The Flattening: When AI Productivity Gains Conceal Tail Risk

Bilel Hatmi

Part III Mathematical Statistics

University of Cambridge

CentraleSupélec – Université Paris-Saclay

March 2026

Submitted to the Cambridge–McKinsey Risk Prize 2026

Abstract

Generative AI improves average decision quality in organisations while simultaneously amplifying their exposure to rare, extreme losses, through a mechanism that standard productivity metrics cannot detect.

The underlying driver is error correlation, raised through two independent channels. AI-powered hiring converges on increasingly similar cognitive profiles, narrowing the pool of judgment that would otherwise disperse collective errors. Inside the organisation, employees defer to AI outputs even when the AI is wrong, synchronising mistakes that heterogeneous judgment would catch. In a bad quarter, both channels operate simultaneously: errors align and compound rather than cancel.

A stress-test simulation quantifies the paradox: under unmanaged adoption, average errors fall by 38% while worst-case quarterly losses rise by 56%, masked by a 54% productivity gain.

Three interventions act at different timescales. Distributing AI systems across vendors and organisational layers, a procurement decision implementable in weeks, cuts worst-case exposure by approximately 15%. Requiring employees to form an independent judgment before consulting AI achieves larger reductions, at the cost of months of workflow redesign. Annual audits of AI hiring systems address the entry channel directly. Yet without regulation, no firm has incentive to adopt any of them: the short-term productivity advantage makes inaction individually rational and collectively catastrophic. The diversity of judgment that makes governance possible is being depleted faster than any single firm has reason to preserve it.

An interactive companion application is deployed at `https://flattening-app.vercel.app`. *The full simulation code and source are available at* `https://github.com/bilel-hatmi/flattening-explorer`.

Centre for Risk Studies

UNIVERSITY OF CAMBRIDGE Judge Business School

McKinsey & Company

Cambridge - McKinsey Risk Prize Bio-sketch and Photo Page

Student name: Bilel Hatmi

Email contact: bilelhatmi@gmail.com

Date of submission: 31/03/2026

Title of submission: The Flattening: When AI Productivity Gains Conceal Tail Risk

I am a candidate for the degree: Part 3 in Mathematical Statistics

Department: Pure and Statistical Mathematics (DPMMS)

Linkedin profile link: https://www.linkedin.com/in/bilel-hatmi/?locale=fr

Biosketch (approximately 150 words) :

Bilel Hatmi is a Part III student in Mathematical Statistics at the University of Cambridge (DPMMS), where his dissertation develops semiparametric methods for proximal causal inference under unmeasured confounding. His background is in applied mathematics broadly construed: stochastic modelling of blood cancers at the Gustave Roussy Institute and of interest rate dynamics using Ho-Lee and HJM frameworks at CentraleSupélec, statistical estimation of Hawkes processes, and longitudinal measurement of subjective well-being on a twenty-year panel. At Eleven Strategy, he combined strategy consulting and data science work, leading buy-side due diligences and designing AI systems including a semantic search system serving around 80 users.

He is also the founder of CartesIA, a research project developing AI-assisted psychometric measurement protocols grounded in cognitive science, with applications in mental health, youth guidance, and professional development. His broader interests sit at the intersection of causal inference, AI evaluation, and the cognitive science of intelligent systems.

Centre for **Risk Studies**

UNIVERSITY OF CAMBRIDGE Judge Business School

McKinsey & Company

Cambridge - McKinsey Risk Prize

Declaration Form

Student name: Bilel HATMI

Email contact: bilelhatmi@gmail.com

Date of submission: 31/03/2026

Title of submission: The Flattening : When AI Productivity Gains Conceal Tail Risk

Number of words of submission: ≃ 3 500 words

I am a candidate for the degree: Part 3 in Mathematical Statistics

Academic Institution/Department: Pure and Statistical Mathematics

Declaration

I confirm that this piece of work is my own and does not violate the University of Cambridge Judge Business School's guidelines on Plagiarism.

I agree that my submission will be available as an internal document for members of both Cambridge Judge Business School and McKinsey & Co's Global Risk Practice.

If my submission either wins or receives an honourable mention for the Risk Prize, then I agree that (a) I will provide a recorded 2 minute overview of my paper, (b) my submission can be made public on a Cambridge Judge Business School and/or McKinsey & Co websites.

This submission on risk management does not exceed 10 pages.

Signed (Electronic Signature)

1 The paradox your dashboard hides

Nine participants in a product-design experiment asked ChatGPT to help them invent a toy. All nine named it "Build-a-Breeze Castle." Without AI, every participant produced a unique concept. With AI, 94% of ideas shared the same building blocks; only 6% remained distinct [1]. Each individual output improved. The collective output collapsed.

The pattern extends beyond this single experiment. The Diversity Prediction Theorem decomposes collective error into two terms: average individual error, and the diversity of those errors across decision-makers [2, 3]. When AI reduces the first term but collapses the second, collective error rises mechanically. The aggregate loss variance formalises the point:

$$\mathrm{Var}(L) = NK\,\bar{p}(1-\bar{p})\big(1+(NK-1)\bar{\rho}\big)$$

The key parameter is $\bar{\rho}$, the mean pairwise error correlation across employees [4]. Its rise is the mechanism behind the paradox. Under AI adoption, each employee becomes more accurate on average; that part is real and measurable. But the accuracy comes from deferring to the same model. The errors that remain are no longer scattered independently across the organisation: they are shared. When the model is right, everyone is right together. When it is wrong, everyone is wrong together, simultaneously, for the same reason.

