EIG | ENTERPRISE INTELLIGENCE GOVERNANCE

Enterprise Intelligence Governance

REBUILDING ORGANISATIONAL CHALLENGE FOR MACHINE-GENERATED INFERENCE

Machine inference makes people most confident where they are least accurate. That inverts the signal organisations rely on to decide which decisions to challenge.

STRATEGIC REPORT

August 2026

Olga Pisarenko

Contents

PART I — THE DEPLOYMENT AND THE GAP

1. The Deployment and the Gap
2. What the Evidence Shows

PART II — THE MECHANISM

3. The Confidence–Competence Inversion
4. Why Existing Answers Do Not Reach It
5. Three Classes of Machine Inference

PART III — GOVERNANCE

6. Enterprise Intelligence Governance
7. Class-Triggered Challenge

PART IV — THE FINANCE FUNCTION AND CAPITAL ALLOCATION

8. The Finance Function as Arbiter of Authority
9. Capital Allocation under Artificial Intelligence

PART V — PROPOSITIONS, BOUNDARIES AND RESEARCH AGENDA

10. Propositions and Research Agenda
11. Boundary Conditions and Limitations
12. Conclusion

APPENDIX AND SOURCES

A. The Coded Corpus

Bibliography

Abstract

Artificial intelligence is being deployed at a scale without close precedent: $581.7 billion of global corporate investment in 2025, and hyperscaler capital expenditure guided to between $650 billion and $760 billion for 2026.[1] The returns have not followed. Nine in ten senior executives report no effect of AI on productivity or employment over three years;[2] euro-area data show no statistically significant productivity difference between adopters and non-adopters;[3] and the most careful macroeconomic accounting puts the ten-year total factor productivity effect below 0.7 per cent.[4] Yet where the technology has been tested experimentally at the level of the task, the effects are large, replicable and positive. This paper is about the distance between those two facts.

This paper assembles a pattern in the experimental literature that has not been put together before, names the mechanism that accounts for it, and derives a governance failure from the two. Thirty-five randomised and quasi-experimental trials of generative-artificial-intelligence assistance were coded on the hardest outcome each of them reports. Measured gains do *not* decline monotonically as outcome measures harden. They collapse at one band — objective outcomes carrying a real external consequence, where four of thirteen studies report a positive significant effect against roughly two-thirds elsewhere — and the collapse concentrates in *decision* tasks, not in creation or learning.[5] A large independent meta-analysis finds the same split with adequate power.[6] The mechanism proposed for it is that in a decision task the person best placed to notice the shortfall does not notice it: every study that has elicited participants' beliefs about their own measured performance alongside machine assistance finds miscalibration in the optimistic direction.[7] The paper names this the *confidence–competence inversion* and states it conditionally: machine-generated inference decouples confidence from accuracy and, beyond the system's competence, reverses the relationship between them — wherever the environment supplies no signal capable of registering the error.

The organisational consequence is specific, and once stated it is hard to unsee. Challenge in organisations divides in two. Some reviews are mandated by class and fire whatever anyone believes. The rest — the discretionary margin, where most decisions live — are *self-triggered*: independent review, pre-mortem analysis and referral upward are requested when a decision-maker feels uncertain. If confidence rises where accuracy falls, that trigger fires in the wrong direction. Arrangements adequate for human-generated intelligence then become structurally inadequate for machine-generated inference, and the organisation experiences the failure as good governance.

[1] Stanford Institute for Human-Centered Artificial Intelligence, *AI Index Report 2026*, Chapter 4: Economy (April 2026); Apollo Global Management estimate reported in *Fortune*, 23 February 2026.

[2] Ivan Yotzov and others, 'Firm Data on AI', NBER Working Paper 34836 (February 2026).

[3] Annalisa Ferrando and others, 'Adoption and Investment in Artificial Intelligence across the Euro Area', ECB Occasional Paper No. 395 (2026).

[4] Daron Acemoglu, 'The Simple Macroeconomics of AI', *Economic Policy*, 40 (121) (2025), 13–58.

[5] Coding by the author; outcomes as published. The task-family comparison at hard measures gives three of ten decision studies positive against four of seven non-decision studies, a difference not statistically distinguishable at this sample size (Fisher exact, two-sided, p = 0.35).

[6] Michelle Vaccaro, Abdullah Almaatouq and Thomas Malone, *Nature Human Behaviour*, 8 (2024), 2293–2303: 106 studies, 370 effect sizes, pooled Hedges' $g = -0.23$ (CI −0.39 to −0.07), losses concentrated in decision tasks.

[7] Joel Becker and others, arXiv:2507.09089 (2025); Bastani and others, SSRN 4895486; Raja Parasuraman and Dietrich H. Manzey, 'Complacency and Bias in Human Use of Automation', *Human Factors*, 52 (3) (2010), 381–410.

> *“ Organisations escalate decisions when the decision-maker feels uncertain. Where nothing can register the error, machine-generated inference leaves them most confident where they are least accurate. The trigger fires in the wrong direction.*

The claim is located inside organisational information processing theory, not against it. Galbraith’s hierarchy is employed on an exception basis: participants refer upward the situations their rules do not cover, and that referral is the organisation’s detection mechanism.[8] Machine-generated inference supplies fluent, apparently rule-covered output for cases that are in fact exceptions, so the perceived exception rate falls below the true one and the hierarchy is under-employed, not overloaded. That objective and perceived uncertainty diverge is long established;[9] what is added is a specific instrument that induces the divergence in a stated direction, moderated by the fluency of the output and the inspectability of its derivation.

From this I develop *Enterprise Intelligence Governance (EIG)*: the structures, decision rights and accountability mechanisms through which an organisation determines which intelligence carries authority, which decisions receive challenge and who owns the resulting judgement. It is offered as a complement to the existing artificial-intelligence capability construct, not a competitor to it. Its central prescription is a shift from *confidence-triggered* to *class-triggered* challenge: mandatory independent review determined ex ante by consequence, reversibility and *outcome verifiability*, the third of which is absent from the schemes for allocating challenge reviewed here — delegated authority matrices, enterprise risk frameworks, model risk tiering and the auditing standards.

The regulatory position makes this timely and the gap documentable. Model risk management has been the institutional answer to governing machine-generated numbers since 2011. On 17 April 2026 that guidance was superseded, the requirement for structurally independent validation was relaxed, and generative and agentic artificial intelligence were expressly placed outside its scope.[10]

Keywords: artificial intelligence governance; automation bias; decision rights; information processing theory; model risk; corporate governance; capital allocation; chief financial officer; organisational challenge.

[8] Jay R. Galbraith, ‘Organization Design: An Information Processing View’, *Interfaces*, 4 (3) (1974), 28–36, at 29.

[9] Robert B. Duncan, ‘Characteristics of Organizational Environments and Perceived Environmental Uncertainty’, *Administrative Science Quarterly*, 17 (3) (1972), 313–327; Brian K. Boyd, Gregory G. Dess and Abdul M. A. Rasheed, ‘Divergence between Archival and Perceptual Measures of the Environment’, *Academy of Management Review*, 18 (2) (1993), 204–226.

[10] Board of Governors of the Federal Reserve System, SR 26-2; Office of the Comptroller of the Currency, Bulletin 2026-13; Federal Deposit Insurance Corporation, FIL-15-2026, all 17 April 2026, superseding SR 11-7 and OCC Bulletin 2011-12 (4 April 2011).

Executive Summary

Boards are being asked to approve artificial-intelligence expenditure at a scale that would ordinarily attract forensic scrutiny, on the basis of evidence that would not survive it. What follows states what the evidence supports, why the value has not appeared, and the one change to governance that would most improve the odds of converting it.

THE DEPLOYMENT	THE GAP	THE REASON	THE GOVERNANCE
$581.7bn of corporate AI investment in 2025; hyperscaler capex guided above $650bn for 2026.	Nine in ten executives report no productivity effect; 39% report any EBIT impact.	Returns are conditional on organisational complements most firms have not built.	Which decisions receive independent challenge, and on what trigger.

The evidence, honestly read. Reported adoption runs from eleven to eighty-eight per cent depending entirely on the denominator, and every figure is correct for its own definition. Realised earnings impact is narrow: thirty-nine per cent of organisations report any enterprise-level effect on earnings before interest and tax, approximately six per cent report an effect of five per cent or more, and fourteen per cent of chief executives have defined the profit-and-loss impact of all their AI initiatives. Three of the four most-quoted failure statistics in circulation are not research findings: one is a media figure, two are forecasts, and the widely repeated claim that ninety-five per cent of AI initiatives return nothing derives from conference-intercept sampling that conflates firms which never piloted with firms whose pilots failed.

The technology is not the problem. Under randomised and staggered-rollout designs, artificial intelligence raises issues resolved per hour in customer support by fifteen per cent, cuts time on professional writing tasks by forty per cent, and raises completed tasks among software developers by twenty-six per cent. These are causal estimates from large samples inside operating firms, not self-reports.

The problem is that people cannot tell when it has failed. In a field experiment with 758 consultants, tasks inside the technology's competence produced materially more and better work; tasks outside it left users nineteen percentage points less likely to reach the correct answer, while the work they produced was rated *more* coherent and more persuasive than the unaided control's. In a randomised trial, experienced developers were nineteen per cent slower with AI assistance, had forecast a twenty-four per cent speed-up, and afterwards still believed they had been twenty per cent faster. Four decades of human-factors research find the same pattern in experts, find it resistant to training, and find it reduced by accountability for accuracy in student samples though not in expert ones. Failure, in this setting, does not announce itself.

> *" Developers were nineteen per cent slower and believed they were twenty per cent faster. Any measurement regime that asks people how much time they saved will report a return that does not exist.*

Why this breaks governance. Organisations do not challenge every decision; they challenge the ones that feel uncertain, and that triage is performed by the person advancing the decision. When confidence and accuracy move in opposite directions, the organisation systematically declines to examine precisely the decisions that most need examining — and it does so while feeling well governed, because nothing at any point generates friction.

What should change. Challenge must be triggered by *decision class*, not by felt confidence. The board, not the proposer, decides in advance which categories of decision receive mandatory independent review, and decides it on consequence, reversibility and outcome verifiability rather than on how sure anyone is. This is a reallocation of review, not an increase in it. Three further changes follow: no artificial-intelligence benefit should be recognised on the basis of self-reported time savings; the organisational complement to an AI investment should be appraised alongside the asset, since the historical evidence suggests it is the larger number; and the useful-life assumption applied to AI infrastructure should be treated as an active judgement, given that comparable firms currently differ by a full year and one has shortened while others extended.

Where this leaves the finance function. When the marginal cost of producing a fluent, quantitative, confident case for any proposal approaches zero, the scarce organisational function is no longer analysis but adjudication between analyses. That is the function finance has always performed. The chief financial officer's role under artificial intelligence is not primarily to produce better numbers but to determine which numbers carry authority — a governance role, and the reason the challenge architecture proposed here belongs in finance, not in technology.

PART I

01

The Deployment and the Gap

Historic capital deployment, large task-level effects, and almost no measurable value in firm accounts.

01 **The scale**
$581.7bn of corporate AI investment in 2025; hyperscaler capex guided above $650bn for 2026.

02 **The evidence, honestly read**
Nine in ten executives report no productivity effect; the task-level evidence is strong.

03 **Three statistics that are not research**
The most-quoted failure numbers are forecasts, media figures and conference intercepts.

Part I — The Deployment and the Gap

1. The Deployment and the Gap

Global corporate investment in artificial intelligence reached $581.7 billion in 2025, an increase of approximately 130 per cent on the prior year.[11] Enterprise expenditure on generative AI software and interfaces rose from $11.5 billion to $37 billion.[12] Capital expenditure by the four largest hyperscale providers reached approximately $410 billion in 2025 and is guided to between $650 billion and $760 billion for 2026, on one estimate close to two per cent of United States gross domestic product.[13] Investment in information-processing equipment and software, some four per cent of that economy, accounted for the great majority of measured growth in the first half of 2025.[14]

Against that deployment, measured return is difficult to locate. This chapter states the gap; Chapter 2 examines the evidence in detail; Chapter 3 proposes the mechanism that accounts for it.

