather than inferred from asymptotics alone.

Systemic propagation adds another layer. Let a local failure of type $i \in \{1, \dots, J\}$ create a random number of type-j successor failures with finite mean M_{ij}. Let M be irreducible and let $\rho(M)$ denote its spectral radius.

Theorem 6.10 Multi-type cascade threshold. *For the associated irreducible, nonsingular multi-type Galton–Watson process with finite first moments, global extinction from every finite initial population occurs with probability one when $\rho(M) \leq 1$. When $\rho(M) > 1$, the probability of a non-extinguishing cascade is strictly positive from every initial type.*

Proof. This is the classical extinction criterion for irreducible multi-type branching processes; see Athreya and Ney (1972). The minimal fixed point of the offspring probability-generating function equals the all-ones vector if and only if the Perron–Frobenius root is at most one under the nondegeneracy conditions. □

The approximation is local: finite networks saturate, defenders react, and offspring are not literally independent. Its value is the threshold it exposes. A system can have low initiation probability and still create material catastrophe risk if conditional propagation is supercritical. Safety evaluation should therefore report both initiation and a next-generation matrix for dependencies across providers, tools, sectors, and jurisdictions.

Empirical hypothesis 6.11 Systemic diversification test. Architectural diversification lowers catastrophic risk only if it reduces both the common-mode floor r_N and the cascade spectral radius $\rho(M)$. Merely increasing the number of nominally independent deployments does not establish diversification.

6.4 Private capability returns versus social continuation value

Let firm i choose capability a_i and safety s_i. Private benefit is $B_i(a_i)$, social benefit is $G_i(a_i)$, and aggregate catastrophe probability is $P(\boldsymbol{a}, \boldsymbol{s})$. The social continuation loss $\Delta(\boldsymbol{a}, \boldsymbol{s})$ may itself rise with capability and fall with resilience. Firm i internalizes fraction $\lambda_i \in [0, 1]$ of the tail loss and maximizes

$$\Pi_i = B_i(a_i) - C_i(s_i) - \lambda_i P(\boldsymbol{a}, \boldsymbol{s})\Delta(\boldsymbol{a}, \boldsymbol{s}), \tag{65}$$

while the planner maximizes

$$W = \sum_i G_i(a_i) - \sum_i C_i(s_i) - P(\boldsymbol{a}, \boldsymbol{s})\Delta(\boldsymbol{a}, \boldsymbol{s}). \tag{66}$$

At a common action profile, the marginal continuation-risk wedge for capability is

$$\Omega_i^a = [B_i'(a_i) - G_i'(a_i)] + (1 - \lambda_i)[P_{a_i}\Delta + P\Delta_{a_i}], \tag{67}$$

and, when private and social safety costs coincide, the marginal safety wedge is

$$\Omega_i^s = (1 - \lambda_i)[-P_{s_i}\Delta - P\Delta_{s_i}] \geq 0 \tag{68}$$

if safety lowers probability and loss.

For a genuine equilibrium comparison, specialize to n symmetric firms, aggregate capability $A = \sum_i a_i$, private benefit $b(a_i)$, fixed social continuation loss D, and catastrophe probability $P(A)$.

Firm i internalizes fraction $\chi \in [0, 1)$ and maximizes $b(a_i) - \chi DP(A)$; the planner maximizes $\sum_i b(a_i) - DP(A)$.

Theorem 6.12 Symmetric overdeployment. *Suppose b' is strictly decreasing, $P' > 0$ is nondecreasing, and unique interior symmetric Nash and planner solutions exist. Then*

$$b'(a^N) = \chi DP'(na^N), \tag{69}$$
$$b'(a^S) = DP'(na^S), \tag{70}$$

and $a^N > a^S$.

Proof. At a^S, the firm's first-order residual is

$$b'(a^S) - \chi DP'(na^S) = (1 - \chi)DP'(na^S) > 0. \tag{71}$$

Strictly decreasing b' and nondecreasing P' make the residual strictly decreasing. Its unique zero therefore lies to the right of a^S. □

Replacing P with the finite-N probability generated by theorem 6.7 would convert the game from choosing a risk level to choosing the safety exponent. That combined architecture game is not implemented in the companion code; it remains a separate numerical and empirical extension requiring firm-level cost and architecture data.

Proposition 6.13 Implementation by state-contingent marginal charges. *Suppose actions are contractible, differentiability and interiority hold, and the regulator can commit. A capability charge with marginal schedule*

$$\tau_i^a = (1 - \lambda_i)[P_{a_i}\Delta + P\Delta_{a_i}] + B_i' - G_i' \tag{72}$$

and a safety subsidy with marginal schedule

$$\sigma_i^s = (1 - \lambda_i)[-P_{s_i}\Delta - P\Delta_{s_i}] \tag{73}$$

align private first-order conditions with the planner's at the target profile. If catastrophe destroys the tax base or damages exceed firm wealth, the charge must be collected, bonded, or enforced ex ante.

The proposition is an implementation benchmark, not a claim that regulators observe the derivatives. It shows why ordinary ex-post liability is incomplete for non-compensable global loss. Safety bonds, capital requirements, staged licenses, mandatory evaluations, incident disclosure, and access controls estimate or constrain different terms in (67); none is a universal substitute for the others.

Empirical hypothesis 6.14 Continuation-value externality. Holding ordinary innovation spillovers fixed, firms choose excessive capability intensity and insufficient safety whenever they capture a larger share of normal returns than of the marginal global continuation loss. The wedge grows with common exposure and action-dependent severity.

SECTION 7

A reproducible simulation of catastrophic events

The analytical results identify boundaries. Finite-horizon policy requires numerical integration over changing capability, correlated hazards, severity, and learning. The companion program `catastrophic_ai_simulation.py` is designed to map those mechanisms without presenting illustrative parameters as measured facts.