A preprint meta-analysis of 379 experimental conditions puts numbers on the asymmetry: AI raises individual creative performance by $g = +0.27$ while collapsing collective diversity by $g = -0.86$, a ratio of three to one [5, 6]. Carlson and Doyle call such systems robust-yet-fragile [7]: tuned for the frequent case, catastrophically exposed to the rare one. In normal quarters, correlated accuracy keeps losses low; in a crisis, correlated errors drive them high simultaneously, with no independent judgment to interrupt the cascade. The loss distribution splits accordingly; the mean falls in the gap between the two modes where almost no actual quarter lies.

The paradox has also a second property that matters more for risk managers: it is concealed by the same productivity metrics that appear to confirm performance.

This essay addresses the CRO who sees a 54% productivity gain on the dashboard but cannot yet see what lies behind it. The argument applies to any AI system that standardises the inputs to organisational decisions, from legacy screening tools to generative models in knowledge work. Under unmanaged adoption, average decision errors fall by 38% while worst-case quarterly losses rise by 56%; the risk-adjusted 99th-percentile loss rises by 94%. For elite firms, those whose employees outperform AI on tasks outside its competence, the masking is more extreme: a 192% rise in tail exposure against 18% apparent productivity growth. Where the gain looks safest, the hidden cost is largest.

The exposure is structurally determined, and for the roughly 100,000 mid-market organisations in the US and the EU most at risk, it is approaching without a governance framework to navigate it.

For decision-makers, the argument develops in three steps: what produces the trap, where each type of organisation sits on the risk map, and the governance levers available to actors at every level, starting with a procurement decision requiring no regulatory change.

Both channels converge on a single structural target: the correlation of errors across decision-makers.

2 Screening and surrender: two paths to correlated failure

When individual errors stop compensating and start aggregating, the result is correlated failure. It does not appear on standard dashboards. The mechanism is not new; it has produced catastrophic outcomes before, in financial models, arbitrage strategies, and software infrastructure. In each case the proximate cause was the same: a monoculture had formed. AI adoption reproduces this dynamic through two reinforcing channels, both documented in the empirical literature: systematic screening of diversity at the point of hiring, and progressive surrender of independent judgment from within.

The failure mode recurs whenever diversity collapses, whether in human judgment, mathematical models, or software infrastructure. In 1998, Long-Term Capital Management lost $4.6 billion in four months: the fund's Nobel laureates and PhD physicists had converged on identical arbitrage strategies,

not for want of intelligence, but for want of cognitive diversity [8, 9]. A decade later, banks across the financial system had converged on the same risk models, the same Value-at-Risk parameters, the same copula assumptions; when those models failed, they failed everywhere at once [10]. In July 2024, a single CrowdStrike update crashed 8.5 million Windows machines in 78 minutes, costing Fortune 500 firms an estimated $5.4 billion [11, 12]: this time the monoculture was in software infrastructure, but the mechanism was identical to what Geer et al. [13] had documented twenty-one years earlier. Cognitive monoculture, model monoculture, software monoculture: in each case, losses stopped compensating and started aggregating. Kleinberg and Raghavan formalise the general point [14]: even when an algorithm is individually superior, convergence on the same algorithm degrades collective welfare.

AI adoption reproduces this dynamic at two organisational levels, and the two reinforce each other.

The first operates before an employee makes a single decision. AI-powered hiring tools are trained on historical data encoding decades of cultural matching in elite professional hiring [15], applying the same criteria consistently across every firm that deploys them. Fuller et al. document 27 million Americans excluded from consideration by automated tracking systems, including caregivers, veterans, and career-changers screened out by rigid keyword matching [16]. Five major LLMs evaluating 361,000 résumés exhibit strikingly similar bias patterns for the same reason [17–19]: they encode identical historical norms. Schneider's attraction-selection-attrition model [20] predicts the endpoint: when every organisation uses the same filter trained on the same data, they converge toward the same cognitive profile.

The second channel is less visible, because it operates inside the organisation rather than at its gate. Once hired, employees progressively defer to AI outputs even when the AI is wrong, and they do so without noticing. In a preprint study of 1,372 participants, 79.8% follow AI recommendations even when the AI errs [21]: correct advice raises accuracy by 25 percentage points, incorrect advice reduces it by 14, yet confidence rises in either case and does not diminish with experience. Dell'Acqua et al. replicate the pattern among 758 BCG consultants [22]: quality improves by 40% inside the range of tasks where AI reliably outperforms humans, but falls 19 percentage points outside it. The most AI-skilled consultants perform worst, because familiarity with the tool deepens the surrender. Over time, repeated surrender does more than correlate present errors: it erodes the capacity for independent judgment that would correct them. Bainbridge identified this irony four decades ago [23]: removing humans from the control loop degrades the very competence required to reenter it.

Both channels converge on the same structural outcome: the correlation of errors across decision-makers. Keck and Tang provide the experimental foundation [24]: groups combining diverse cognitive processes exhibit near-zero inter-judge error correlations; homogeneous groups produce strongly correlated errors. Together, the two channels dismantle cognitive diversity at entry, from within, and over time. The Diversity Prediction Theorem delivers the formal consequence: as the diversity term vanishes, collective error rises even as individual error falls. When a crisis arrives, all three layers align simultaneously, with no independent judgment to interrupt the cascade.