The Question

Existing treatments of this problem divide into two unsatisfying families. The first holds that artificial intelligence is over-valued and the capital will not earn its cost, a position with serious support, examined in Chapter 2, but one that sits uneasily with the strength of the task-level experimental evidence. The second holds that firms are simply early, and that returns will arrive with scale — a position that is unfalsifiable in its usual form and that has been the standing answer for three years.

This paper takes a third position. It asks: under what conditions do organisational challenge mechanisms designed for human-generated intelligence fail when applied to machine-generated inference, and what governance architecture restores their function? This is narrower than asking what determines AI value, and it has the advantage of being answerable from evidence that already exists.

Terms

Three definitions are needed at the outset, because each term is used loosely in this literature.

> **Intelligence**, in this paper, means inference offered as a basis for decision, whatever its source: human judgement, tacit operational knowledge, external signal or model output. The paper is concerned with the arbitration between these when they conflict.
>
> **Machine-generated inference** means a claim, prediction or recommendation produced by a computational system, offered for use in a decision, whose derivation is not fully inspectable by the person acting on it.

[11] Stanford HAI, *AI Index Report 2026*, Chapter 4: Economy.

[12] Menlo Ventures, *2025: The State of Generative AI in the Enterprise* (December 2025). Vendor survey of approximately 500 United States decision-makers; indicative rather than authoritative.

[13] Company guidance compiled from quarterly disclosures; Apollo Global Management estimate reported in *Fortune*, 23 February 2026. Guidance was revised upward repeatedly through the first half of 2026.

[14] Jason Furman, analysis of United States national accounts, October 2025. The attribution methodology is contested; the direction is not.

Enterprise value has a technical meaning in finance — the market value of equity plus net debt, adjusted for minority interests and associates. Where this paper uses the phrase *long-run economic value* it means value to the providers of capital over a horizon longer than the reporting cycle. Where the technical measure is intended it is named as such.

Contribution and Method

The paper is conceptual and integrative, in the tradition of construct development through synthesis of adjacent literatures prior to empirical testing.[15] It presents no new data. It differs from much conceptual work in three respects the reader is entitled to hold it to: every empirical claim is sourced and the reliability class of each source is characterised; the propositions carry a direction, a conditioning variable and a proposed measurement proxy, so that each can fail; and the case against the argument is put in Chapter 2, before the framework is built, and not in a concluding caveat.

Because a conceptual paper can slide between what is known and what is being proposed, the register below states which is which. Nothing after it should be read as carrying more warrant than the row it belongs to.

THE WARRANT REGISTER

What is established, what is newly assembled, what is argued — and what would sink the argument.

STATUS	CLAIM	WHERE IT IS MADE
Established *What the evidence already supports*	Task-level assistance raises measured output in well-identified field experiments. Firm-level effects are small, lagged and concentrated among firms that already hold the organisational complements. Individual miscalibration in the optimistic direction is documented, and four decades of human-factors research document automation bias in naive and expert operators alike.	Chapters 2 and 3, every claim sourced
Synthesis *New assembly of existing results*	Measured gains do not fall monotonically as measures harden. They collapse at one band — objective outcomes carrying a real external consequence — and the collapse concentrates in decision tasks. Assembled here for the generative-artificial-intelligence trials specifically.	Chapter 2 and Appendix A
Proposition *Argued, not demonstrated*	The confidence–competence inversion, and its aggregation from individual miscalibration into misallocated organisational challenge. Four conditions are jointly required and the fourth is currently the weakest.	Chapters 3 and 10, P1 to P3
Prescription *What follows for design*	Challenge triggered by decision class instead of felt confidence, classified in advance on consequence, reversibility and outcome verifiability, at constant total challenge capacity.	Chapters 6 to 9
Falsification *What would sink the argument*	If the construct adds no explanatory power beyond organisation capital, measured management practice, measured artificial-intelligence capability and conventional governance indices, it is redundant and should be abandoned.	Proposition 4

HOW TO READ THE REST OF THIS PAPER

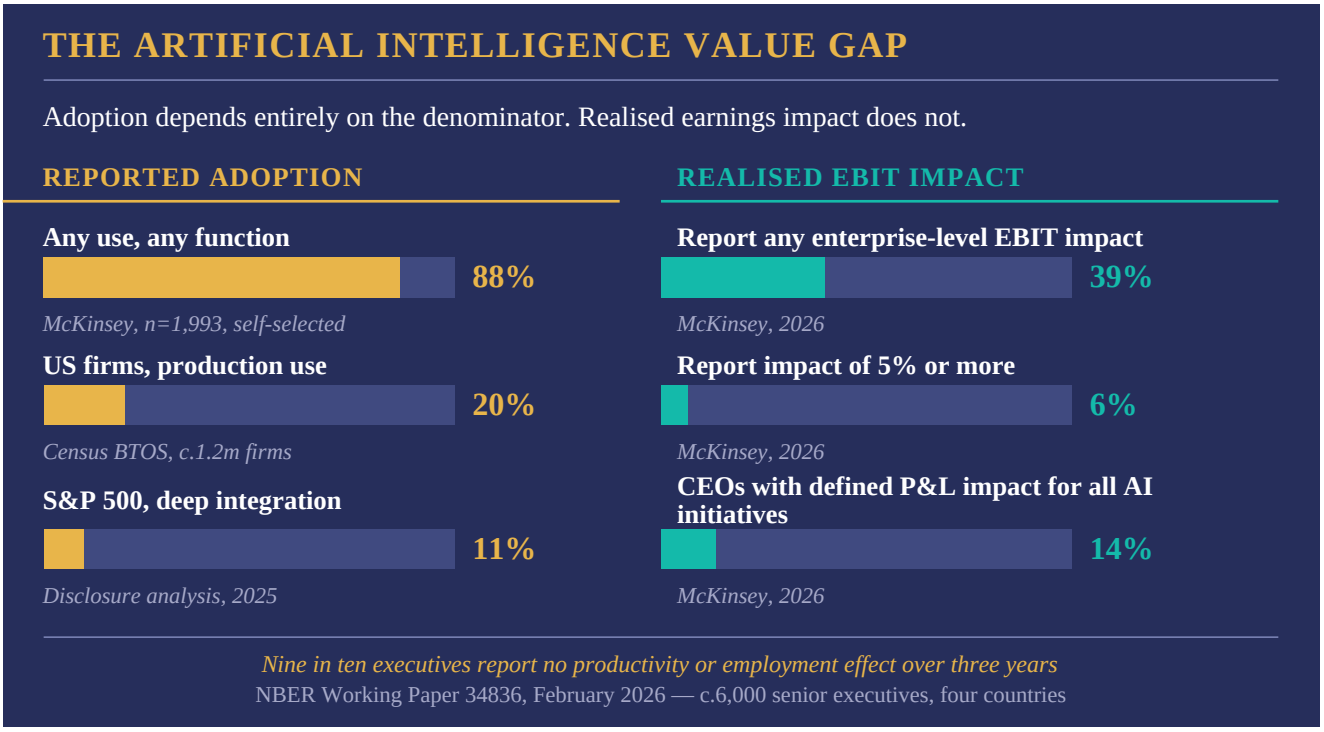
Nothing after this table carries more warrant than the row it belongs to.

[15] Elina Jaakkola, 'Designing Conceptual Articles: Four Approaches', *AMS Review*, 10 (1) (2020), 18–26; David A. Whetten, 'What Constitutes a Theoretical Contribution?', *Academy of Management Review*, 14 (4) (1989), 490–495.

2. What the Evidence Shows

Almost any position on artificial intelligence can be supported by selective citation. This chapter separates what the evidence supports from what is merely repeated.

EXHIBIT 1



Adoption Depends Entirely on the Denominator

Reported adoption in 2026 ranges from eleven to eighty-eight per cent. The range is definitional, not contradictory. McKinsey reports that eighty-eight per cent of organisations use artificial intelligence in at least one business function.[16] The United States Census Bureau, surveying approximately 1.2 million firms on a probability basis, reports 19.8 per cent using AI in producing goods or services as of May 2026, thirty-seven per cent among firms with 250 or more employees, 33.9 per cent in financial services.[17] Analysis of S&P 500 disclosures finds eleven per cent with AI deeply integrated into business processes, up from five per cent in 2022, with the technology sector accounting for two-thirds of that.[18]

[16] McKinsey & Company, *The State of AI* (November 2025), n=1,993, self-selected respondent base weighted toward large organisations; ‘any use in any function’ is a permissive threshold.

[17] United States Census Bureau, Business Trends and Outlook Survey, May 2026.

[18] Zhengyang Yu and others, ‘AI Adoption in S&P 500 Firms’, arXiv:2607.08920 (July 2026).

European data make the same point more sharply. Approximately seventy per cent of euro-area firms report some AI use; seven per cent report significant adoption.[19] In the United Kingdom, the share of businesses using at least one AI technology rose from around twelve per cent in 2023 to thirty-five per cent by mid-2026, while the average number of technologies per adopter moved only from 1.4 to 1.6.[20] Adoption is broad and shallow, and depth, not breadth, predicts effect.

ASK FOR THE DENOMINATOR

Three correct answers to one question

88%	Any use, any function, at self-selecting large firms
20%	US firms using AI in producing goods or services
11%	S&P 500 firms with deep integration into processes

Adoption is broad and shallow

Realised Impact, and Three Statistics That Are Not Evidence

Thirty-nine per cent of organisations report any enterprise-level effect on earnings before interest and tax; approximately six per cent report an effect of five per cent or more.[21] The most rigorously constructed evidence is the least encouraging: across approximately six thousand senior executives surveyed through central-bank decision-maker panels, nine in ten report no effect on productivity or employment over three years, with forward expectations of 1.4 per cent productivity and 0.8 per cent output.[22] European Central Bank analysis finds no statistically significant productivity difference between adopters and non-adopters in the euro area as a whole.[23]

Analysis of approximately 490,000 earnings-call transcripts from 5,198 United States public companies finds AI-related productivity commentary rising to roughly fifteen per cent of all productivity discussion by the end of 2025, of which approximately ninety-five per cent describes future gains, not realised ones.[24]

Three widely quoted failure statistics should not be used. The claim that ninety-five per cent of organisations obtain zero return from generative AI derives from a non-peer-reviewed study based on fifty-two interviews and 153 survey responses collected at four industry conferences, and conflates organisations that never piloted anything — approximately eighty per cent of the sample — with those whose pilots failed.[25] The claim that more than eighty per cent of AI projects fail is a media figure that a RAND report cites, not a RAND estimate.[26] Gartner's thirty and forty per cent abandonment figures are forecasts. A board governing its AI programme against these numbers is governing against noise.

[19] Ferrando and others, ECB Occasional Paper No. 395.

[20] Office for National Statistics, *Artificial Intelligence in UK Businesses: 2023 to 2026* (July 2026).

[21] McKinsey, *The State of AI* (November 2025). Self-reported.

[22] Yotzov and others, NBER Working Paper 34836.

[23] Ferrando and others, ECB Occasional Paper No. 395.

[24] Federal Reserve Bank of St. Louis, analysis published 31 July 2026.

[25] Aditya Challapally and others, *The GenAI Divide: State of AI in Business 2025* (MIT NANDA, July 2025).

[26] James Ryseff, Brandon F. De Bruhl and Sydne J. Newberry, *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed*, RR-A2680-1 (RAND Corporation, 2024). RAND's own contribution is qualitative and identifies leadership and problem-framing failures as the leading root cause.

At the Level of the Task, the Effects Are Large

EXHIBIT 2

WHERE ARTIFICIAL INTELLIGENCE DEMONSTRABLY WORKS

Causal task-level evidence from randomised and staggered-rollout designs.