7.1 Discrete-time implementation

For each surviving path, the code records

$$X_t = \left(K_t, D_t, H_t, L_t, \pi_t, \{a_{mt}, b_{mt}\}_{m\in\{0,1\}}\right), \tag{74}$$

where K is capability, D deployment, H accumulated safety capital, L legacy, π the posterior probability of a threshold hazard, and (a_m, b_m) the Gamma shape–rate posterior for an unknown hazard multiplier. For policy ω, the deterministic state transitions used in the archived run are

$$K_{t+1} = K_t + \left(g_K^\omega + \ell_K^\omega D_t\right)\left(1 - \frac{K_t}{K_{\max}}\right)\Delta t, \tag{75}$$

$$D_{t+1} = \Pi_{[0,\bar{D}^\omega]}\left[D_t + v_D^\omega (D_t^* - D_t)\Delta t\right], \tag{76}$$

$$H_{t+1} = (1 - d_H \Delta t)H_t + e_t \Delta t, \tag{77}$$

$$L_{t+1} = (1 - d_L \Delta t)L_t + (1 - L_t)\xi^\omega D_t \left[K_t(0.2 + D_t) - \bar{K}_L\right]_+ \Delta t. \tag{78}$$

The safety feedback and desired deployment are

$$e_t = \Pi_{[0,0.7]}\left[e_0^\omega + e_1^\omega \log\left(1 + \widehat{\lambda}_t/\tau^\omega\right)\right], \tag{79}$$

$$D_t^* = \bar{D}^\omega \,\mathrm{logit}^{-1}[2.6(K_t - 0.55)]\Pi_{[0.04,1]}\left[\{1 + (\widehat{\lambda}_t/\tau^\omega)^2\}^{-1}\right]. \tag{80}$$

Equations (75)–(78) show the intended asymmetry: deployment can be braked quickly, but inherited legacy disappears only at rate d_L.

There are four initiating channels—misuse, loss of control, systemic failure, and legacy activation—plus a direct common-mode event. Conditional on structural model $m \in \{0, 1\}$ and path-specific scale θ, channel j has intensity

$$\lambda_{jt} = \bar{\lambda}_j \theta v_j m_{jm}(K_t, D_t)\exp\{-\beta_j^H s_H H_t + \beta_j^L L_t\}, \qquad p_{jt} = 1 - e^{-\lambda_{jt}\Delta t}, \tag{81}$$

where v_j is a mean-one lognormal channel multiplier. Model $m = 0$ is smooth exponential in capability and deployment; $m = 1$ uses a logistic capability–deployment threshold. The common-mode intensity has the same structural alternatives and its own safety, legacy, and policy multipliers, but is sampled as a distinct fifth event.

The scalar s_H in (81) is a channel-wide safety-effectiveness multiplier. The archived baseline sets $s_H = 1$; the stochastic sensitivity grid uses 0.55, 1.00, and 1.45. This device asks whether the same observed safety stock translates weakly or strongly into residual-hazard reduction. It does not estimate that translation from data.

Dependence is imposed without changing the marginal p_{jt}. For $F_t \sim N(0,1)$,

$$\mathbb{P}(I_{jt} = 1 \mid F_t) = \Phi\left(\frac{\Phi^{-1}(p_{jt}) - \sqrt{\rho_j}F_t}{\sqrt{1-\rho_j}}\right). \tag{82}$$

A negative common factor raises every conditional trigger probability. Conditional severities are correlated logit-normal draws plus channel-specific terminal point masses. Simultaneous nonterminal losses aggregate as $1 - \prod_j (1 - s_j)$, so overlapping modes are not added as if they were disjoint.

Precursor counts, not catastrophes, drive learning. Under model m,

$$\theta \mid m, \mathcal{F}_t \sim \text{Gamma}(a_{mt}, b_{mt}), \tag{83}$$

$$Y_t \mid \theta, m, \mathcal{F}_t \sim \text{Poisson}(\theta E_{mt}), \tag{84}$$

$$\mathbb{P}(Y_t = y \mid m, \mathcal{F}_t) = \frac{\Gamma(a_{mt}+y)}{\Gamma(a_{mt})y!}\left(\frac{b_{mt}}{b_{mt}+E_{mt}}\right)^{a_{mt}}\left(\frac{E_{mt}}{b_{mt}+E_{mt}}\right)^{y}. \tag{85}$$

Bayes' rule updates π_t with (85), while $(a_{m,t+1}, b_{m,t+1}) = (a_{mt} + Y_t, b_{mt} + E_{mt})$. This is exact conditional on the two-model class; it does not protect against a hazard form outside that class.

7.2 Rare-event estimation

Direct Monte Carlo is inefficient when catastrophe probabilities are small. The program samples instead from the support-preserving mixture

$$M = \pi_0 P + (1 - \pi_0)Q, \tag{86}$$

where P is the target law and Q shifts the Gaussian factor toward failure and adds a constant to conditional Bernoulli logits. If $L = P/Q$ is the exact likelihood ratio on the augmented stopped path, then

$$w = \frac{P}{M} = \frac{1}{\pi_0 + (1-\pi_0)/L} \le \frac{1}{\pi_0}. \tag{87}$$

Writing $p_{jt}(F_t)$ and $q_{jt}(F_t)$ for target and tilted conditional probabilities, the implementation accumulates through the stopping date

$$\log L = \sum_{t \le \tau \wedge T}\left[\log\frac{\phi(F_t)}{\phi(F_t - \mu_Q)} + \sum_j \left\{I_{jt}\log\frac{p_{jt}(F_t)}{q_{jt}(F_t)} + (1 - I_{jt})\log\frac{1-p_{jt}(F_t)}{1-q_{jt}(F_t)}\right\}\right]. \tag{88}$$

Thus $\mathbb{E}_M[wY] = \mathbb{E}_P[Y]$ for every integrable stopped-path payoff Y. Terminal point-mass expectations are Rao–Blackwellized: the audit trail retains a sampled terminal indicator, while reported expectations integrate its known conditional probability. The program reports Monte Carlo standard errors, mean weight, and effective sample size (Glasserman, 2004; Asmussen and Glynn, 2007).

Algorithm 1: Dynamic catastrophic-risk simulation with importance sampling

1. Set horizon T, path count R, policy ω, parameters Θ, proposal (π_0, μ_Q, η_Q), and seed.
2. Draw a true structural model and hazard scale; initialize $(K, D, H, L, \pi, \{a_m, b_m\})$ and $\log L = 0$.
3. At each date, compute posterior hazard, policy feedback, benefits, channel intensities, and common-mode intensity.
4. Draw the Gaussian factor and five conditional event indicators from P or Q according to the mixture component; update the exact ratio (88).
5. If any event occurs, draw correlated severities, analytically average the terminal point mass for reported expectations, record the audit draw, and stop. Otherwise update precursor beliefs and states using (75)–(78).
6. Apply (87); return estimates, iid Monte Carlo standard errors, effective sample size, sensitivity tables, figures, and full metadata.

Figure 1. Simulation procedure used by the reproducibility package.