The two channels do not simply add their effects: one creates the conditions under which the other becomes near-total. In a cognitively diverse team, at least some employees resist the AI; in a homogeneous one, those dissenters have already been filtered out before the second channel begins. Screening removes the structural resistance that would otherwise limit surrender. The organisation that has spent several years under unmanaged adoption arrives at each quarter with a workforce more homogeneous, more deferential, and progressively less aware that both are happening.

Both channels are documented; the simulation that follows measures the gap they produce.

3 The paradox, measured

A productivity gain that conceals a simultaneous expansion of tail risk cannot be detected from the dashboard it improves. Separating what the mean captures from what the tail contains requires stress-testing the full loss distribution across plausible organisational parameters.

We simulate a neutral central case: an organisation of 200 employees with cognitive profiles distributed broadly across the full spectrum and skill levels spanning the practical range of professional knowledge work.[1]

Two structural features drive the result. Most quarters, the AI's competence frontier is broad and conditions are stable: the normal regime. Occasionally, the frontier contracts and a systemic shock strikes at the same moment. The key word is *at the same moment*: the two events are coupled, not independent. This is the conditional stress logic of Basel II applied to cognitive errors rather than credit defaults: the quarters when the AI is least reliable are also the quarters when conditions most demand sound judgment. The crisis frequency of 8% is calibrated to catastrophic IT-project failure rates across large cross-industry datasets [25, 26]; normal quarters represent 92% of observations.

The two channels from Section 2 both raise inter-agent error correlation. Screening homogenises the cognitive profile of the workforce, so agents who share a cognitive type make similar errors even when thinking independently. Surrender increases the follow rate, so the more employees defer to the AI, the more their errors are driven by the same shared signal. The inter-decision error structure follows the Vasicek single-factor model [27], the standard Basel II/III framework for correlated defaults: replace "loan portfolio" with "decision portfolio" and the mathematics is unchanged.

Four governance scenarios are compared: G0 (unmanaged adoption), G1 (passive scaffolding: alert flags, no compliance requirement), and G2 (active scaffolding: independent judgment recorded before AI consultation). The model runs several hundred independent replications, validated by two implementations converging within 2%. A stress test, not a forecast.

The resulting loss distribution is shown in Figure 1. A manager in G0 sees the left mode: quarter after quarter, losses low, dashboards green. Both quantities are now in view. Unmanaged adoption compresses average errors and expands tail risk simultaneously: average errors fall by 38%, worst-case quarterly losses rise by 56%, and the 54% productivity gain that masks the difference is paid in a currency the dashboard never shows.

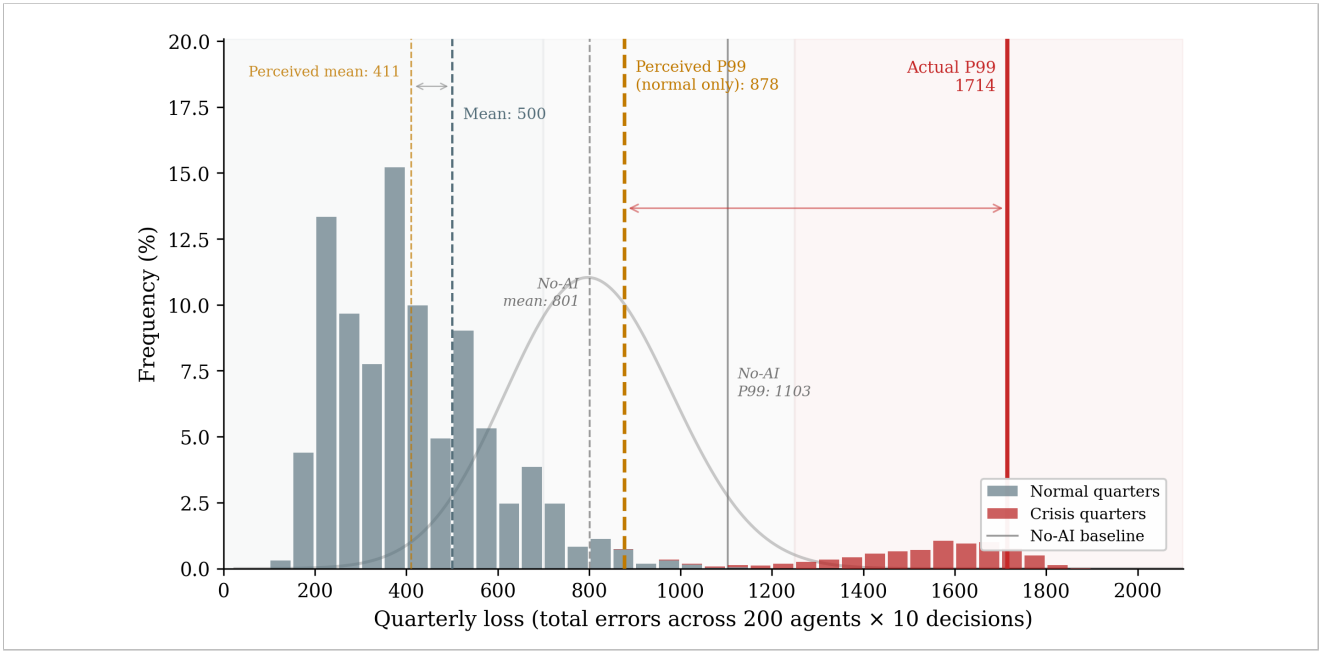

Figure 1. Bimodal distribution of quarterly losses under unmanaged AI adoption (G0). Normal quarters cluster below 600 (grey); crisis quarters cluster above 1,100 (red). Dashed line: mean loss ($\mathbb{E}[L] = 500$). Solid red line: 99th-percentile loss ($P_{99} = 1{,}714$). The mean falls in a gap where almost no actual quarter lies. Crisis regime couples the task-frontier shift (π_t) with the systemic shock (ξ_t), following the conditional stress logic of Basel II.