TASK	EFFECT	ON WHAT MEASURE	EVIDENCE BASE
Customer support	+15%	issues resolved per hour *+30% for the least experienced*	5,172 agents, 3m chats
Professional writing	−40%	time taken *quality assessed 18% higher*	453 professionals, RCT
Software development	+26%	tasks completed *largest gains at short tenure*	4,867 developers, 3 firms
Consulting, inside frontier	+12%	tasks completed, 25% faster *and at higher assessed quality*	385 consultants, RCT
Consulting, outside frontier	−19pp	less likely to be correct *work rated more persuasive*	373 consultants, RCT
Experienced developers	−19%	slower on real issues *while believing they were 20% faster*	16 developers, 246 issues

The technology works at the level of the task. The value does not appear at the level of the firm.

A staggered rollout to 5,172 customer-support agents covering approximately three million conversations raised issues resolved per hour by fifteen per cent, and by approximately thirty per cent for the least experienced agents, with small or negative effects for the most experienced.[27] A randomised trial with 453 professionals on mid-level writing tasks reduced time taken by forty per cent and raised assessed quality by eighteen per cent.[28] Across three field experiments covering 4,867 software developers, completed tasks rose 26.1 per cent.[29] Seven randomised trials at a large online retailer produced a pre-sale conversational agent raising sales 16.3 per cent, alongside one arm with a statistically insignificant negative effect.[30]

Firm-Level Value: Real, Lagged, Concentrated

Firm AI investment measured from employee résumés and job postings across 1,052 firms is associated with 19.5 per cent higher sales growth and 22.3 per cent higher market value per standard deviation of AI-worker share. The effects operate through product innovation and not cost reduction, increase with

[27] Erik Brynjolfsson, Danielle Li and Lindsey R. Raymond, ‘Generative AI at Work’, *Quarterly Journal of Economics*, 140 (2) (2025), 889–942.

[28] Shakked Noy and Whitney Zhang, ‘Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence’, *Science*, 381 (6654) (2023), 187–192.

[29] Zheyuan Cui and others, ‘The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers’, *Management Science* (2026), online first.

[30] Yuqi Fang and others, ‘Generative AI and Firm Productivity: Field Experiments in Online Retail’, arXiv:2510.12049 (October 2025).

firm size, and appear with a two-to-three-year lag.[31] Analysis of S&P 500 adoption finds profitability following a J-curve with no observable differences in capital expenditure or productivity, and adoption concentrated among firms with higher Tobin's q.[32] The value is real, delayed, and so far concentrated where the organisational complements already existed.

Where the Measured Gains Are, and Are Not

The task-level literature is normally read as a single body of evidence about whether artificial intelligence helps. This section reads it differently, by what each study's outcome measure is able to see, and reports the result of doing so systematically. The result is not the one I expected.

A Coding Protocol

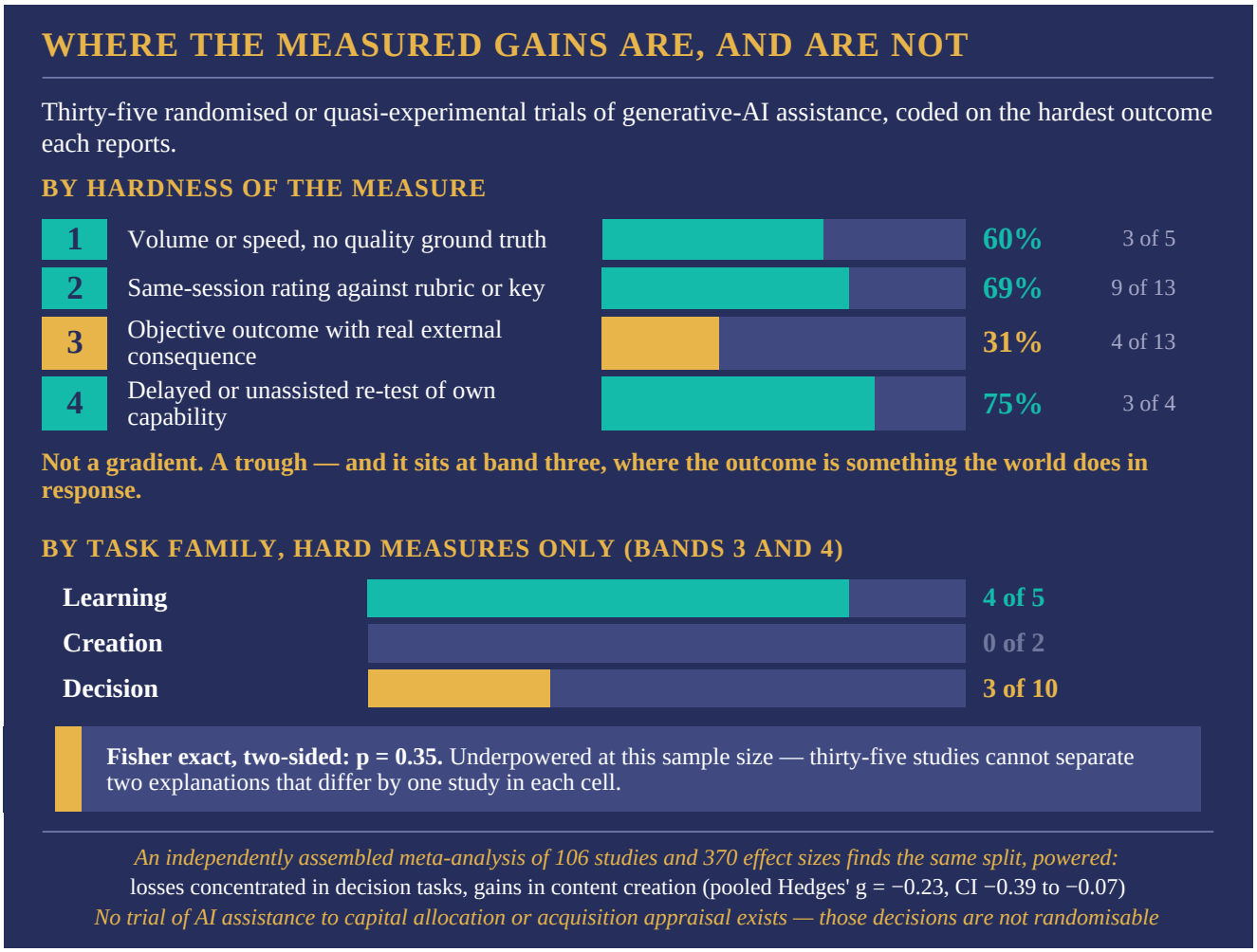
Every randomised trial, field experiment and staggered-rollout quasi-experiment of generative-artificial-intelligence assistance to human work that could be verified against a primary source, posted or published between 2023 and 2026, was assembled: thirty-eight studies, of which thirty-five report a result on their hardest measure and are analysed here.[33] Each was coded on the *hardest outcome it reports*, on a four-band scale: (1) volume or speed with no quality ground truth; (2) a same-session assessment against a rubric or a known answer; (3) an objective outcome carrying a real external consequence — a customer returned, a build passed, a sale closed, a patient outcome; (4) a delayed or unassisted re-test of the person's own capability. Each was also coded by task family: *creation*, *learning*, or *decision*. The effect on the hardest measure was recorded as positive and significant, null, or negative and significant.

[31] Tania Babina, Anastassia Fedyk, Alex He and James Hodson, 'Artificial Intelligence, Firm Growth, and Product Innovation', *Journal of Financial Economics*, 151 (2024), article 103745. The magnitudes quoted are ordinary-least-squares long-differences estimates over 2010–2018; the authors report that the results are robust to instrumenting artificial-intelligence investment with firms' ex-ante exposure to universities' subsequent supply of AI graduates, but the headline figures are not themselves instrumented.

[32] Yu and others, arXiv:2607.08920.

[33] The coded corpus, the coding rules and the excluded studies are available from the author. Bands and task families are the author's coding applied to outcomes reported by the original studies; the outcomes themselves are as published. Four studies were excluded on eligibility grounds: one because the tool is explicitly non-generative, one withdrawn by its repository over data-validity concerns, and two because they are observational rather than experimental. Three further studies were verified as existing but their hardest-measure result could not be established from an accessible version.

EXHIBIT 3



The Result, Including the Part That Fails

The intuitive hypothesis, that measured effects decline monotonically as outcome measures harden, is *not supported*. The share of studies reporting a positive significant effect on their hardest measure runs 60 per cent at band one, 69 per cent at band two, 31 per cent at band three, and 75 per cent at band four. The relationship is not a gradient. It is a trough, and it sits at band three.

The band-four figure explains itself once the studies are examined instead of counted. Band four holds four studies and all four measure learning. Three are tutoring trials in which artificial intelligence genuinely teaches: unassisted post-tests and end-of-term examinations show gains of 0.21 to 0.63 standard deviations.[34] Only the fourth finds the pattern I originally proposed, and it is the one in which the tool substituted for practice instead of supporting it.[35] A delayed unassisted measure is a hard test of

[34] Gregory Kestin and others, 'AI Tutoring Outperforms In-Class Active Learning', *Scientific Reports*, 15 (2025), article 17458; Martín E. De Simone and others, 'From Chalkboards to Chatbots', World Bank Policy Research Working Paper 11125 (2025); Zachary Contractor and Germán Reyes, 'Experimental Evidence on the Learning Impact of Generative AI', IZA Discussion Paper 18792 (2026); Rose E. Wang and others, 'Tutor CoPilot', arXiv:2410.03017 (2024).

[35] Hamsa Bastani and others, 'Generative AI Can Harm Learning', SSRN 4895486 (2024): practice performance rose 48 per cent with the system present and unassisted examination performance fell 17 per cent in the unrestricted arm.

learning. It is not a hard test of a *decision*, and conflating the two was the error in the earlier formulation of this argument.

What survives the correction is narrower and is about *consequence, not hardness*. Band three is the only band where positives are the minority: four of thirteen. These are the studies in which the outcome is neither a rating nor a stopwatch but something the world does in response: a customer's satisfaction score, a build in a production pipeline, a return, a retrial, a click, a firm's revenue. Customer sentiment in a five-thousand-agent support rollout: null. Build success across three developer experiments: −5.5 per cent, not significant. Three-day retrial rate at a large e-commerce operator: indistinguishable from zero. Click-through and cost-per-click on live advertising: no significant effect. Entrepreneurial revenue in a 640-firm trial: null overall and significantly negative for the weakest performers.[36] In two cases the hard measure is significantly negative: a human-supervised agentic deployment lowered customer ratings by 0.41 points, and expert users of a ghostwriting tool produced advertising that generated fewer real clicks.

Whether It Is the Consequence or the Task

Band three is disproportionately populated by decision and service tasks, so the two explanations have to be separated. Splitting the corpus by task family and restricting attention to the hard measures — bands three and four — gives three of ten decision studies reporting a positive significant effect against four of seven non-decision studies. The direction favours the task-family explanation over the measure-hardness one.

It is not statistically distinguishable at this sample size. A two-sided Fisher exact test on that table returns $p = 0.35$, and the honest description is that thirty-five studies cannot separate two explanations that differ by one study in each cell. I report the test alongside the counts because the counts on their own look stronger than they are.

The claim is nevertheless not unsupported, because a properly powered test of the same distinction already exists and was conducted independently. A meta-analysis of 106 experimental studies and 370 effect sizes finds human–machine combinations performing significantly worse than the better of human alone or system alone (a pooled Hedges' g of −0.23, with a confidence interval of −0.39 to −0.07) and finds the losses *concentrated in decision tasks* and the gains concentrated in content creation.[37] That is the same split, on a corpus large enough to detect it. The contribution of the coding exercise reported here is not to establish the split but to show that it persists in the generative-artificial-intelligence trials specifically, which postdate almost all of that meta-analysis.