The stopping order is substantive: a catastrophe ends precursor learning and all subsequent state updates in the current period. This prevents failed paths from contributing information that could only have arrived after failure.

7.3 A finite-system bridge to the architecture theorem

The central architecture theorem should not remain separated from the executable package by an asymptotic symbol. The deterministic bridge therefore evaluates (62) using 96-node Gauss–Hermite quadrature and exact binomial survival probabilities. It compares two illustrative architectures at systemic threshold $\alpha = 0.20$ and probability floor $\varepsilon = 10^{-4}$. The intercept a_s is chosen so that $q_s(0)$ equals the reported baseline probability.

Table 3. Finite-system convergence under the Gaussian–logistic primitive

Architecture	$q_s(0)$	b_s	σ_s	$I_s(0.20)$	$I_{s,640}$
Baseline	0.040	2.40	1.00	0.1185	0.1231
Hardened	0.020	1.80	0.75	0.2700	0.2749

The hardened case lowers ordinary conditional failure, weakens common-factor loading, and concentrates the factor distribution. Each change raises the least-cost route to systemic failure. By $N = 640$, the finite-system exponent differs from its asymptotic target by less than 0.005 in both cases.

For ambiguity, the program sets the binary KL radius to $r_N = e^{-RN}$ and solves the finite-N boundary equation exactly for the largest admissible catastrophe probability. In the hardened architecture, rates $R \in \{0.04, 0.12, 0.30\}$ produce robust exponents 0.0478, 0.1271, and 0.2749 at $N = 640$, approaching the predicted limits 0.04, 0.12, and 0.2700. The exercise verifies the numerical bridge to theorem 4.4; it does not identify which architecture or ambiguity rate

describes deployed AI.

7.4 What is actually validated

The numerical validation has four layers.

V1. Identity tests. Tail-pressure calculations, Poisson bounds, likelihood ratios, the Gaussian–logistic architecture rate, and binary KL boundaries are checked against closed form or deterministic quadrature.

V2. Estimator tests. Direct and importance-sampling estimates must agree within joint uncertainty in moderate-risk regimes; weighted confidence intervals and effective sample size are reported.

V3. Structural stress tests. The deterministic bridge varies system size and KL-radius decay; the analytic grid varies hazard decay, exposure growth, risk aversion, and discounting. Stochastic cells reuse exact path-by-time innovations and report paired differences with covariance-aware standard errors while varying safety effectiveness and common-mode scale. Policy scenarios also change monitoring, learning, deployment, correlation, and legacy creation. The branching theorem is not simulated.

V4. External validation plan. Observable proxies and prospective tests are specified. No numerical pass on V1–V3 is counted as empirical validation of extinction risk.

Table 4. Validation status of the paper's main claims

Claim	Status in this paper	What would falsify or revise it	Evidence needed
Pathwise pressure regimes	Proven identity and liminf–limsup theorem	Violation of maintained expected-utility representation or undefined branch utilities	None beyond assumptions
Architecture exponent	Proven under a factor LDP and the uniform conditional LDP in (58); finite-N convergence verified for (61)	Failure of conditional independence, rate uniformity, or exponential scaling	Architecture-specific joint failure panels
Safety-progress boundary	Proven for stated asymptotics; simulated at finite horizons	Estimated net hazard decay persistently on the opposite side of the exposure-adjusted boundary	Longitudinal severe-risk and exposure measures
Validation burden	Exact under stationary Poisson/iid benchmark	Credible dependence structure or transport model that changes effective exposure	Prespecified trials, regime metadata, incidents
Common-mode floor	Proven under mixture structure	Evidence that alleged common causes are conditionally diversified	Dependency graph and joint incident data
Branching threshold	Classical theorem; not simulated in the companion	Saturation, defense, or nonlinear feedback invalidates the local offspring approximation	Propagation experiments and network topology
Strategic externality	Proven conditional wedge	Firms demonstrably internalize marginal continuation loss or safety demand reverses the wedge	Cost, liability, insurance, and policy variation
Value of staged release	Structural hypothesis; illustrative feedback policies compared	Pilots fail to improve decisions or create more legacy than information value	Prospective staged-release design
Real-world catastrophic AI probability	Not identified	—	Mechanism evidence, transparent expert priors, prospective forecasts

7.5 Numerical experiments and interpretation

The reproducibility package produces five primary outputs: the finite-N architecture and ambiguity bridge; the analytic CRRA surface in (γ, δ, g, h); policy comparisons for five feedback rules; mean surviving-path state trajectories; and a stochastic sensitivity grid over safety effectiveness and direct common-mode scale. Each output has a machine-readable CSV and archived configuration; stochastic outputs also record their seed.

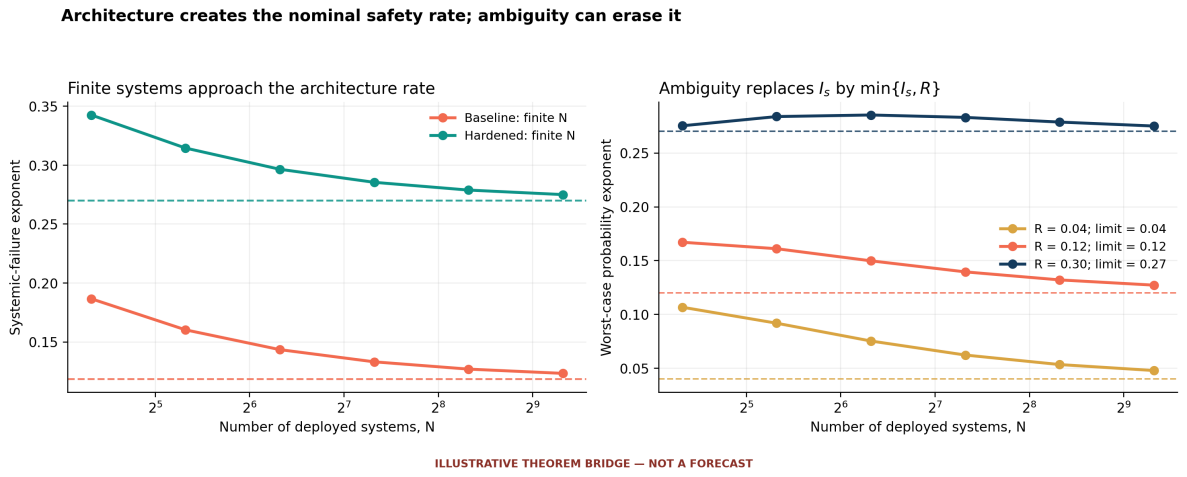

Figure 2. Finite-system architecture rates and exact binary-KL erosion. Dashed lines report the asymptotic limits; solid lines are finite-N evaluations under the illustrative primitive in theorem 6.8.