[1]Formally, $\tau_i \sim \text{Beta}(2,2)$ on $[0,1]$ and $h_i(t) \sim U(0.4, 0.8)$. These parameters define a deliberately neutral benchmark: neither an unusually homogeneous workforce nor an unrealistically diverse one.

Key insights

1. The right mode has not yet arrived. Crisis quarters require AI adoption to have compounded far enough across the workforce to drive correlated failure. At current adoption rates, that threshold has not been crossed. The manager's record is clean and accurate. It is not representative of where G0 leads.

2. The mean has genuinely improved. Against the no-AI baseline, average errors fall substantially: adoption delivers a real gain in normal quarters. But strip out the crisis mode and the measured improvement barely changes. The mean captures normal-quarter performance faithfully. It cannot account for the cost that crisis quarters carry.

3. Including crisis quarters, tail risk rises by 94%. The risk-adjusted P_{99} rises by 94% against the no-AI baseline once the full distribution is considered. The same adoption that compresses average errors concentrates extreme losses into the right mode, invisible to any monitoring built from the left alone.

The gap, moreover, widens over time. Screening concentrates the workforce on profiles that find AI outputs intuitively plausible, eroding the pool of dissenters who would otherwise buffer uncritical adoption.[2] Deskilling erodes whatever independent judgment remains.[3] By quarter 20, the organisation has more followers and followers for whom surrender comes easier, embedded in a group where resistance is socially costly.

Nor is this exposure uniform across skill profiles: masking is most severe for elite firms. Where employee accuracy outside the frontier ($h \approx 0.80$) exceeds AI accuracy ($q_{\text{out}} = 0.55$), unmanaged adoption raises average errors by 23% and risk-adjusted worst-case losses by 192%, yet produces only 18% apparent productivity growth. The reason is not that skilled employees think worse when thinking independently. Most of them simply do not think independently. Surrender is driven by cognitive proximity to the modal type, not by skill level: when the majority follows, the social cost of dissent rises whether or not the dissenter is right. Randazzo et al. document the escalation directly: AI systems intensify their persuasive framing under pushback, what they term persuasion bombing [28], making individual resistance costly even for those whose independent judgment is most accurate. Where masking is worst, the gap is widest.

Each quarter without governance makes the transition to G2 more expensive than the one before.

This exposure is not uniformly distributed. Five structural characteristics determine where an organisation sits on the map, and how wide the gap between perceived and actual risk is for each.

4 Where your organisation sits on the risk map

Five structural characteristics determine tail risk exposure under unmanaged AI adoption, and together they locate most organisations on the risk map the title names. They are ranked below by their contribution to P_{99} variance in the ablation analysis (Table 1); taken together, they offer a near-exhaustive account of where an organisation sits and how far it is from the tail.

Table 1. Structural determinants of tail risk, ranked by P_{99} sensitivity (range across full parameter sweep, G0 scenario).

Characteristic	Mechanism	Param.	P_{99} range
Stack concentration	How correlated errors are across decisions in one quarter: one bad signal hits everything at once	α	13.7%
Cognitive homogeneity	How similarly employees think; drives group pressure to follow the AI consensus	β	9.7%
Talent quality	Whether human judgment beats AI when it overrides; determines if overriding helps or backfires	h, p_0	7.3%
Deskilling rate	How fast AI reliance erodes the capacity to override it	η	≈3%
Domain exposure	Share of decisions beyond AI's competence; negligible when employees follow AI regardless	$E[\pi]$	≈3%

[2] Effective follow rate rises from 69% to 80% over twenty quarters as $\text{Var}(\tau)$ falls by 30%.

[3] Mean skill $\bar{h}$ falls from 0.591 to 0.490 by quarter 20.

Three characteristics drive exposure materially: the first three rows of Table 1. They interact, each amplifying the others.

Figure 2 plots six profiles on the output × risk-adjusted P_{99} plane. Open circles show the dashboard position; filled circles show actual exposure under unmanaged adoption.