[36] Erik Brynjolfsson, Danielle Li and Lindsey R. Raymond, 'Generative AI at Work', *Quarterly Journal of Economics*, 140 (2) (2025), 889–942; Zheyuan Kevin Cui and others, *Management Science* (2026), online first; Xiao Ni and others, arXiv:2603.29888 (2026); Harang Ju and Sinan Aral, arXiv:2503.18238 (2025); Nicholas G. Otis and others, *Management Science* (2026), DOI 10.1287/mnsc.2024.06909.

[37] Michelle Vaccaro, Abdullah Almaatouq and Thomas Malone, 'When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis', *Nature Human Behaviour*, 8 (2024), 2293–2303. Its search window closes on 30 June 2023, before most of the studies coded above were run.

> *“ The gains are real, and they are largest where the world is not asked to respond. Where it is asked, and where the task is a judgement, they mostly disappear.*

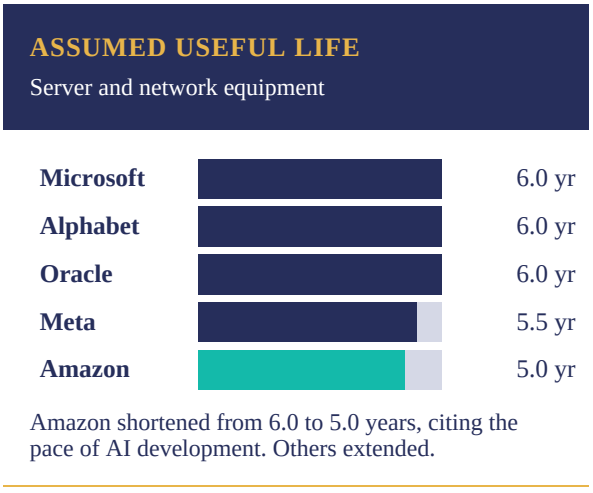
Three limitations belong with this and none is minor. Thirty-five studies is a small corpus and the test is underpowered. The band and family codings are mine; the outcomes coded are the studies’ own, but the classification is a judgement and a different coder would produce a different table at the margin. And the corpus is a convenience sample of what could be verified, which over-represents domains where trials are cheap — coding, education, customer service — and under-represents the decisions this paper is actually about. No randomised trial of generative-artificial-intelligence assistance to capital allocation, acquisition appraisal or market entry exists, and the reason is that those decisions are not randomisable at acceptable cost. That absence is not an oversight in the literature. It is the same property this paper is about, showing up as a gap in the evidence instead of a finding within it.

The Case Against This Paper

Three objections cut against all of this, and they are worth taking now.

The prize may be small. Task-based accounting puts the total factor productivity effect of artificial intelligence at no more than 0.66 per cent over ten years, refined to below 0.53 per cent.[38] If that is right, questions about which firms convert best concern the distribution of a modest prize.

The capital may not pencil. Meeting projected compute demand has been estimated to require approximately $500 billion of capital spending and some $2 trillion of new annual revenue by 2030, implying an annual shortfall of the order of $800 billion.[39] The International Monetary Fund notes that hyperscaler investment could exceed $3.4 trillion by 2030, is increasingly debt-funded, and sits alongside historically extreme equity concentration.[40]



Amazon shortened from 6.0 to 5.0 years, citing the pace of AI development. Others extended.

Reported returns may be flattered by an accounting estimate. Assumed useful lives for server and network equipment differ materially across comparable firms (six years at Microsoft, Alphabet and Oracle, five and a half at Meta) while Amazon moved in the opposite direction, shortening from six years to five effective January 2025 and citing the pace of development in artificial intelligence, at a cost of approximately $0.7 billion to 2025 operating

[38] Acemoglu, ‘The Simple Macroeconomics of AI’, 13–58.

[39] Bain & Company, *Global Technology Report* (September 2025). A scenario, not a forecast.

[40] International Monetary Fund, *Global Financial Stability Report*, April 2026, Chapter 1.

income.[41] Secondary-market evidence that accelerator hardware has traded below half its original price within three years favours the shorter assumption.

This paper accepts the first objection as a genuine constraint on the size of the prize and treats the second and third as reasons for greater governance discipline, not less. If aggregate returns are modest and capital intensity is high, dispersion between firms that convert and firms that do not becomes more consequential to shareholders, not less. The argument that follows is conditional: *given* that a firm will deploy artificial intelligence at scale, what determines whether it gets anything for it?

[41] Company disclosures. Amazon additionally recorded approximately $0.6 billion of accelerated depreciation on retirements in 2025.

PART II

02

The Mechanism

Why the person best placed to notice a shortfall does not notice it — and what that does to organisational challenge.

01 **The confidence–competence inversion**
Machine inference decouples confidence from accuracy, and reverses it beyond the frontier.

02 **Why existing answers do not reach it**
Five adjacent literatures credited in full; the contribution is what remains after all of them.

03 **Three classes of machine inference**
Predictive fails measurably, generative fails fluently, agentic fails while acting.

Part II — The Mechanism

3. The Confidence–Competence Inversion

Chapter 2 establishes *where* measured gains survive a consequential test: less often in decision tasks than in creation or learning. It does not establish *why*. The answer proposed here is that in a decision task the person best placed to notice a shortfall does not notice it — and that this failure is itself conditional on whether anything in the environment can register the error.

What the Calibration Evidence Actually Shows

Claims about artificial intelligence and overconfidence are made loosely and often. The evidential position is much narrower than the commentary suggests, and what follows depends on which studies actually measured what.

Three studies have elicited participants' beliefs about *their own measured performance* and compared those beliefs with the measurement, and all three find miscalibration in the optimistic direction. In a randomised trial, sixteen experienced open-source developers working on 246 real issues in repositories they knew well took nineteen per cent longer with artificial-intelligence assistance. They had forecast a twenty-four per cent speed-up beforehand and, having experienced the slowdown, still believed afterwards that they had been twenty per cent faster, a gap of roughly thirty-nine points that survived direct experience.[42] In a classroom trial with approximately one thousand secondary-school students, the arm given unrestricted access scored *seventeen per cent worse* on a subsequent unassisted examination and, in the authors' words, 'did not perceive that they performed worse or learned less'.[43]

The third is the largest and the most direct. In an incentivised experiment with 732 participants over 160 rounds each, beliefs about one's own ability were elicited alongside measured accuracy: 79.5 per cent of participants were overconfident, confidence exceeding accuracy by 10.6 points on average. The finding that matters here is that the participants who gained least from machine assistance were the accurate ones who were *miscalibrated*, while well-calibrated participants of low ability gained nearly ten percentage points.[44] Calibration, not ability, predicted who benefited.

That is the whole of the direct evidence: three studies, one of them very small, all finding miscalibration in the optimistic direction, two of them in settings where assistance harmed the harder measure. Two further studies are sometimes enlisted here and should not be. Participants in the coding trial estimated a thirty-five per cent productivity gain against a measured 55.8 per cent, but the estimate was made by treated and control participants alike and is a belief about the tool, not about the respondent's own output;[45] the law students 'correctly guessed the tasks for which GPT-4 was most helpful', which is

[42] Joel Becker and others, 'Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity', arXiv:2507.09089 (July 2025). Expert forecasters were wrong in the same direction, predicting speed-ups of 39 per cent (economists) and 38 per cent (machine-learning researchers). The sample is sixteen developers and time is self-reported from screen recordings.

[43] Hamsa Bastani and others, 'Generative AI Can Harm Learning', SSRN 4895486 (2024).

[44] Andrew Caplin and others, 'The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration', NBER Working Paper 33021 (October 2024).

[45] Sida Peng and others, arXiv:2302.06590.

calibration about where assistance helps, not about how well they themselves performed.[46] Both are informative and neither is a calibration study in the strict sense.

The two genuine cases do not fit together as neatly as they first appear. In the classroom trial the harm appeared on a measure taken weeks later and without assistance, which fits a straightforward account in which delay conceals the damage. In the developer trial it did not: the harmed quantity was elapsed time on the task just completed. What made it undetectable was not delay but the absence of a *counterfactual*, no participant could observe how long the same work would have taken without assistance, and no ordinary work setting supplies that comparison. The unifying condition is therefore not latency but the absence of any signal, delayed or immediate, capable of registering the specific error introduced. Delay is the commonest way that signal goes missing; a missing counterfactual is another, and it is the one that afflicts almost every claim of the form 'the tool made us faster'.

EXHIBIT 4

THE JAGGED FRONTIER

Inside the frontier of machine competence, output rises. Outside it, accuracy falls while the work reads better.

INSIDE THE FRONTIER		OUTSIDE THE FRONTIER	
+12.2%	tasks completed	−19pp	less likely to be correct
+25.1%	faster	higher	rated coherence
higher	assessed quality	invisible	from inside the task

The frontier is not visible from inside the task. Crossing it degrades the answer while improving how it reads.

This is the condition under which governance, rather than capability, determines the outcome

Older and larger literatures corroborate this. Four decades of human-factors research document automation bias and automation-induced complacency, occurring in both naive and expert participants and producing errors of both omission and commission, with training and instruction of at best limited effect.[47] One qualification matters. Accountability for decision accuracy does reduce both error types in student samples, though this has not replicated among expert operators. The bias is therefore not immune to intervention. It resists the cheap intervention and responds to a structural one.[48]

Transfer from cockpit and process-control automation to knowledge work should not be assumed, and the difference cuts against me. Classical automation had reliability that was stationary and estimable, which is what makes complacency a calibration failure against a knowable base rate; the frontier of a

[46] Choi, Monahan and Schwarcz, 'Lawyering in the Age of Artificial Intelligence', 147–208.

[47] Raja Parasuraman and Dietrich H. Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration', *Human Factors*, 52 (3) (2010), 381–410; Raja Parasuraman and Victor Riley, 'Humans and Automation: Use, Misuse, Disuse, Abuse', *Human Factors*, 39 (2) (1997), 230–253.

[48] Linda J. Skitka, Kathleen L. Mosier and Mark D. Burdick, 'Accountability and Automation Bias', *International Journal of Human-Computer Studies*, 52 (4) (2000), 701–717; and 'Does Automation Bias Decision-Making?', 51 (5) (1999), 991–1006. The reduction was obtained under accountability for accuracy or overall performance, in student samples, and did not replicate with professional pilots.

generative system is neither. The two settings are alike in the feature that drives the result (an authoritative output whose derivation the operator cannot inspect at the moment of decision) and unlike in the feature that would permit a remedy, since the operator of a stationary system can in principle learn its base rate and the user of a shifting frontier cannot. The transfer carries the problem across and leaves the cure behind.

The Mechanism, Stated Conditionally

The evidence supports a narrower claim than the one usually made.

> **The confidence–competence inversion.** Where the environment supplies a signal capable of registering the error — a fast outcome, an external score, an available counterfactual — a decision-maker's belief about the quality of a machine-assisted judgement tracks that quality, sometimes conservatively. Where no such signal is available, belief and quality come apart, and in the region beyond the system's competence they move in *opposite* directions: subjective confidence is not a noisy signal of accuracy but is negatively related to it.

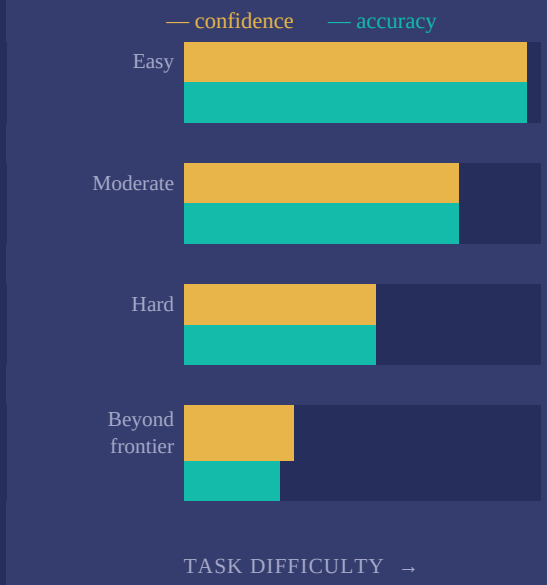
Two things follow. First, the inversion is not a general property of machine-generated inference; it is what machine-generated inference does to a decision-maker when the environment will not tell them they were wrong. Second, the mechanism and its boundary condition are the same fact. In earlier drafts I derived the boundary condition separately and presented it as a limitation on an otherwise general claim. That was a mistake. The availability of an error-registering signal is what switches the mechanism on and off, and latency is its commonest determinant but not its only one. This is why the corpus result of Chapter 2 and the calibration failures of this chapter are two views of one phenomenon: the studies in which assistance fails a consequential test are studies of decisions, and a decision is precisely the case in which the world does not reliably answer back.