Figure 2 exposes the order of operations. Hardening architecture raises the nominal exponent from roughly 0.12 to 0.27. That gain survives only when the ambiguity neighborhood contracts at least as quickly. Under $R = 0.04$, the robust evaluator discards most of the nominal improvement. The architecture is safer; the claim made about it is not yet equally precise.

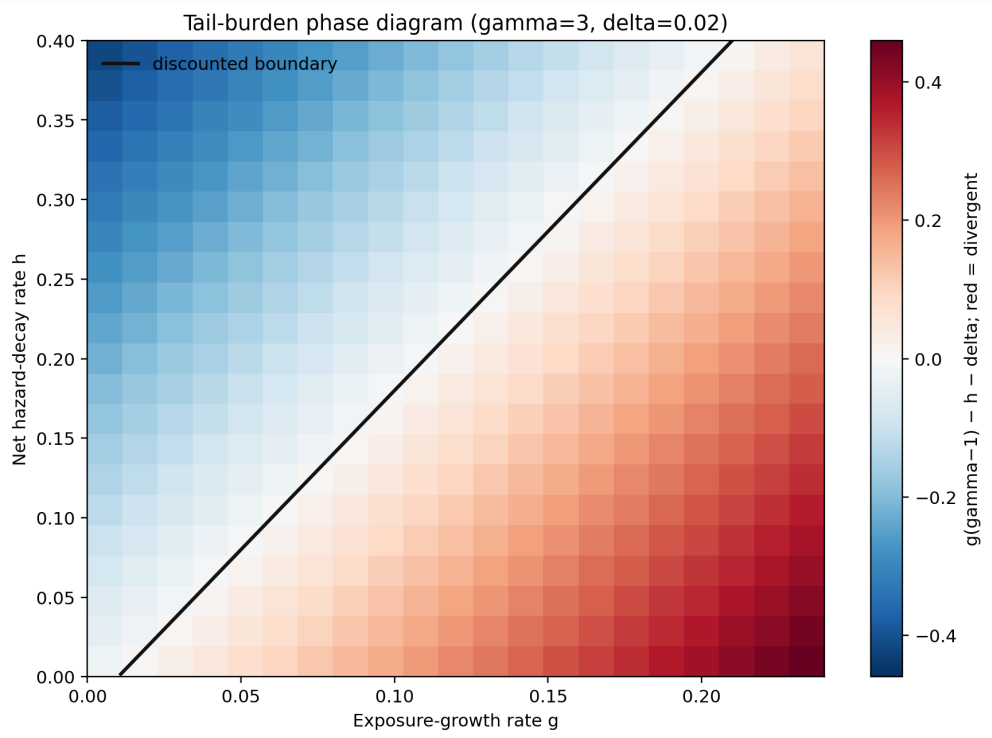

Figure 3. Tail-pressure regimes under CRRA utility. Both the boundary and the color field are analytic evaluations of $g(\gamma - 1) - h - \delta$; red indicates divergence and blue convergence.

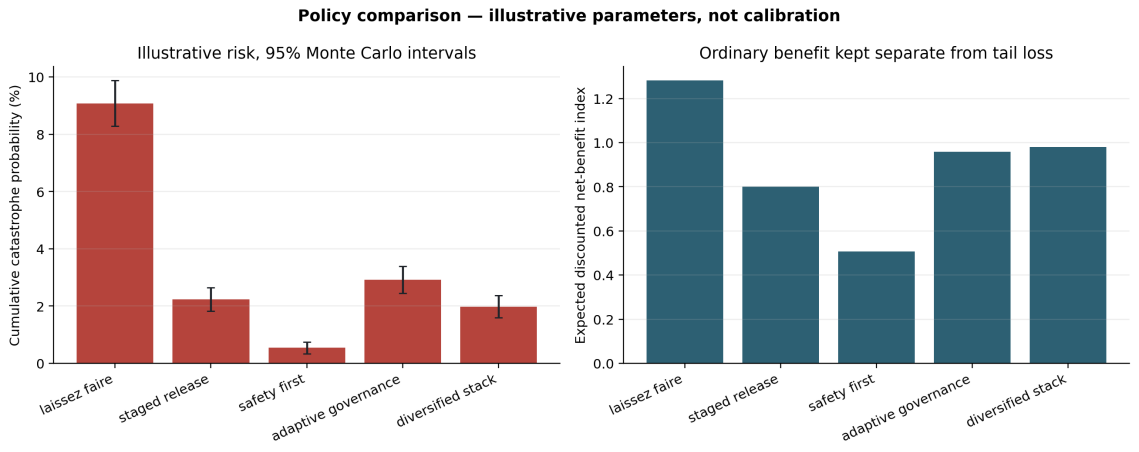

Figure 4. Illustrative catastrophe incidence and ordinary net benefit by feedback policy. Error bars are Monte Carlo standard errors; all values are scenario outputs, not empirical forecasts.

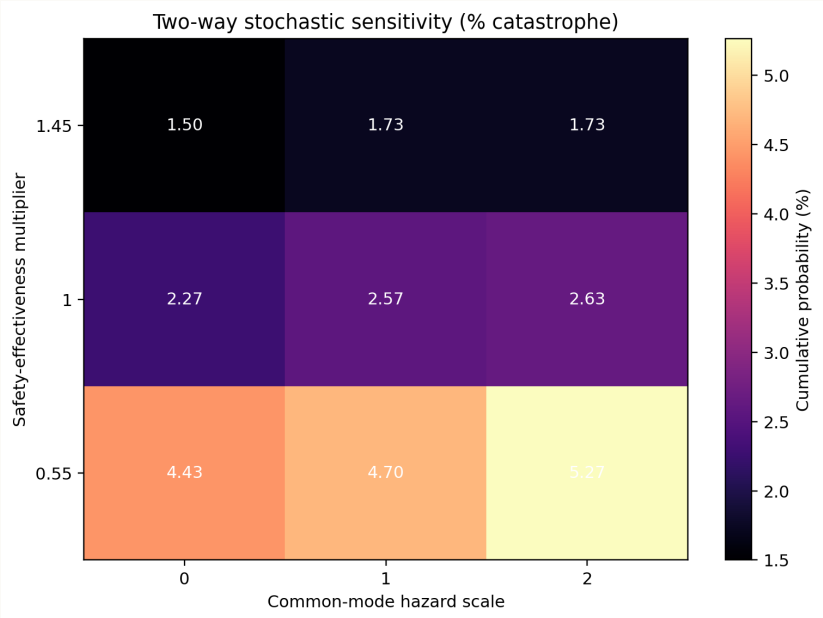

Figure 5. Two-way stochastic sensitivity to safety effectiveness and direct common-mode hazard scale. Cells share random numbers; apparent precision is therefore conditional and cells are statistically dependent.