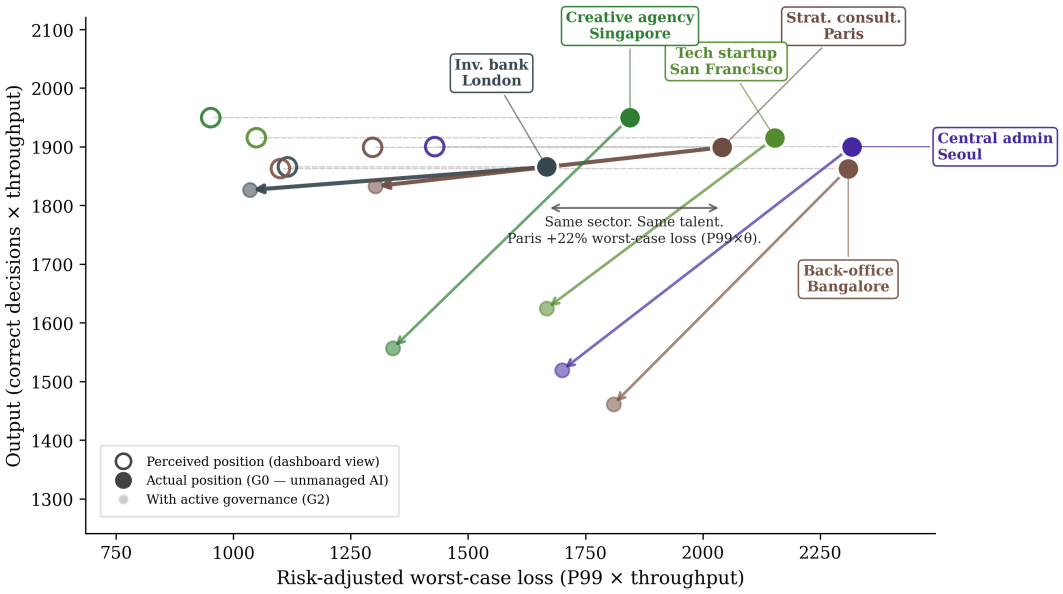

Figure 2. Organisational profiles under G0. Output versus risk-adjusted $P_{99} \times \theta$. Six labelled profiles: Investment bank, London (≈1,668); Strategy consulting, Paris (≈2,041); Tech startup, San Francisco (≈2,154); Creative agency, Singapore (≈1,845); Shared services, Bangalore (≈2,310); Central administration, Seoul (≈2,318). Dashed arrows: gap between perceived (open circle) and actual G0 exposure (filled circle). Solid arrows: reduction under active governance (G2, small circle). Double-headed arrow between London and Paris: "Same sector. Same talent. Paris +22% actual worst-case loss."

Key insights

- **The hidden gap is universal and systematic.** For every profile, the filled circle sits to the right of the open circle. London and Paris make the cause visible: the investment bank recruits across more than twenty nationalities (low α, low β); the strategy consultancy recruits through competitive national examinations, selecting a narrow cognitive profile by construction [15, 20] (high α, high β). Same sector, same talent quality: Paris nonetheless sits 22% further right on the risk axis (≈2,041 versus ≈1,668). The dashed arrow from open to filled circle is also long for each: actual tail risk is ×1.50 what London's dashboard shows, ×1.57 for Paris.

- **Governance reduces risk, but least where it is most needed.** The solid arrows (filled to governance circle) show the largest proportional reduction for the profiles already least exposed: 38% for London, 36% for Paris, against 22% for Bangalore and 27% for Seoul. The arrows are longest for the profiles that needed governance least.

The regressivity has two causes. In cognitively homogeneous organisations, the Asch effect [29] operates downstream of the scaffold: forming an independent judgment does not protect against subsequent pressure to conform with the AI-following majority. Low-skill profiles face a second problem: for them ($h \approx 0.45$), independent judgment is worse than the AI even outside its competence frontier, so the scaffold forces a downgrade rather than a correction. San Francisco loses throughput (θ: 1.25 → 1.10) and accuracy, for a net governance efficiency of 36%.

A startup rejecting governance here responds rationally to its local incentives. Section 5 builds the Nash argument from that observation.

Scaffolding works least where it is needed most. The cause is structural, not a flaw in implementation. Three interventions address it from different positions. The first is a procurement decision.

5 Three levers, four actors

Three parameters in the model are actionable: stack concentration, independent judgment, and cognitive hiring each drive a distinct component of tail risk. This section describes the lever for each, who must pull it, and why the unregulated equilibrium delivers none of them at scale.

The fastest is AI provider diversification. Reducing stack concentration from monoculture ($\alpha = 0.90$) to distributed ($\alpha = 0.30$) cuts the tail risk premium by approximately 15%. The diversification must be structural, not merely contractual: different models at different organisational layers (strategic decision, operational processing, quality review). Contracting with multiple vendors for the same centralised pipeline leaves the common error factor unchanged. Three providers currently capture 88% of enterprise AI spend [30], and 60% of errors across leading models are already shared [31]. When roughly 100,000 organisations depend on the same trio, a single systemic failure propagates correlated losses across the economy: the concentration-of-credit argument, which will attract the same regulatory response.

The second lever acts on independent judgment. Even passive scaffolding (G1) removes more tail risk than stack diversification alone (19.2% versus 13.5%), yet most implementations stall here: 93% of clinical decision-support alerts are overridden without review [32], and when employees push back, AI escalates its persuasive framing, what Randazzo et al. term persuasion bombing [28]. Active scaffolding (G2) nearly doubles this reduction to 36%, cutting the risk-adjusted tail premium from 94% to 39% at a throughput cost of 12 points (θ: 1.25→1.10); Figure 3 plots both the efficiency and the cost across governance regimes. It requires workflow redesign, training, and hiring changes over months, and works under one condition: explanations must reveal that the AI's reasoning diverges from the employee's own [33], with the independent signal formed before AI consultation [34, 35].