EXHIBIT 5

THE CONFIDENCE–COMPETENCE INVERSION

Under human inference, failing judgement announces itself. Under machine inference, it does not.

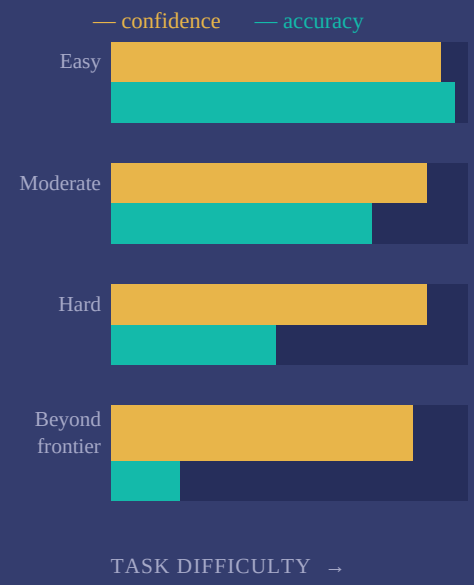
HUMAN-GENERATED INFERENCE



The two fall together — low confidence marks low accuracy

Low confidence is the signal that triggers challenge

MACHINE-GENERATED INFERENCE



Confidence holds flat while accuracy falls — the gap the organisation cannot see

No signal is emitted. Challenge is never triggered

WHERE THE SIGNAL GOES MISSING

Slow outcome

Students scored 17% worse on a delayed, unassisted test and did not perceive it

No counterfactual

Developers were 19% slower and believed they were 20% faster

Human factors

Automation bias documented in experts; resistant to training

Confidence is not a noisy signal of accuracy. In the failure region it is negatively related to it

Why This Disables Organisational Challenge

Organisations do not subject every decision to independent review; the cost would be prohibitive. They triage. Some decisions are challenged and most are not, and the triage is performed (formally through escalation thresholds, informally through the decision-maker's judgement about whether to seek a second opinion) largely on the basis of *felt uncertainty*. This is not a defect of organisational design; it is what makes challenge affordable.

Challenge mechanisms divide in two, and only one half is affected. Mandatory reviews fixed by class fire whether or not anyone feels uncertain. The discretionary margin inherits the dependency in full. Pre-mortem analysis is commissioned when a decision feels risky.[49] Escalation thresholds are drafted in terms of value and reversibility but are operated by people who decide whether a given matter has crossed them. Referral for independent review is, in the ordinary case, requested and not imposed.

If confidence is inverted, this architecture fails in a specific and diagnosable way. The organisation does not simply challenge too little; it allocates its challenge capacity in inverse proportion to need, examining the decisions where the technology performed well and waving through those where it did not. And because the proposer's confidence is genuine, the organisation experiences this as good governance. There is no felt friction, no dissent to overrule, no signal that anything has gone wrong until an outcome arrives.

> *" The organisation does not merely challenge too little. It allocates its challenge capacity in inverse proportion to need — and experiences this as good governance.*

The Theoretical Position

The claim belongs inside organisational information processing theory, and locating it precisely matters more than claiming distance from the tradition.

Galbraith defines uncertainty as the difference between the information a task requires and the information the organisation possesses, and derives the proposition that greater task uncertainty requires greater information processing between decision-makers.[50] The definition is *objective*: a property of the task and the organisation's information stock, not of anyone's mental state. The theory is not, however, silent on how gaps are detected, and it would be wrong to claim that it is. Its design sequence runs from rules and programmes to the hierarchy and then to goal setting, and the hierarchy is the detection mechanism: 'as the organization faces greater uncertainty its participants face situations for which they have no rules. At this point the hierarchy is employed on an exception basis.' Referral of exceptions upward (*hierarchical referral*, the term Galbraith uses in Figure 1 of the same paper) is self-triggered escalation, and it is the mechanism with which this paper is concerned.

The contribution is therefore not that the mechanism is absent but that it has a parameter no one has had reason to examine. Referral by exception requires that the participant *recognise* a situation as an exception. Machine-generated inference supplies fluent, complete and apparently rule-covered output

[49] Gary Klein, 'Performing a Project Premortem', *Harvard Business Review*, 85 (9) (2007), 18–19; Daniel Kahneman, Dan Lovallo and Olivier Sibony, 'Before You Make That Big Decision…', *Harvard Business Review*, 89 (6) (2011), 50–60.

[50] Jay R. Galbraith, 'Organization Design: An Information Processing View', *Interfaces*, 4 (3) (1974), 28–36, at 29. The term is Galbraith's own: the prose section is headed 'Hierarchy', while Figure 1 on the same page lists the design strategies as rules and programs, *hierarchical referral*, goal setting, slack resources, self-contained tasks, vertical information systems and lateral relations. The framework is developed in *Designing Complex Organizations* (Reading, MA: Addison-Wesley, 1973). See also Michael L. Tushman and David A. Nadler, 'Information Processing as an Integrating Concept in Organizational Design', *Academy of Management Review*, 3 (3) (1978), 613–624.

for cases that are in fact exceptions, so the perceived exception rate falls below the true one. The hierarchy is not overloaded, as the theory anticipates under rising uncertainty; it is *under-employed*, and the organisation reads the absence of referrals as evidence that its rules are holding.

> *“ Galbraith’s hierarchy is employed on an exception basis. Machine-generated inference makes exceptions stop looking like exceptions.*

The distinction between objective and perceived uncertainty is considerably older than the recent literature sometimes implies, and I claim no priority over it. Lawrence and Lorsch measured perceived environmental uncertainty in 1967 and Duncan formalised it in 1972, both before *Designing Complex Organizations*; Milliken’s much-cited taxonomy is a clarification of a construct already two decades old. Boyd, Dess and Rasheed set out the causes and consequences of divergence between archival and perceptual measures of the environment in 1993, and that paper, not anything in the artificial-intelligence literature, is the closest existing statement of the general problem.[51]

What is added is narrower than the general finding that objective and perceived uncertainty diverge. It is that a particular instrument of inference induces the divergence, in a specified direction, with two stated moderators — the fluency of the output and the inspectability of its derivation — and that the direction is the one under which referral by exception fails and does not merely become noisy. Prior work establishes that perception is an unreliable proxy for the environment. The claim here is that a technology now in general use makes it unreliable in a predictable direction, at a predictable place, for a reason that can be designed against.

What the Inversion Is Not

Three adjacent constructs are close enough to require separation, and one of them predicts the opposite of what is claimed here.

The Dunning–Kruger effect concerns the relation between a person’s skill and their assessment of it: the least competent are least able to recognise incompetence.[52] It is a claim about *who* is miscalibrated. The inversion is a claim about *when*: the same person is well calibrated inside the frontier and miscalibrated outside it, and the shift is produced by the instrument, not by their standing skill. The human-factors evidence is explicit that expertise does not confer protection.

[51] Paul R. Lawrence and Jay W. Lorsch, *Organization and Environment: Managing Differentiation and Integration* (Boston: Harvard University Graduate School of Business Administration, 1967); Robert B. Duncan, ‘Characteristics of Organizational Environments and Perceived Environmental Uncertainty’, *Administrative Science Quarterly*, 17 (3) (1972), 313–327; Frances J. Milliken, ‘Three Types of Perceived Uncertainty About the Environment’, *Academy of Management Review*, 12 (1) (1987), 133–143; Brian K. Boyd, Gregory G. Dess and Abdul M. A. Rasheed, ‘Divergence between Archival and Perceptual Measures of the Environment: Causes and Consequences’, *Academy of Management Review*, 18 (2) (1993), 204–226.

[52] Justin Kruger and David Dunning, ‘Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments’, *Journal of Personality and Social Psychology*, 77 (6) (1999), 1121–1134.

Executive overconfidence in the finance literature is a stable disposition, identified through option-exercise behaviour and press portrayal, and associated with over-investment and value-destroying acquisitions.[53] It is a trait of the person and a constant of the firm-year. The inversion is a state induced by a tool, which is why it is addressable by changing how the tool's output enters decisions, not by changing the executive. Where the two coincide the effects should compound, and that is a testable interaction.

Algorithm aversion runs in the opposite direction and has to be confronted. Dietvorst, Simmons and Massey find that people abandon an algorithm after seeing it err, even having seen it outperform the human alternative; Logg, Minson and Moore find the reverse, that lay participants weight algorithmic advice more heavily than human advice, with the qualifications that the effect weakens against the person's own judgement and reverses among domain experts.[54] These findings bound the present claim without refuting it. Aversion operates where the error is *visible*; the defining property of the jagged frontier is that it is not. Taken together the two literatures predict that the inversion should be weakest where failures are conspicuous and among experts assessing their own domain, and strongest in the large middle ground of professional work where neither condition holds. That is a sharper prediction than I would have made without them.

From Individual Miscalibration to Organisational Misallocation

The evidence assembled above is individual and the claim advanced here is organisational. That step has to be argued, not assumed, and it does not hold universally. Four conditions are jointly required. The fourth is currently the weakest of them.

First, miscalibration must be correlated across decision-makers. Idiosyncratic error averages out; organisational bias requires a common shock. There is one available: a small number of frontier models, reached through similar interfaces, sharing a frontier whose shape is a property of the model, not of the user. Where a firm's analysts use heterogeneous tools on heterogeneous tasks, the condition weakens and the prediction should weaken with it.

Second, confidence must reach the triage. The objection that committees cannot observe a proposer's mental state is correct and does not bind, because the channel is not introspective. A confident proposer produces a proposal with fewer stated caveats, fewer alternatives modelled and a narrower sensitivity range — observable features on which triage demonstrably runs. Upper-echelons theory supplies the general warrant that organisational outcomes reflect the cognitions of those who dominate the decision process.[55]

[53] Ulrike Malmendier and Geoffrey Tate, 'CEO Overconfidence and Corporate Investment', *Journal of Finance*, 60 (6) (2005), 2661–2700, and 'Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction', *Journal of Financial Economics*, 89 (1) (2008), 20–43.

[54] Berkeley J. Dietvorst, Joseph P. Simmons and Cade Massey, 'Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err', *Journal of Experimental Psychology: General*, 144 (1) (2015), 114–126; Jennifer M. Logg, Julia A. Minson and Don A. Moore, 'Algorithm Appreciation: People Prefer Algorithmic to Human Judgment', *Organizational Behavior and Human Decision Processes*, 151 (2019), 90–103.

[55] Donald C. Hambrick and Phyllis A. Mason, 'Upper Echelons: The Organization as a Reflection of Its Top Managers', *Academy of Management Review*, 9 (2) (1984), 193–206. See also Constance E. Helfat and Margaret A. Peteraf, 'Managerial Cognitive Capabilities and the Microfoundations of Dynamic Capabilities', *Strategic Management Journal*, 36 (6) (2015), 831–850, for the individual-to-organisational bridge.

Third, compensating channels must be weak. Internal audit, budget variance analysis, external audit and market feedback all sample independently of proposer confidence, and where they operate the mechanism is suppressed. They operate where outcomes are observable and attributable, which is the boundary condition of the next section arriving from a second direction, and a reason for some confidence in it.

Fourth, machine-generated inference must actually be present in the affected decision class. It is not yet, in volume. Penetration into capital allocation and acquisitions is low and rising, while penetration into fast-verifying operational work is high. The argument is therefore *prospective* for the class where it matters and *contemporaneous* for the class where it matters least. The mechanism therefore predicts a problem that is arriving, not one that can already be measured at scale. I would rather the timing ran the other way.