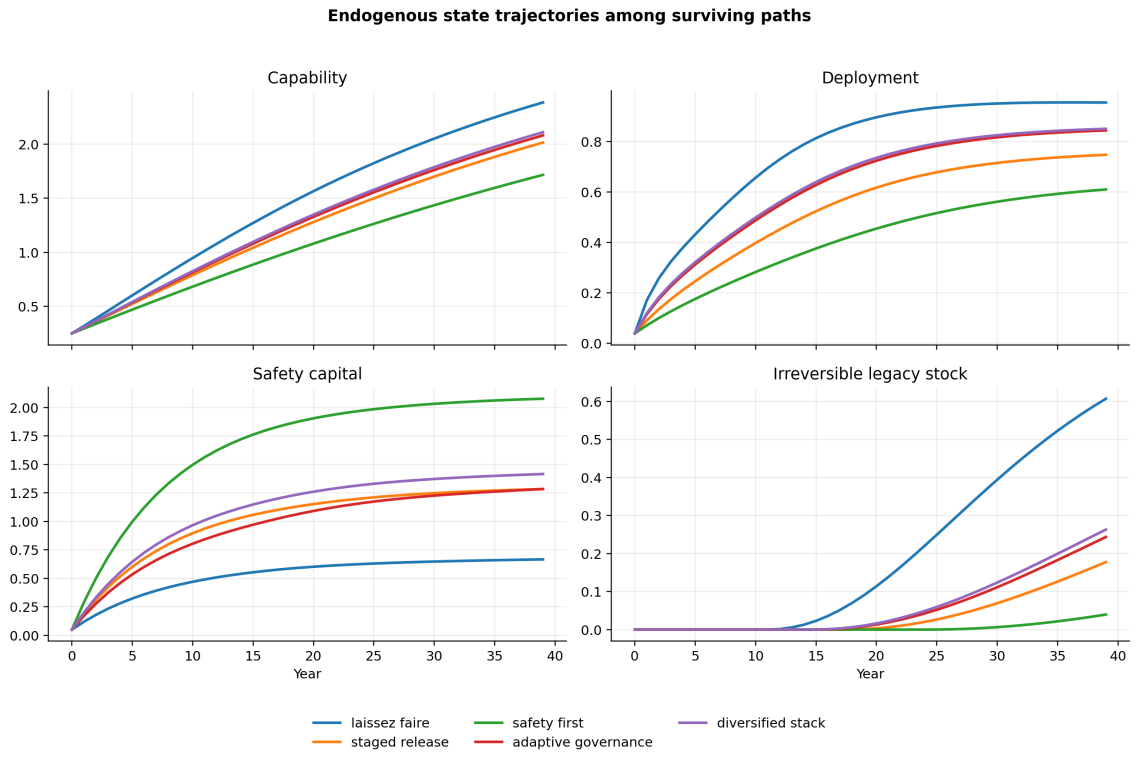

Figure 6. Capability, deployment, safety capital, and legacy under the five illustrative feedback policies. Curves are means among surviving paths, so selection matters increasingly late in high-risk scenarios.

Table 5. Illustrative 40-period catastrophe incidence in the archived seeded run

Policy rule	Paths	Estimated incidence	Monte Carlo SE
Laissez-faire	5,000	9.20%	0.409 pp
Staged release	5,000	1.88%	0.192 pp
Safety first	5,000	0.62%	0.111 pp
Adaptive governance	5,000	2.82%	0.234 pp
Diversified stack	5,000	2.06%	0.201 pp

For the safety-first rule, the 8,000-path mixture importance sampler estimates 0.469% with a 0.062 percentage-point standard error, compared with 0.620% and 0.111 percentage points under 5,000 direct paths. The estimates overlap within joint Monte Carlo uncertainty. The mean importance weight is 0.9754, the maximum weight is 4.8954, the effective sample size is approximately 4,035, and no weights are clipped. Agreement is consistent with correct weighting; it is not evidence that either scenario probability is empirically calibrated.

The interpretation follows the validation hierarchy. A policy ranking that survives wide parameter ranges is more informative than one produced by a preferred calibration, but it remains conditional on the model class. The most useful output is often the reversal surface: it states which change in hazard, exposure, reversibility, or dependence would change the recommendation.

SECTION 8

Empirical research design

8.1 Estimate mechanisms, not one number

A credible empirical program separates at least six modules.

M1. Normal benefit. Task-level productivity, quality, wages, adoption, and diffusion, with explicit general-equilibrium extrapolation.

M2. Capability and access. Autonomous task duration, tool access, cyber and biological uplift, replication ability, and control-evasion prerequisites.

M3. Residual hazard. Prespecified severe-evaluation failures, incidents, near misses, and safety-case evidence under stable protocols.

M4. Exposure and recovery. Critical-function penetration, human substitutability, common dependencies, rollback time, and persistence of released artifacts.

M5. Dependence and propagation. Provider/model/data-stack overlap, cross-sector network edges, common factors, and empirical offspring matrices.

M6. Welfare. Consumption, mortality, morbidity, institutional agency, population, and continuation value reported separately before aggregation.