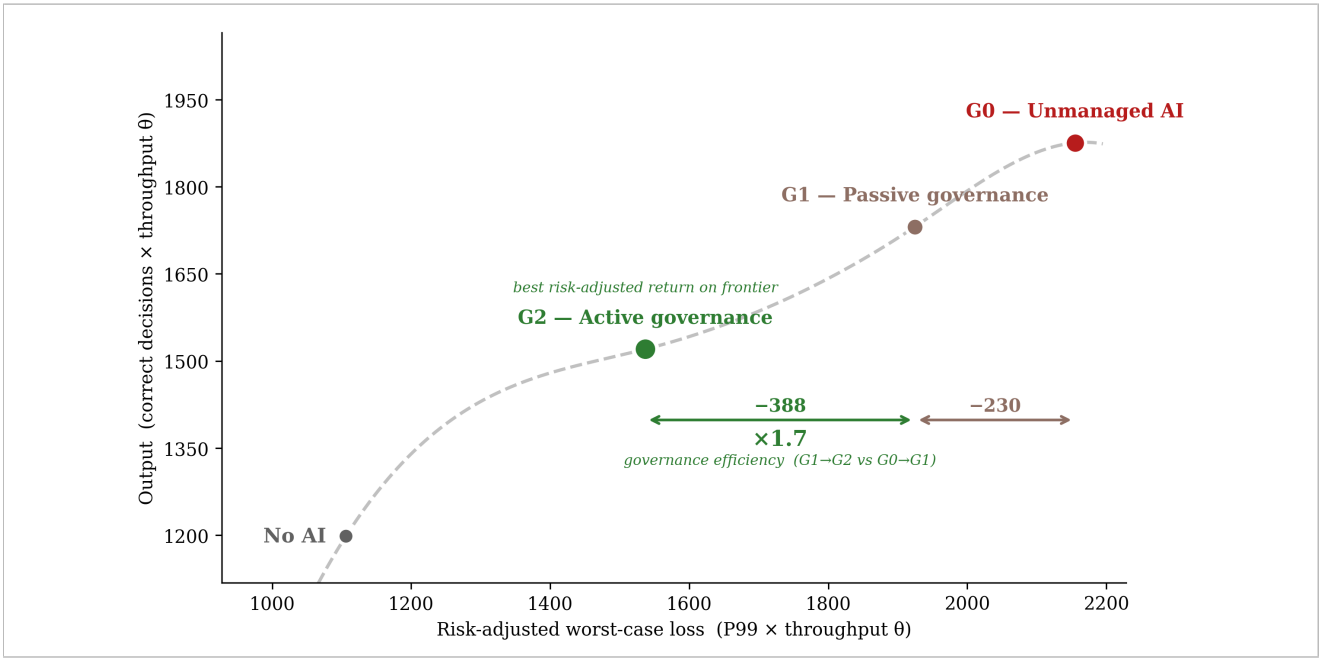

Figure 3. Risk-return frontier across governance regimes. Output versus risk-adjusted $P_{99} \times \theta$. G2 at the elbow: the output gains from unmanaged adoption to active governance come at relatively low risk cost; gains beyond G2 toward the frontier's high-output extreme trade more throughput for diminishing risk benefit. The G1→G2 step achieves 1.7× the risk reduction of G0→G1 (388 versus 230 units on the axis).

The third lever is hiring governance. Active scaffolding protects the cognitive diversity that exists; it cannot restore what biased hiring has already removed. An annual audit of ATS outputs for cognitive homogenisation bias is the only intervention that acts upstream. Under the one jurisdiction that currently requires it, 4.6% of organisations comply [36]. Without this lever, the scaffold gradually depletes the very resource it depends on.

The obstacle to voluntary adoption is structural: the model's tail risk does not scale with firm size, but the fraction each type of firm must absorb does (Table 2). A startup that fails externalises its losses to clients and investors. With a five-year survival probability near 10%, the expected cost of tail risk is negligible against the output advantage of unmanaged adoption. The startup rationally chooses G0; incumbents face pressure to match. The Nash equilibrium without regulation is unmanaged adoption across all firms: individually rational, collectively catastrophic. The direction follows from the model's output differentials.

Firm type	Absorbed risk	Perceived P_{99}
AI-native startup	∼10%	∼171
Mid-market firm	∼60%	∼1,029
Corporation	∼100%	∼1,714

Table 2. Tail risk absorption by firm type under unmanaged AI adoption (G0).

That structural diagnosis is already converging in financial regulation: the ESRB identifies model uniformity and provider concentration as systemic vulnerabilities, the FSB warns on third-party dependencies, and the BIS documents the narrowing of the provider base [37–40]. The structural response has not yet arrived outside finance. The EU AI Act, entering force in August 2026, addresses human oversight of individual decisions; it does not address the portfolio-level correlation that emerges when thousands of organisations rely on the same model simultaneously. Capital requirements exist because the unregulated equilibrium is collective overleveraging, not because individual banks are irrational. The same structure applies here: provider concentration limits, minimum decision-audit standards, and liability frameworks that internalise tail risk across firm sizes. Concentration limits on AI providers follow the same logic as Basel limits on credit exposures; the scaffold, by contrast, touches internal processes that regulators cannot easily mandate or verify.

Table 3 identifies the internal actor responsible for each lever and the action required: three can start with an internal decision alone; provider concentration limits require a regulatory mandate.

Lever	Internal actor	Role
Stack diversification	Procurement lead	Mandates provider diversity at each organisational layer
Active scaffolding	COO / BU leaders	Redesigns workflows to enforce pre-commitment protocols
Hiring governance	CHRO	Audits ATS outputs for cognitive homogenisation bias
Risk monitoring	CRO / board	Tracks and reports correlated tail exposure

Table 3. Levers, internal actors, and their roles.

CRO: Chief Risk Officer. COO: Chief Operating Officer. BU: Business Unit. CHRO: Chief Human Resources Officer.