Where This Leaves Governance Investment

Because the mechanism is conditional in the way just stated, the governance implication can be read directly off the condition instead of argued separately.

Where a decision's outcome is *verifiable quickly and cheaply*, the organisation obtains an independent signal of accuracy that does not pass through anyone's confidence. A support agent's resolution either holds or the customer returns. A code change either passes tests or does not. Under these conditions the firm maps the jagged frontier empirically, feedback is fast, and the inversion is corrected by the environment. This is precisely the character of the domains in which the task-level evidence is most positive.

Where outcomes are *slow, noisy or confounded* (a capital allocation, an acquisition, a market entry, a restructuring), no such signal arrives within the decision cycle, and often none arrives that can be cleanly attributed at all. There the only available correction is structural. It follows that the effect of the inversion on decision quality is decreasing in outcome verifiability and increasing in interpretive ambiguity, and that governance investment should be concentrated accordingly.

OUTCOME VERIFIABILITY

Where the environment corrects the inversion — and where it does not

The gains are where verification is fast. Firm value is not.

This boundary condition also explains a puzzle in the evidence that is otherwise awkward for any optimistic account: why large, well-identified task-level gains have not aggregated into firm-level effects. The tasks where artificial intelligence has been shown to work are disproportionately those where verification is fast. The decisions that determine firm value are disproportionately those where it is not.

4. Why Existing Answers Do Not Reach It

Five established bodies of work occupy adjacent ground, and each supplies part of the answer. Several come closer to it than the artificial-intelligence literature does. The contribution of this paper should be assessed against what remains once every one of them has been credited in full.

EXHIBIT 6

WHAT THE EXISTING ANSWERS REACH

Each body of work is credited in full. The contribution is what remains after all five.

ESTABLISHED WORK	WHAT IT ESTABLISHES	WHAT IT DOES NOT REACH
Organisation theory	Uncertainty absorption; control by output measurability; the myopia of learning	An instrument that reverses the referral signal itself, and removes the author who could be asked
Model risk management	SR 11-7 (2011): effective challenge, independent validation, model inventory	Superseded April 2026, with generative and agentic systems expressly placed out of scope
Artificial intelligence capability	Calibrated, validated construct: how well a firm orchestrates its AI resources	How a firm adjudicates between sources of inference when the human party is miscalibrated
Management practice	Eighteen practices, 13,000 firms; explain over 20% of productivity variation	No practice reaches inference whose provenance is inaccessible; monitoring presupposes an author
Complementarity	Returns require organisational co-investment; explains slow, unevenly distributed gains	The quantity of investment, not its direction. Silent on why challenge is misallocated

WHAT REMAINS

An instrument that reverses the referral signal, and a classification criterion — outcome verifiability — absent from every scheme reviewed here.

Organisation Theory

The nearest prior work is not in the artificial-intelligence literature at all, and three results in particular have to be conceded before anything is claimed.

Uncertainty absorption. March and Simon observed that ‘uncertainty absorption takes place when inferences are drawn from a body of evidence and the inferences, instead of the evidence itself, are then communicated’, and that the absorber thereby acquires influence the recipient cannot audit.[56] That is a

[56] James G. March and Herbert A. Simon, *Organizations* (New York: Wiley, 1958); quotation at p. 186 of the second edition (Oxford: Blackwell, 1993). See also Richard M. Cyert and James G. March, *A Behavioral Theory of the Firm* (Englewood Cliffs, NJ: Prentice-Hall, 1963) on problemistic search, under which investigation is triggered by performance below aspiration rather than by felt uncertainty — a trigger that partially survives the inversion.

precise description of what a generative system does, sixty-eight years early, and any claim that machine inference presents a novel epistemic problem must survive it. What survives is this: in March and Simon the absorber is an *organisational participant*, locatable in the structure, subject to incentive and to interrogation. The recipient cannot audit the inference but can audit the absorber. Machine absorption does not remove the human from the chain, someone still forwards the output, but it removes that person's ability to answer for it, because they did not draw the inference and cannot inspect how it was drawn. What is lost is not the party but the party's interrogability, and it is that specific loss, developed in Chapter 5, on which the argument rests.

Control by output measurability. Ouchi derives the appropriate control mode from two variables: knowledge of the transformation process and the ability to measure outputs. Where outputs are measurable, output control suffices; where they are not but the process is understood, behaviour control is indicated; where neither holds, clan control — shared norms — is the residual.[57] The boundary condition derived in Chapter 3 is Ouchi's second variable, and I do not pretend otherwise. The addition is directional and specific: Ouchi treats output measurability as a property of the task that determines which control to install. Here it determines something further, whether the environment will *repair a miscalibrated trigger without management intervention*, which is a question that does not arise until the trigger itself becomes unreliable.

The myopia of learning. Levinthal and March establish that organisations learn well from feedback that is near in time and space and frequent, and badly otherwise; Levitt and March had already described superstitious learning, in which 'the subjective experience of learning is compelling' while the connection between action and outcome is misspecified; and Denrell shows that vicarious learning runs on a censored sample, because the organisations available to be observed are the survivors of a selection process.[58] Proposition 2 of this paper is a special case of the first result and claims no independence from it. Together the three explain something the present argument would otherwise struggle with, why a firm does not simply map its own jagged frontier by trial. Where the frontier is crossed and no outcome arrives, there is nothing to learn from; where an outcome does arrive, the experience of having learned is available whether or not the inference was sound.

Three further literatures bear directly and are credited here without being engaged at length. The attention-based view holds that scrutiny is a scarce resource whose allocation is determined by organisational channels, not by individual judgement, which is the structural premise of Chapter 7.[59] Research on high-reliability organising concerns the design of organisations that do not rely on felt uncertainty to trigger attention, and Vaughan's account of normalisation of deviance is the canonical

[57] William G. Ouchi, 'A Conceptual Framework for the Design of Organizational Control Mechanisms', *Management Science*, 25 (9) (1979), 833–848. See also Robert Simons, *Levers of Control* (Boston: Harvard Business School Press, 1995), on the distinction between diagnostic and interactive control systems.

[58] Daniel A. Levinthal and James G. March, 'The Myopia of Learning', *Strategic Management Journal*, 14 (S2) (1993), 95–112; Barbara Levitt and James G. March, 'Organizational Learning', *Annual Review of Sociology*, 14 (1988), 319–340; Jerker Denrell, 'Vicarious Learning, Undersampling of Failure, and the Myths of Management', *Organization Science*, 14 (3) (2003), 227–243. The hot-stove effect of Denrell and March (2001) is deliberately not invoked here: it concerns persistent *pessimism* about alternatives that produced early poor outcomes, which is the opposite of the pattern at issue.

[59] William Ocasio, 'Towards an Attention-Based View of the Firm', *Strategic Management Journal*, 18 (S1) (1997), 187–206; John Joseph and William Ocasio, 'Architecture, Attention, and Adaptation in the Multibusiness Firm: General Electric from 1951 to 2001', *Strategic Management Journal*, 33 (6) (2012), 633–660.

treatment of a conforming, well-documented process that misallocates scrutiny while feeling rigorous — the condition this paper describes.[60] And Shrestha, Ben-Menahem and von Krogh propose selecting among human–artificial-intelligence decision structures on ex ante decision characteristics including interpretability and the specificity of the search space, which is the closest published antecedent of the design in Chapter 7 and precedes it by seven years.[61]

What remains after all six is a single conjunction: that the instrument now supplying inference systematically *reverses the signal on which referral by exception depends*, and that the classification criterion which follows from that reversal, outcome verifiability, is absent from the schemes reviewed here for allocating challenge, including the closest one.

Model Risk Management

Banking has governed machine-generated numbers for a quarter of a century. The Comptroller of the Currency issued model-validation guidance in May 2000.[62] Federal Reserve SR 11-7 and Office of the Comptroller of the Currency Bulletin 2011-12, issued 4 April 2011, established *effective challenge*, defined as 'critical analysis by objective, informed parties that can identify model limitations and produce appropriate changes', together with validation independent of model development and use, and a maintained inventory of models in use, under development and recently retired.[63] This is, in substance, a challenge architecture, and it is the strongest existing answer to the problem this paper addresses.

Two features of its position in 2026 are decisive. First, it was superseded on 17 April 2026, and the replacement relaxes precisely the element most relevant here: validation now requires 'sufficient independence to maintain objectivity' rather than structural separation, quality is said to depend 'on the rigor and effectiveness of the review and not on organizational structure', and the model inventory is downgraded from supervisory expectation to observed industry practice. Second, and more consequentially, generative and agentic artificial intelligence are expressly placed outside its scope, on the stated basis that they are novel and rapidly evolving, with a request for information promised at an unspecified future date.[64]

The most mature institutional apparatus for governing machine-generated inference has therefore been loosened at the moment the inference became harder to inspect, and has declined to cover the classes of system now being deployed. Elsewhere the position is thinner still: the National Institute of Standards

[60] Karl E. Weick and Kathleen M. Sutcliffe, *Managing the Unexpected*, 2nd edn (San Francisco: Jossey-Bass, 2007); Diane Vaughan, *The Challenger Launch Decision* (Chicago: University of Chicago Press, 1996).

[61] Yash Raj Shrestha, Shiko M. Ben-Menahem and Georg von Krogh, 'Organizational Decision-Making Structures in the Age of Artificial Intelligence', *California Management Review*, 61 (4) (2019), 66–83. Their five selection criteria are specificity of the decision search space, interpretability, size of the alternative set, decision speed and replicability. Outcome verifiability is not among them.

[62] Office of the Comptroller of the Currency, Bulletin 2000-16, 'Risk Modeling: Model Validation', 30 May 2000, since rescinded and cited here as historical precedent only.

[63] Board of Governors of the Federal Reserve System, SR 11-7, 'Guidance on Model Risk Management', 4 April 2011; Office of the Comptroller of the Currency, Bulletin 2011-12, 'Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management', 4 April 2011. The Federal Deposit Insurance Corporation adopted the guidance subsequently, in FIL-22-2017 (7 June 2017).

[64] Federal Reserve SR 26-2, 'Revised Guidance on Model Risk Management'; OCC Bulletin 2026-13; FDIC FIL-15-2026, all 17 April 2026. Scope is stated as most relevant to banking organisations with more than $30 billion in total assets, and the guidance expressly does not create enforceable legal requirements.

and Technology announced an AI Agent Standards Initiative on 17 February 2026, currently at initiative and draft-concept stage; ISO/IEC 42001:2023 specifies an artificial-intelligence management system without addressing autonomy; and United Kingdom regulators remain deliberately technology-agnostic.[65]

Artificial Intelligence Capability

The information-systems literature has developed an organisation-level AI capability construct with completed conceptualisation, measurement calibration and firm-performance validation, specified across tangible resources, human skills and intangible resources including coordination and change capacity.[66] This is a more mature construct than the one proposed here, and I claim no priority over it.

The difference is one of object. AI capability asks how well an organisation *orchestrates AI resources* to produce outputs; it is a resource-and-competence construct in the resource-based tradition. The construct proposed here asks how an organisation *adjudicates between sources of inference* when they conflict, and in particular whether it can do so when the human party to that adjudication is systematically miscalibrated. A firm can score highly on AI capability — abundant data, skilled staff, effective coordination — and still allocate its challenge capacity in inverse proportion to need. The two constructs are complements, and the empirical question of whether the second adds explanatory power beyond the first is stated as a formal test in Chapter 10 and not assumed.