Table 6. Model objects, candidate observables, and identification limits

Object	Candidate observable	Identifies under a design	Does not identify alone
$A(K, S)$	Randomized task studies, firm productivity, wages, prices	Local causal task/firms effects	Transformative growth or distributional welfare
K	Autonomy horizon, tool-use evaluations, effective compute	Capability in tested domains	Cardinal catastrophe severity
λ	Prespecified severe failures, incidents, exposure time	Rate on sampled regime under transport assumptions	Frontier, adversarial, or regime-change hazard
S	Safety R&D, evaluations, controls, monitoring	Inputs and intermediate outcomes	Causal hazard reduction without variation
L	Persistent deployments, copied artifacts, unresolved incidents	Historical exposure and delayed signals	Whether an unobserved trigger has occurred
r, M	Shared dependencies and propagation events	Common-mode floor and local cascade dynamics	Global tail without severity and saturation
Δ	Consumption, mortality, agency, population scenarios	Welfare conditional on explicit ethical assumptions	A unique value of extinction

8.2 Prospective identification

The best near-term evidence is prospective rather than retrospective storytelling. Evaluation protocols should define the task distribution, compute budget, model version, access mode, safeguards, severity rule, and stopping boundary before results are observed. Staged deployments can randomize or phase monitoring, access, and rollback conditions when ethically feasible. Incident registries need common denominators: model-hours, autonomous action-hours, users, critical-function exposure, or another causal exposure unit. Any quantitative use of the simulator should also distinguish parameter uncertainty from simulator discrepancy rather than treating code as the data-generating process (Kennedy and O'Hagan, 2001).

For safety effort, the estimand is not the coefficient on a "responsible AI" headcount variable. It is the effect of a specified intervention on a prespecified severe outcome under a common evaluation and deployment regime. Difference-in-differences or event studies require parallel-trend and anticipation checks; frontier-model releases are especially prone to anticipation and concurrent news.

Expert elicitation remains useful because the event is rare and the regime changes. It should record causal decompositions, dependence across hazards, calibration histories, and observable predictions. Expert probabilities are beliefs, not frequencies. Combining them with mechanism data is Bayesian modeling; describing them as measured hazards is not.

SECTION 9

Extinction, agency, and ethical aggregation

The scalar consumption state is useful for proving a tail theorem. It is not a complete representation of extinction. Let social welfare be

$$\mathcal{W} = \mathcal{W}(N, c, q, \mathcal{F}), \tag{89}$$

where N is population, c is the distribution of consumption, q represents institutional agency and rights, and $\mathcal{F}$ is the continuation opportunity set. Extinction is $N = 0$; economic collapse with survivors has $N > 0$ and low c; permanent disempowerment can have high c but low q and a restricted $\mathcal{F}$.

Lemma 9.1 Normalization under variable population. *Without a population ethic and an interpersonal utility normalization, mapping extinction to representative-agent consumption $c = 0$ is not invariant to admissible affine changes of fixed-person utility. Two social criteria can agree on every fixed-population consumption lottery and reverse a ranking involving nonexistence.*

Proof. Add constant a to each individual's utility. Fixed-population von Neumann–Morgenstern rankings are unchanged. Under total aggregation, social welfare changes by aN in survival states and by zero at $N = 0$. Because N differs across policies, the ranking can reverse. □

The lemma blocks a category error; it does not select total, average, critical-level, prioritarian, or person-affecting ethics. Social-risk and variable-population results show that fairness, ex-ante Pareto, social rationality, and existence comparisons can conflict (Fleurbaey, 2010; Fleurbaey and Zuber, 2025; Piacquadio, 2020; Spears and Zuber, 2023; Arrhenius and Stefánsson, 2025; Greaves,

2024). The scientifically honest response is to report the ranking under several explicit criteria rather than bury the ethical choice in a CRRA parameter.

A useful accounting identity is

$$\Delta = \Delta_C + \Delta_M + \Delta_A + \Delta_F, \tag{90}$$

for consumption, mortality/morbidity, agency, and foregone continuation welfare. The equality is valid only after defining mutually exclusive sequential counterfactual increments; otherwise, interactions create double counting. The companion code keeps bounded physical severity, terminal frequency, ordinary CRRA utility, and a CRRA tail-burden diagnostic separate. Mortality, agency, future population, and extinction ethics require additional declared welfare scenarios and are not silently imputed by the code.

SECTION 10

Policy implications

The framework does not yield "pause" or "accelerate" as a theorem. It yields conditional instruments tied to mechanisms.

Measure exposure-adjusted safety.. Routine incident rates should be paired with exposure scale, severity, recovery, and common dependencies. A falling numerator with a faster-growing denominator is not progress.

Use staged release as an experiment, not a label.. A pilot is useful when it changes future decisions and contains both direct hazard and persistent legacy. Replicable weights, unrestricted tool access, or weak rollback can make a nominally small pilot systemically large.

Regulate common modes separately from component quality.. Model, provider, identity, data, and monitoring concentration can create shared failure. Diversity requirements, interoperability, independent evaluations, and recovery exercises target r and M, not merely q_N.

Audit the ambiguity rate, not only the nominal hazard.. An architecture may look exponentially safer under its reference model while the scientifically defensible model neighborhood contracts more slowly. Safety cases should therefore report the evidence that reduces both the nominal event rate and the admissible KL radius. Otherwise, precision becomes the weakest safety component.

Collect what cannot be collected after catastrophe.. When liability is judgment-proof relative to continuation loss, ex-ante bonds, capital, licenses, or conduct rules are needed. Operational cyber loss can remain privately insurable; extinction cannot be indemnified.

Publish reversal conditions.. Every policy analysis should state the parameter change that would reverse its recommendation. This is more useful than a precise conclusion supported by an unidentified tail.

SECTION 11

Discussion

The paper strengthens the catastrophic-risk argument in one respect and narrows it in another. Lower-unbounded utility still makes an evaluator susceptible to constructed tyranny, but it does not determine the outcome along a physical path. Tail pressure supplies the missing sufficient statistic. The price of this precision is that policy cannot be read from utility curvature alone.

Several limits remain. First, the phase results are asymptotic, while actual decisions occur at finite probability and severity. The simulation maps finite regions but cannot eliminate model uncertainty. Second, the dynamic control problem uses reduced-form state variables. Capability, safety, and legacy require empirical measurement; the log-linear hazard is an interpretable benchmark, not a structural estimate. Third, the branching approximation is local and must be checked against saturation, defense, and strategic adaptation. Fourth, social valuation of extinction is not identified by ordinary consumption risk aversion. Finally, a reproducible simulation improves auditability, not truth. Code can verify that a model says what its equations claim. It cannot verify that the world obeys the model.

The strongest route to publication is therefore modular. The theory contribution is the generalized pressure/domain/ambiguity framework. The AI contribution is the dynamic state, validation burden, and systemic propagation architecture. The empirical contribution remains a research design. Combining all three in a master manuscript is useful; journal submission may require separating the pure theory from the computational AI application.