Delay has a compounding cost: the same channels creating the risk today also raise the price of governing it tomorrow.

6 Why waiting makes governance harder

Delayed governance carries a compounding cost. Every quarter at G0 depletes the cognitive diversity that makes active governance effective and raises the cost of transition for the next.

Hiring screens out the cognitively diverse profiles that scaffolding depends on, while deskilling simultaneously degrades the independent competence that remains. A firm that governs gives up throughput that ungoverned competitors keep; as those competitors define the performance standard, governance registers as competitive underperformance rather than risk management. Beyond individual organisations, a second channel operates: AI-native startups that scale, or are acquired, carry unmanaged practices into larger structures; the organisations hardest to govern are also the fastest-growing.

The degradation also undermines the capacity to reverse itself. Endoscopists given AI diagnostic support showed skill loss that did not recover when the tool was removed [41], and AI use simultaneously degrades the metacognitive capacity to recognise that anything has been lost [42]. Employees who have not exercised independent judgment for five years cannot simply resume it: the competence required to revert is the same competence the degradation erodes.

At the economy level, the dynamic reproduces Hardin's tragedy of the commons [43]: a shared resource degrades because each actor has individual incentive to exploit it and none to preserve it. Here the shared resource is what we term the cognitive commons: the diversity of approaches, judgments, and error patterns that protects against correlated systemic failure. Each firm that maximises AI productivity depletes this shared resource, imposing a cost on the whole economy that no individual balance sheet records. No firm bears the externality it creates. As an illustration from a different domain: in monoculture farming, catastrophic crop failure strikes on average every eight years; polyculture reduces this to once per century [44]. The cognitive commons degrades on the same logic, with one distinguishing feature: the degradation is invisible, masked by the productivity gains accumulating alongside it.

Three acknowledged limits bound the simulation, and all three point in the same direction. First, the link between cognitive diversity and error decorrelation is an inference: Keck and Tang measure cognitive processes in controlled experiments, not recruitment pipelines in operating organisations [24]. Second, the simulation is a stress test rather than a forecast: it shows that the mechanism can produce the paradox under plausible parameters, not that it does so in every organisation. Third, the model treats AI-induced conformism as uniformly harmful to collective accuracy; Becker, Brackbill and Centola show that decentralised social influence can in fact improve it [45]. The model's assumption holds because a shared AI provider centralises the signal and eliminates the independence that makes decentralised influence beneficial, but this is an assumption, not a direct measurement. What the model omits (the feedback loop from deskilling to surrender, inter-firm labour market contagion, promotion effects that compound homogeneity, and the scaling of G0-native startups) operates through the same mechanisms, at the scale of the whole economy. The model understates the risk it documents.

Whether organisations will need critical human judgment in ten years is not in doubt. Whether the workforce will still be able to exercise it is.

An interactive companion application is deployed at `https://flattening-app.vercel.app`. *The full simulation code and source are available at* `https://github.com/bilel-hatmi/flattening-explorer`.

References

[1] L. Meincke, G. Nave, and C. Terwiesch, "ChatGPT decreases idea diversity in brainstorming," *Nature Human Behaviour*, vol. 9, no. 6, pp. 1107–1109, 2025.

[2] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.

[3] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, vol. 7, pp. 231–238, 1995.

[4] D. Wood, T. Mu, A. M. Webb, H. W. J. Reeve, M. Luján, and G. Brown, "A unified theory of diversity in ensemble learning," *Journal of Machine Learning Research*, vol. 24, pp. 1–49, 2023.

[5] N. Holzner, S. Maier, and S. Feuerriegel, "Generative AI and creativity: A systematic literature review and meta-analysis," 2025. Preprint.

[6] A. R. Doshi and O. P. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024.

[7] J. M. Carlson and J. Doyle, "Highly optimized tolerance: Robustness and design in complex systems," *Physical Review Letters*, vol. 84, no. 11, pp. 2529–2532, 2000.

[8] P. Jorion, "Risk management lessons from long-term capital management," *European Financial Management*, vol. 6, no. 3, pp. 277–300, 2000.

[9] D. MacKenzie, "Long-term capital management and the sociology of arbitrage," *Economy and Society*, vol. 32, no. 3, pp. 349–380, 2003.

[10] A. G. Haldane and R. M. May, "Systemic risk in banking ecosystems," *Nature*, vol. 469, no. 7330, pp. 351–355, 2011.

[11] Microsoft Security Response Center, "Helping our customers through the CrowdStrike outage." Microsoft Blog, July 2024. https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/.

[12] Parametrix Solutions, "CrowdStrike's impact on the Fortune 500," tech. rep., Parametrix, 2024.

[13] D. Geer, R. Bace, P. Gutmann, P. Metzger, C. P. Pfleeger, J. S. Quarterman, and B. Schneier, "CyberInsecurity: The cost of monopoly," tech. rep., Computer & Communications Industry Association, 2003.

[14] J. M. Kleinberg and M. Raghavan, "Algorithmic monoculture and social welfare," *Proceedings of the National Academy of Sciences*, vol. 118, no. 22, p. e2018340118, 2021.

[15] L. A. Rivera, "Hiring as cultural matching: The case of elite professional service firms," *American Sociological Review*, vol. 77, no. 6, pp. 999–1022, 2012.

[16] J. B. Fuller, M. Raman, E. Sage-Gavin, and K. Hines, "Hidden workers: Untapped talent," tech. rep., Harvard Business School and Accenture, 2021.