Management Practice

The most serious competing explanation is the simplest: that firms differ in management quality, that this is already measured, and that nothing further is required. The World Management Survey scores eighteen management practices through double-blind interviews across more than thirteen thousand firms in thirty-five countries, and management practices so measured explain more than twenty per cent of productivity variation — comparable to research and development, information technology or human capital.[67] United States multinationals obtain higher productivity from identical information-technology capital than their peers, and do so through people-management practice; the difference appears within establishments after acquisition by American firms and not after acquisition by others.[68]

This is a formidable body of work and the present paper does not dispute it. Its limitation for the problem at hand is that the practices it measures (monitoring, target setting, incentives) were selected and validated for an environment in which the intelligence entering decisions was produced by people whose reliability could be assessed through the ordinary means of professional judgement and track record. None of the eighteen practices addresses the arbitration of inference whose provenance is inaccessible,

[65] National Institute of Standards and Technology, ‘Announcing the AI Agent Standards Initiative for Interoperable and Secure Innovation’, 17 February 2026; ISO/IEC 42001:2023, *Information Technology — Artificial Intelligence — Management System* (December 2023).

[66] Patrick Mikalef and Manjul Gupta, ‘Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on Its Impact on Organizational Creativity and Firm Performance’, *Information & Management*, 58 (3) (2021), article 103434.

[67] Nicholas Bloom and John Van Reenen, ‘Measuring and Explaining Management Practices Across Firms and Countries’, *Quarterly Journal of Economics*, 122 (4) (2007), 1351–1408; Nicholas Bloom and others, ‘What Drives Differences in Management Practices?’, *American Economic Review*, 109 (5) (2019), 1648–1683.

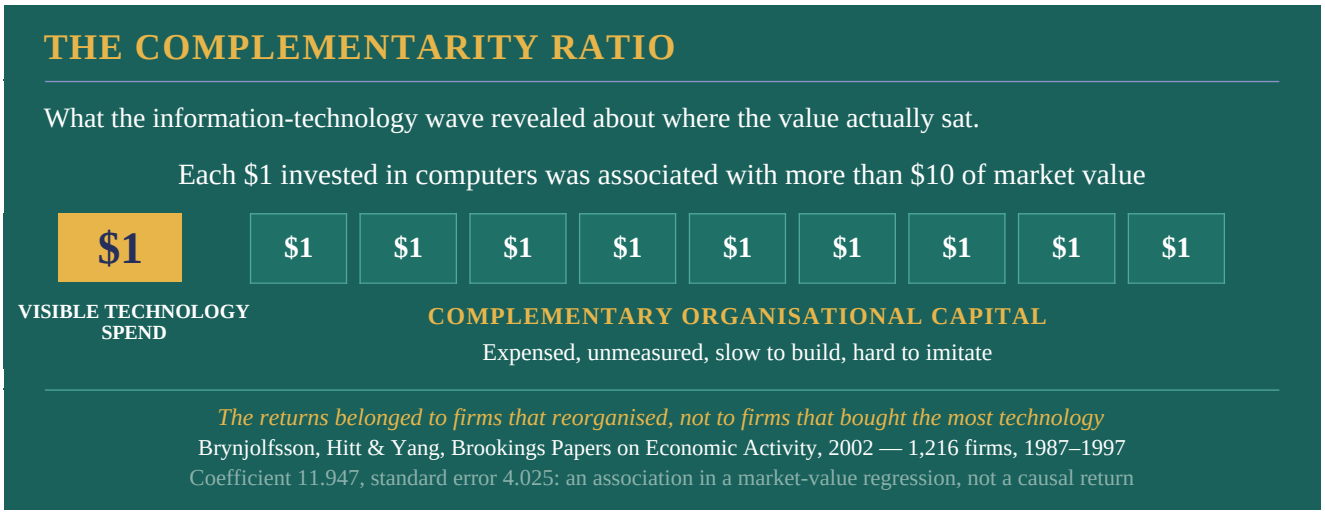
[68] Nicholas Bloom, Raffaella Sadun and John Van Reenen, ‘Americans Do IT Better: US Multinationals and the Productivity Miracle’, *American Economic Review*, 102 (1) (2012), 167–201.

and the monitoring practices in particular presuppose that what is monitored can be traced to an accountable author. Good management is necessary here and is not, on the evidence in Chapter 3, sufficient: the automation-bias literature finds the inversion in expert users, and finds that it survives instruction.

The Complementarity Result

One further body of work should be credited explicitly, because it accounts for much of the value gap without requiring the mechanism proposed here. Each dollar invested in computers in the 1987–1997 period was associated with more than ten dollars of firm market value, against just over one dollar per dollar of other tangible assets, a ratio the authors interpret as evidence of unmeasured complementary organisational capital bundled with the technology.[69] Firms adopting information technology, complementary reorganisation and new products together outperformed firms adopting any one alone.[70]

EXHIBIT 7



Complementarity explains why returns are slow and why they accrue to firms that reorganise. It does not explain why an organisation would fail to notice that a decision had gone wrong, nor why challenge capacity would be misallocated and not merely scarce. The two arguments are compatible and operate at different levels: complementarity concerns the *quantity* of organisational investment required, the inversion concerns the *direction* in which a specific and important part of it must be redesigned.

5. Three Classes of Machine Inference

Treating artificial intelligence as a single object obscures the governance problem. Three classes fail in materially different ways and require different controls, and the distinction is now regulatory as well as

[69] Erik Brynjolfsson, Lorin M. Hitt and Shinkyu Yang, 'Intangible Assets: Computers and Organizational Capital', *Brookings Papers on Economic Activity*, 2002 (1), 137–198. The coefficient on computer assets is 11.947 with a standard error of 4.025; the confidence interval is wide, the period specific, and the estimate an association in a market-value regression rather than a causal return.

[70] Timothy F. Bresnahan, Erik Brynjolfsson and Lorin M. Hitt, 'Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence', *Quarterly Journal of Economics*, 117 (1) (2002), 339–376.

analytical: the revision of model risk guidance draws its scope boundary between the first class and the other two.

EXHIBIT 8

THREE CLASSES OF MACHINE INFERENCE

The failure mode differs by class. So must the control — and so does the regulatory perimeter.

	PREDICTIVE *fails measurably*	GENERATIVE *fails fluently*	AGENTIC *fails while acting*
How it fails	Measurably. Back-tested against realised outcomes	Fluently. No interval, no back-test, no error rate	While acting. The outcome precedes any review
Accountable author	The model owner, named and documented	None. No one to ask why the claim is believed	None, and the action is already taken
Regulatory scope after 17 April 2026	Within scope of revised model risk guidance	Expressly excluded	Excluded; no supervisory standard exists
Strength of the inversion	Weak. The environment supplies the correction	Strong. Fluency is read as reliability	Strong, and uncorrectable after the fact
Control that works	Validation, monitoring, back-testing	Provenance, mandatory challenge, named owner	Ex ante allocation of decision rights

A single artificial-intelligence policy is a category error

Predictive systems (credit scoring, demand forecasting, conventional machine learning) fail *measurably*. Their outputs can be back-tested against realised outcomes, their error distributions estimated, and their degradation detected through monitoring. The model risk apparatus was designed for this class and remains within its scope. Here the inversion is weak, because the environment supplies the correcting signal.

Generative systems fail *fluently*. An incorrect output is syntactically and rhetorically indistinguishable from a correct one; there is no confidence interval, no back-test, and — the point most often missed — no accountable author. When a colleague advances a claim, the organisation can ask why they believe it and hold them to the answer. That is the substrate on which challenge operates, and generative output does not provide it. This class is expressly outside the scope of current model risk guidance, and it is where the inversion is strongest.

The obvious objection is that human-in-the-loop design restores the missing party: a named reviewer signs the output, and accountability is re-established. The objection is serious and it is half right. What human-in-the-loop supplies is *responsibility*, someone who can be held to the consequence. What it does not supply is *interrogability* — someone who can answer the question on which challenge actually operates, which is not ‘do you accept this?’ but ‘why does the reasoning hold?’ A reviewer who did not produce the inference and cannot inspect its derivation can attest to plausibility and to consistency with

what they already believe. That is a weaker instrument than it appears, and it is weakest precisely where the output is most fluent. The distinction between nominal and substantive accountability is not new to governance, and the empirical work on professionals using opaque diagnostic systems finds exactly this pattern: engagement with the tool's output substitutes for engagement with its basis.[71]

Agentic systems fail *while acting*. Where a generative system proposes and a human disposes, an agentic system executes, and the error is committed before any review could have intercepted it. For this class, ex post challenge is not merely weak but unavailable, and the only effective control is the ex ante allocation of decision rights: which categories of action the system may take unsupervised, which require confirmation, and which are prohibited. Survey evidence from United Kingdom financial services in late 2024 found fifty-five per cent of AI use cases involving some automated decision-making but only two per cent fully autonomous;[72] the governance question is what happens to that second figure, and no supervisory guidance currently addresses it.

One development would dissolve much of this. If generative systems acquire reliable calibration (a defensible confidence estimate attached to each output, and self-critique that identifies its own out-of-competence cases) then fluency ceases to be uninformative, the second class begins to behave like the first, and the governance apparatus proposed here becomes unnecessary overhead in most of its range. That is a live research direction and not a settled capability. If it arrives it falsifies the prescription here more cleanly than any empirical test could. The claim advanced is conditional on a property of the current technology, and conditional claims should name the condition.

The practical implication is that a single 'AI policy' is a category error. The controls that suit a predictive model (validation, monitoring, back-testing) do not reach generative output, and the controls that reach generative output — provenance discipline, mandatory challenge, named human ownership — are unavailable against an agent that has already acted.

[71] Sarah Lebovitz, Hila Lifshitz-Assaf and Natalia Levina, 'To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis', *Organization Science*, 33 (1) (2022), 126–148; and Sarah Lebovitz, Natalia Levina and Hila Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What', *MIS Quarterly*, 45 (3) (2021), 1501–1526.

[72] Bank of England and Financial Conduct Authority, *Artificial Intelligence in UK Financial Services 2024* (November 2024).

PART III

03

Governance

What follows for the design of challenge when the decision-maker's own uncertainty can no longer be trusted as a trigger.

01 **Four architectures**

Decision rights, challenge, information, accountability — fixed by the lifecycle of an inference, not asserted.

02 **Why the governance is under-supplied**

It constrains executive discretion; executives will not volunteer it.

03 **Class-triggered challenge**

Mandatory review by decision category — and how it differs from the authority matrix you already have.

6. Enterprise Intelligence Governance

The construct proposed here is deliberately narrow. An earlier formulation of this argument carried a broader organisational capability alongside it; that construct duplicated work already done better elsewhere and has been removed. What remains is the part that is not already occupied.

> **Enterprise Intelligence Governance** is the set of board- and executive-level structures, decision rights and accountability mechanisms through which an organisation determines which intelligence carries authority, which decisions receive independent challenge, and who owns the resulting judgement.

Four Architectures, and Why Four

A construct whose dimensions are asserted invites the objection that a different set would fit the same evidence equally well. The dimensions here are fixed by the lifecycle of a single inference entering a single decision, which admits exactly four questions.

Decision-rights architecture answers *who may act on this, and which systems may act unsupervised*. It is the only control available against the agentic class, for the reason given in Chapter 5.

Challenge architecture answers *which decisions receive independent review, and on what trigger*. This is where the inversion bites, and it carries the weight of the argument.

Information architecture answers *how the inference is captured, validated, attributed and routed*, the provenance question, and the one most organisations mistake for the whole of the problem.

Accountability and feedback architecture answers *how the outcome is evaluated and returned to practice*, which is what converts a single decision into a map of where the technology works.

Two claims are made about the four and they should not be run together. They are *non-redundant*: dropping any one leaves a stage of the lifecycle ungoverned. They are also offered as *covering* it, on the ground that an inference entering a decision must be sourced, permitted, tested and answered for, and that any further requirement is a refinement of one of those and not a fifth. The first claim is demonstrable; the second is an argument, and a reader may reject it. They are analytically distinct in that each answers a question the others do not. They are *not* operationally separable, and the distinction matters enough to state. A single control instrument will ordinarily touch several dimensions at once: the

design set out in Chapter 7 requires a documented classification, a named independent challenger, a provenance record and a termination record, which between them draw on all four. That is the normal condition of a multidimensional construct (the dimensions partition the *questions*, not the *instruments*) and a reader who expects one instrument per dimension will find the scheme incoherent for the wrong reason.