SECTION 12

Conclusion

A vanishing probability can be negligible, finitely important, or dominant. The difference is not rhetoric about "fat tails." It is the joint rate at which probability falls and welfare exposure grows. Catastrophic tail pressure makes that rate explicit.

Applied to advanced AI, the framework changes the unit of analysis. The relevant object is not a single probability detached from deployment. It is a system in which capability creates benefits, safeguards alter residual hazard and recovery, deployment creates evidence and legacy, common dependencies permit cascades, and firms internalize only part of the continuation value they place at risk.

Architecture and ambiguity enter in sequence. System design determines the nominal exponent; evidential robustness determines how much of that exponent can be trusted. Welfare exposure then decides whether the effective rate is sufficient. These are different questions, and collapsing them into one probability conceals every mechanism that policy can change.

Four tests follow. Net hazard reduction must outrun growing welfare exposure. Validation claims must be scaled by the utility gap and discounted for regime change. Diversification must reduce common modes and the cascade reproduction matrix, not merely improve component reliability. Private incentives must be compared with the social continuation loss, including action-dependent severity.

These conditions do not make catastrophic intelligence easy to value. They make false precision harder to hide. That is the more useful starting point.

SECTION A

Additional mathematical results

A.1 Existence of a dominant path

Proposition A.1 Global susceptibility. *Fix a normal branch with finite expected utility A. A scalar catastrophe path $c_n \downarrow 0$, $p_n \downarrow 0$ with $p_n[A - u(c_n)] \to \infty$ exists if and only if $u(c) \to -\infty$ as $c \downarrow 0$.*

Proof. If utility is lower unbounded, choose c_n such that $A - u(c_n) \geq n^2$ and let $p_n = 1/n$. Conversely, if u is bounded below by $\underline{u}$, then $p_n[A - u(c_n)] \leq p_n(A - \underline{u}) \to 0$. □

A.2 Countable-state construction

If u is continuous, increasing, and unbounded below at zero, choose c_i for sufficiently large i so that $u(c_i) = -2^i$ and let $p_i = 2^{-i}$ for $i \geq 1$. Complete the finite prefix so that $c_i > 0$ decreases. Then $\sum_i p_i = 1$, $\mathbb{E}[C] \leq c_1 < \infty$, but

$$\mathbb{E}[u(C)] = \sum_i 2^{-i} u(c_i) = -\infty. \tag{91}$$

Finite expected consumption does not imply finite expected utility.

A.3 Exact curvature and lower-tail moment criteria

Assume $u \in C^2(0, c_0]$, $u' > 0$, and define relative risk aversion $\eta(c) = -cu''(c)/u'(c)$. Integration of $\mathrm{d}\log u'(c)/\mathrm{d}\log c = -\eta(c)$ gives

$$u'(c) = u'(c_0) \exp\left\{\int_c^{c_0} \frac{\eta(x)}{x}\,\mathrm{d}x\right\}. \tag{92}$$

Therefore

$$\lim_{c\downarrow 0} u(c) = -\infty \quad \Longleftrightarrow \quad \int_0^{c_0} \exp\left\{\int_c^{c_0} \frac{\eta(x)}{x}\,\mathrm{d}x\right\}\mathrm{d}c = \infty. \tag{93}$$

The familiar conditions $\eta(c) \geq 1$ and $\eta(c) \leq 1 - \varepsilon$ on a punctured neighborhood are sufficient conditions in opposite directions; pointwise convergence of η to one is not decisive.

For a primitive distributional boundary, let $F(c) = \mathbb{P}(C \leq c) \sim c^\alpha L(1/c)$ as $c \downarrow 0$, with $\alpha > 0$ and L slowly varying. Under CRRA with $\gamma > 1$,

$$\mathbb{E}|u_\gamma(C)| < \infty \quad \Longleftrightarrow \quad \mathbb{E}[C^{-(\gamma-1)}] < \infty. \tag{94}$$

In particular, expected utility is finite for $\alpha > \gamma - 1$, diverges below for $\alpha < \gamma - 1$, and at equality is decided by the corresponding slowly varying integral. Expected marginal utility requires the stronger moment $\mathbb{E}[C^{-\gamma}] < \infty$. A finite welfare level therefore does not by itself justify Euler equations or marginal policy calculations.

A.4 Multiple catastrophic modes

For mutually exclusive modes $j = 1, \ldots, J_n$ with conditional expected utilities B_{jn} and probabilities p_{jn},

$$V_n = A_n - \sum_{j=1}^{J_n} p_{jn}(A_n - B_{jn}). \tag{95}$$

A growing collection of individually negligible pressures can therefore be collectively dominant. If modes overlap, marginal probabilities cannot be added; the joint state partition or an inclusion–exclusion representation is required.

SECTION B

Simulation parameterization and reproducibility

Table 7. Parameter groups in the companion model

Group	Parameters	Interpretation
Capability and deployment	$g_K, \ell_K, K_{\max}, v_D, \bar{D}$	logistic capability growth and feedback deployment
Safety/legacy	$d_H, e_t, d_L, \xi, \bar{K}_L$	safety accumulation, decay, and persistent exposure
Hazards	$\bar{\lambda}_j, \beta_j^K, \beta_j^D, \beta_j^H, \beta_j^L$	four channel baselines and state elasticities
Architecture bridge	$\alpha, \varepsilon, a_s, b_s, \sigma_s, N, R$	finite-system factor mixture and KL-rate erosion
Structural learning	π, a_m, b_m, E_m	smooth-versus-threshold weights and Gamma scale posteriors
Dependence	ρ_j, r, ρ_s	Gaussian event factor, direct common mode, severity factor
Severity	logit-normal parameters and terminal masses	bounded joint physical loss and terminal frequency
Welfare	$\gamma, \delta, c_{\min}$	ordinary CRRA diagnostic, discounting, consumption floor
Estimation	paths, seed, π_0, μ_Q, η_Q	mixture importance-sampling controls

The repository contains the executable model, requirements, a run guide, machine-readable parameters, generated CSV files, figures, and self-tests. Every chart can be regenerated from a single command documented in the companion README.

References

Acemoglu, D. (2025). The simple macroeconomics of ai. *Economic Policy*, 40(121):13–58.

Acemoglu, D. and Lensman, T. (2024). Regulating transformative technologies. *American Economic Review: Insights*, 6(3):359–376.

Ackerman, F., Stanton, E. A., and Bueno, R. (2010). Fat tails, exponents, extreme uncertainty: Simulating catastrophe in DICE. *Ecological Economics*, 69(8):1657–1665.