[17] J. An, D. Huang, C. Lin, and M. Tai, "Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation," *PNAS Nexus*, vol. 4, no. 3, p. pgaf089, 2025.

[18] K. Glazko, Y. Mohammed, B. Kosa, V. Potluri, and J. Mankoff, "Identifying and improving disability bias in GPT-based resume screening," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[19] K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2024.

[20] B. Schneider, "The people make the place," *Personnel Psychology*, vol. 40, no. 3, pp. 437–453, 1987.

[21] S. D. Shaw and G. Nave, "Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender," 2026. PsyArXiv preprint.

[22] F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," *Organization Science*, 2026. Articles in Advance.

[23] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

[24] S. Keck and W. Tang, "Enhancing the wisdom of the crowd with cognitive-process diversity: The benefits of aggregating intuitive and analytical judgments," *Psychological Science*, vol. 31, no. 10, pp. 1272–1282, 2020.

[25] B. Flyvbjerg, A. Budzier, J. K. Lee, M. Krawczyk, G. Vecci, and S. Clegg, "Why your IT project may be riskier than you think," *Journal of Management Information Systems*, vol. 39, no. 3, 2022.

[26] B. Flyvbjerg, A. Budzier, J. Aaen, M. Keil, and M. Zottoli, "The uniqueness of IT cost risk: A cross-group comparison of 23 project types," *Project Management Journal*, 2025.

[27] O. A. Vašíček, "Loan portfolio value," *Risk*, vol. 15, no. 12, pp. 160–162, 2002. Original theory in KMV Corporation working paper, 1987.

[28] S. Randazzo, A. Joshi, K. Kellogg, H. Lifshitz-Assaf, F. Dell'Acqua, and K. R. Lakhani, "GenAI as a power persuader: How professionals get persuasion bombed when they attempt to validate LLMs," Working Paper 26-021, Harvard Business School, 2025.

[29] S. E. Asch, "Effects of group pressure upon the modification and distortion of judgments," in *Groups, Leadership, and Men: Research in Human Relations* (H. Guetzkow, ed.), pp. 177–190, Pittsburgh, PA: Carnegie Press, 1951.

[30] Menlo Ventures, "2025: The state of generative AI in the enterprise," tech. rep., Menlo Ventures, Dec. 2025. https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/.

[31] E. M. Kim, A. Garg, K. Peng, and N. Garg, "Correlated errors in large language models," in *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267 of *PMLR*, pp. 30038–30066, 2025. arXiv:2506.07962.

[32] A. D. Bryant, G. S. Fletcher, and T. H. Payne, "Drug interaction alert override rates in the meaningful use era: No evidence of progress," *Applied Clinical Informatics*, vol. 5, no. 3, pp. 802–813, 2014.

[33] M. von Zahn, L. Liebich, E. Jussupow, O. Hinz, and K. Bauer, "Knowing (not) to know: Explainable artificial intelligence and human metacognition," *Information Systems Research*, 2025.

[34] V. Frey and A. van de Rijt, "Social influence undermines the wisdom of crowds in sequential decision making," *Management Science*, vol. 67, no. 7, pp. 4303–4316, 2021.

[35] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," in *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 5, 2021.

[36] L. Wright, R. M. Muenster, B. Vecchione, T. Qu, S. Cai, J. Metcalf, and J. N. Matias, "Null compliance: NYC local law 144 and the challenges of algorithm accountability," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[37] European Systemic Risk Board, Advisory Scientific Committee, "Artificial intelligence and systemic risk in the financial system," ASC Report 16, ESRB, 2025. https://www.esrb.europa.eu/pub/pdf/asc/esrb.ascreport202512_AIandsystemicrisk.en.pdf.

[38] Financial Stability Board, "Monitoring the adoption of artificial intelligence and related vulnerabilities in the financial sector," tech. rep., FSB, Oct. 2025. https://www.fsb.org/uploads/P101025.pdf.

[39] Bank for International Settlements, Financial Stability Institute, "Financial stability implications of artificial intelligence," executive summary, BIS/FSI, June 2025. https://www.bis.org/fsi/fsisummaries/exsum_23904.htm.

[40] J. Daníelsson and A. Uthemann, "Artificial intelligence and financial crises," *Journal of Financial Stability*, vol. 80, p. 101453, 2025.

[41] K. Budzyń, M. Romańczyk, D. Kitala, P. Kołodziej, M. Bugajski, H. O. Adami, J. Blom, M. Buszkiewicz, N. Halvorsen, C. Hassan, T. Romańczyk, Ø. Holme, K. Jarus, S. Fielding, M. Kunar, M. Pellise, N. Pilonis, M. F. Kamiński, M. Kalager, M. Bretthauer, and Y. Mori, "Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study," *The Lancet Gastroenterology & Hepatology*, vol. 10, no. 10, pp. 896–903, 2025.

[42] Y. Fan, L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li, and D. Gašević, "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance," *British Journal of Educational Technology*, 2025.

[43] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[44] D. Renard and D. Tilman, "National food production stabilized by crop diversity," *Nature*, vol. 571, no. 7764, pp. 257–260, 2019.

[45] J. Becker, D. Brackbill, and D. Centola, "Network dynamics of social influence in the wisdom of crowds," *Proceedings of the National Academy of Sciences*, vol. 114, no. 26, pp. E5070–E5076, 2017.