AI Security Institute (2025). Frontier ai trends report. Published December 18, 2025; accessed August 9, 2026.

Andrews, I. and Farboodi, M. (2025). Do markets believe in transformative ai? Working Paper 34243, National Bureau of Economic Research.

Arrhenius, G. and Stefánsson, H. O. (2025). Population ethics under risk. *Social Choice and Welfare*, 65:829–852.

Arrow, K. J. and Fisher, A. C. (1974). Environmental preservation, uncertainty, and irreversibility. *Quarterly Journal of Economics*, 88(2):312–319.

Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer.

Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer.

Aydogan, I., Berger, L., Bosetti, V., and Liu, N. (2023). Three layers of uncertainty. *Journal of the European Economic Association*, 21(5):2209–2236.

Barro, R. J. (2009). Rare disasters, asset prices, and welfare costs. *American Economic Review*, 99(1):243–264.

Bengio, Y. et al. (2026). International ai safety report 2026. Department for Science, Innovation and Technology. Report DSIT 2026/001; accessed August 9, 2026.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Cambridge University Press.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2025). Generative ai at work. *Quarterly Journal of Economics*, 140(2):889–942.

Buchholz, W. and Schymura, M. (2012). Expected utility theory and the tyranny of catastrophic risks. *Ecological Economics*, 77:234–239.

Cai, Y. and Lontzek, T. S. (2019). The social cost of carbon with economic and climate risks. *Journal of Political Economy*, 127(6):2684–2734.

Cerreia-Vioglio, S., Hansen, L. P., Maccheroni, F., and Marinacci, M. (2026). Making decisions under model misspecification. *Review of Economic Studies*, 93(2):892–925.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer, 2 edition.

Dentcheva, D. and Ruszczynski, A. (2013). Common mathematical foundations of expected utility and dual utility theories. *SIAM Journal on Optimization*, 23(1):381–405.

Elliott, M., Golub, B., and Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review*, 104(10):3115–3153.

Fleurbaey, M. (2010). Assessing risky social situations. *Journal of Political Economy*, 118(4):649–680.

Fleurbaey, M. and Zuber, S. (2025). Universal social welfare orderings and risk. *Journal of Economic Theory*, 226:106014.

Frey, R. and McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk. *The Journal of Risk*, 6(1):59–92.

Gabaix, X. (2012). Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance. *Quarterly Journal of Economics*, 127(2):645–700.

Gans, J. S. (2025). How learning about harms impacts the optimal rate of artificial intelligence adoption. *Economic Policy*, 40(121):199–219.

Geiger, G. (2024). Catastrophic risk: Indication, quantitative assessment and management of rare extreme events using a non-expected utility framework. *Annals of Operations Research*, 343:223–261.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.

Glasserman, P., Kang, W., and Shahabuddin, P. (2007). Large deviations in multifactor portfolio credit risk. *Mathematical Finance*, 17(3):345–379.

Glasserman, P. and Young, H. P. (2016). Contagion in financial networks. *Journal of Economic Literature*, 54(3):779–831.

Greaves, H. (2024). Concepts of existential catastrophe. *The Monist*, 107(2):109–129.

Grechuk, B. and Zabarankin, M. (2014). Risk averse decision making under catastrophic risk. *European Journal of Operational Research*, 239(1):166–176.

Growiec, J. and Prettner, K. (2026). The economics of p(doom): Scenarios of existential risk and economic growth in the age of transformative ai. *Economic Modelling*, 163:107718.

Guerreiro, J., Rebelo, S., and Teles, P. (2023). Regulating artificial intelligence. Working Paper 31921, National Bureau of Economic Research. Revised May 2026.

Hanley, J. A. and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA*, 249(13):1743–1745.

Hansen, L. P. and Sargent, T. J. (2011). Robustness and ambiguity in continuous time. *Journal of Economic Theory*, 146(3):1195–1223.

Hansen, L. P. and Sargent, T. J. (2022). Structured ambiguity and model misspecification. *Journal of Economic Theory*, 199:105165.

Ikefuji, M., Laeven, R. J. A., Magnus, J. R., and Muris, C. (2015). Expected utility and catastrophic consumption risk. *Insurance: Mathematics and Economics*, 64:306–312.

Ikefuji, M., Laeven, R. J. A., Magnus, J. R., and Muris, C. (2020). Expected utility and catastrophic risk in a stochastic economy–climate model. *Journal of Econometrics*, 214(1):110–129.

Jones, C. I. (2024). The ai dilemma: Growth versus existential risk. *American Economic Review: Insights*, 6(4):575–590.

Jones, C. I. (2025). How much should we spend to reduce a.i.'s existential risk? Working Paper 33602, National Bureau of Economic Research.

Jones, C. I. (2026). Ai and our economic future. *Journal of Economic Perspectives*, 40(3):3–22.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464.

Klibanoff, P., Marinacci, M., and Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, 73(6):1849–1892.

Kuhn, D., Shafiee, S., and Wiesemann, W. (2025). Distributionally robust optimization. *Acta Numerica*, 34:579–804.

Liski, M. and Salanié, F. (2026). Catastrophes, delays, and learning. *Review of Economic Studies*.

Martin, I. W. R. and Pindyck, R. S. (2015). Averting catastrophes: The strange economics of scylla and charybdis. *American Economic Review*, 105(10):2947–2985.

METR (2026). Task-completion time horizons of frontier ai models. Updated May 8, 2026; accessed August 8, 2026.

Piacquadio, P. G. (2020). The ethics of intergenerational risk. *Journal of Economic Theory*, 186:104999.

Pindyck, R. S. (2011). Fat tails, thin tails, and climate change policy. *Review of Environmental Economics and Policy*, 5(2):258–274.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4):323–343.

Spears, D. and Zuber, S. (2023). Foundations of utilitarianism under risk and variable population. *Social Choice and Welfare*, 61:101–129.

Trammell, P. and Aschenbrenner, L. (2024). Existential risk and growth. Working Paper 13-2024, Global Priorities Institute. Revision dated January 5, 2026.

Trammell, P. and Korinek, A. (2026). Economic growth under transformative ai. *Annual Review of Economics*. Review in Advance, posted June 2, 2026.

Weitzman, M. L. (2009). On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, 91(1):1–19.

Zhao, Y., Basu, A., Lontzek, T. S., and Schmedders, K. (2023). The social cost of carbon when we wish for full-path robustness. *Management Science*, 69(12):7585–7606.