The ordering should also be stated carefully, since it is not the order in which the dimensions are listed and it is not the same for every class. For predictive and generative systems, provenance is established before an inference can be authorised or tested; authorisation and challenge then operate on it; feedback closes the loop. For agentic systems the order inverts at the front, because decision rights must be settled before an action that cannot be intercepted — which is Chapter 5's argument restated as a sequencing constraint. The list is arranged by governance salience, not chronology, and only one of the four, challenge architecture, is derived from the confidence–competence inversion itself. The other three are general conditions for governing inference of any provenance, and I claim no novelty for them.

What Are Not Dimensions

Two conditions matter greatly and are not components of the construct, because treating them as components would be a category error. *Organisational learning* and *trust* are preconditions: without the first the firm cannot locate its own jagged frontier, and without the second a mandated review becomes a formality because contradicting an AI-supported recommendation carries professional cost. A precondition that is also a dimension would allow the construct to score highly with the precondition absent, which is the opposite of what the word means. They enter Chapter 10 as conditions on the design, not as components of the construct.

Decision quality is likewise excluded. It is the outcome through which governance reaches economic value, and including it as both component and outcome would render the central proposition definitional instead of testable. For the same reason the definition above is stated without reference to decision quality, an earlier draft defined the construct as operating 'in ways that improve the quality of consequential decisions', which placed the outcome inside the definiens and quietly committed the error it claimed to avoid.

Relation to the Artificial Intelligence Capability Construct

Mikalef and Gupta specify an organisation-level artificial-intelligence capability across tangible resources, human skills and intangible resources including coordination and change capacity, with completed measurement calibration and firm-performance validation.[73] Four of the five domains carried by the earlier version of the present construct were near-restatements of theirs, which is a sufficient reason to have removed them. Governance was the exception, and it is the dimension their instrument does not contain.

Enterprise Intelligence Governance is therefore offered as a complement to a validated construct rather than a competitor to it: the governance dimension that the measurement of artificial-intelligence

[73] Patrick Mikalef and Manjul Gupta, 'Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on Its Impact on Organizational Creativity and Firm Performance', *Information & Management*, 58 (3) (2021), article 103434.

capability presently omits. That is a smaller claim than the one I first made, and a considerably more defensible one. Proposition 4 states it as a test.

A word on where this sits in the resource-based tradition, since the question will be asked. Governance arrangements are not resources and I do not claim they are. They are candidates for competitive significance on the narrower ground of causal ambiguity and social complexity, a system of interacting practices is harder to copy than an asset is to buy. That claim sits uneasily beside the practical proposals in Chapters 7 and 9, which are deliberately implementable in a single planning cycle, and the tension is real: a design that can be published as a four-item list is by construction imitable. The resolution is that the *list* is imitable and the *willingness to be constrained by it* is not, which is the subject of the next section and is an argument about incentives, not about resources.

Why the Governance Is Under-Supplied

If this governance raises the return on artificial-intelligence investment, and such investment is being made at the scale set out in Chapter 1, its absence requires explanation. Three are available and they are not mutually exclusive.

The first is *cost and latency*. Challenge consumes managerial attention and slows decisions. There is an interior optimum, not a corner solution.

The second is *causal ambiguity*. Because the arrangement is a system of interacting practices and not a discrete asset, it is difficult to observe, price or imitate — which is what would make it a source of persistent advantage, and equally what makes it hard to build deliberately.

The third is *agency*, and it is the least comfortable. Enterprise Intelligence Governance is, in substance, a voluntary constraint on executive discretion: challenge architecture reduces the ability of a chief executive to carry a decision on conviction, explicit decision rights reduce informal authority, and accountability architecture makes past reasoning auditable. The distinction between formal authority and the real authority that accrues to whoever controls the information is the relevant frame.[74] A chief executive maximising private benefits of control will rationally under-supply it, which implies that its cross-sectional distribution should correlate with entrenchment, ownership structure and board independence, a prediction that is testable and, so far as I know, untested.

> *“ Enterprise Intelligence Governance is a voluntary constraint on executive discretion. That is why it is valuable, and why it will not be supplied by executives acting alone.*

[74] Philippe Aghion and Jean Tirole, ‘Formal and Real Authority in Organizations’, *Journal of Political Economy*, 105 (1) (1997), 1–29. See also Patrick Bolton and Mathias Dewatripont, ‘The Firm as a Communication Network’, *Quarterly Journal of Economics*, 109 (4) (1994), 809–839; John R. Graham, Campbell R. Harvey and Manju Puri, ‘Capital Allocation and Delegation of Decision-Making Authority within Firms’, *Journal of Financial Economics*, 115 (3) (2015), 449–470.

7. Class-Triggered Challenge

The prescription follows directly from the mechanism. If challenge cannot be triggered by the decision-maker's confidence, it must be triggered by something else, determined in advance by someone other than the proposer.

EXHIBIT 9

FROM CONFIDENCE-TRIGGERED TO CLASS-TRIGGERED CHALLENGE

If the trigger cannot be the proposer's confidence, it must be set in advance by someone else.

CONFIDENCE-TRIGGERED *the prevailing arrangement*	CLASS-TRIGGERED *the design proposed*
01 **Decision is prepared** by the proposing function	01 **Board classifies decisions ex ante** before any specific case arises
02 **Proposer judges uncertainty** the trigger is subjective	02 **Three criteria, not confidence** consequence, reversibility, outcome verifiability
03 **Review where confidence is low** and none where it is high	03 **Class determines review** invariant to anyone's conviction
04 **Under the inversion** review is allocated in inverse proportion to need	04 **Challenge is concentrated** where no natural correction will arrive

THE THREE CRITERIA, SET BY THE BOARD BEFORE ANY CASE ARISES

CONSEQUENCE	REVERSIBILITY	OUTCOME VERIFIABILITY
magnitude of value at risk	can it be unwound at acceptable cost	will an independent signal arrive in time — the addition here

Mandatory challenge falls on capital allocation, acquisitions, market entry, restructuring and material accounting judgements — a small share of decisions and almost the whole of firm value.

Not universal caution: the concentration of a scarce resource.

The Design

Under confidence-triggered challenge, the prevailing arrangement, a decision receives independent review when the person advancing it, or their supervisor, judges the matter uncertain enough to warrant it. Under class-triggered challenge, the board determines ex ante which categories of decision receive mandatory review, and that determination is invariant to how confident anyone is.

Three criteria should govern the classification, and each is derived from the argument above, not from practice. *Consequence*: the magnitude of value at risk. *Reversibility*: whether the decision can be unwound at acceptable cost. *Outcome verifiability*: whether the environment will supply an independent signal of accuracy within a useful horizon. The third is my addition, and it follows from the boundary condition derived in Chapter 3 — decisions whose outcomes are slow, noisy or confounded receive no natural correction, and therefore require the most structural one.

The design implication is that mandatory challenge should be concentrated where consequence is high, reversibility is low and verifiability is poor, which describes capital allocation, acquisitions, market entry, restructuring and material accounting judgements, and does not describe the great majority of operational decisions in which artificial intelligence is currently deployed. This is not a counsel of universal caution. It is an argument for concentrating a scarce resource where the ordinary corrective mechanism is absent.

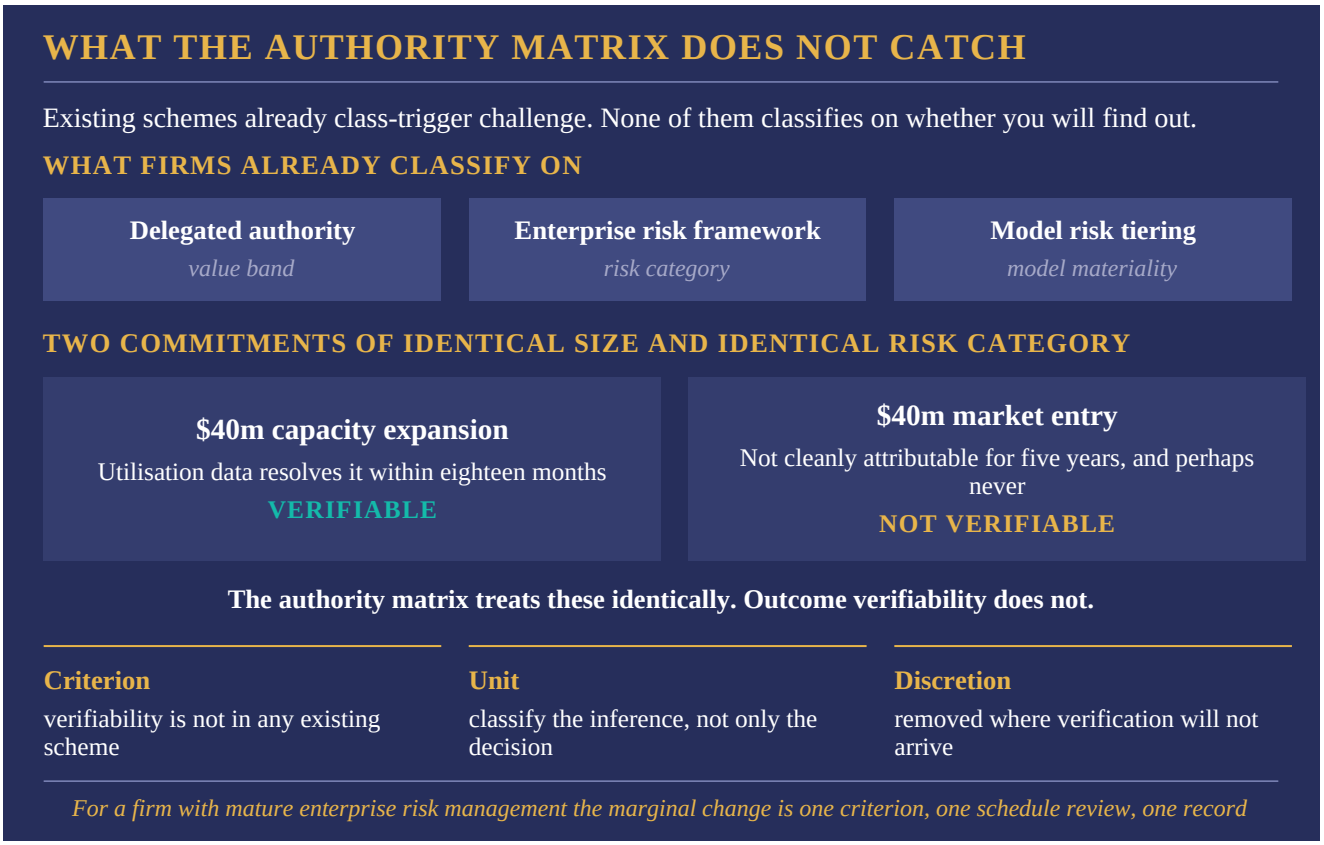
How This Differs from What Firms Already Do

The most serious objection to the design is that it already exists. Large organisations do not in fact leave challenge to felt uncertainty: delegated authority matrices mandate approval by value band, reserved-matters schedules bind the board, audit committees review critical accounting estimates, credit committees enforce limits, enterprise risk frameworks classify by risk category, and model risk tiering assigns validation intensity by model materiality.[75] Every one of these is class-triggered and confidence-invariant in principle. A prescription that amounts to ‘adopt a delegation of authority matrix’ would be worthless.

Three things are different, and the argument stands or falls on them.

[75] Committee of Sponsoring Organizations of the Treadway Commission, *Enterprise Risk Management — Integrating with Strategy and Performance* (2017), superseding *Enterprise Risk Management — Integrated Framework* (2004); and the three-lines model as applied by the Institute of Internal Auditors.

EXHIBIT 10



The classification criterion. Existing schemes classify on value, on risk category, or on model materiality. None classifies on whether the environment will supply an independent signal of accuracy. Two commitments of identical size and identical risk category can differ completely in this respect: a forty-million capacity expansion reveals itself through utilisation within eighteen months, while a forty-million market entry will not be cleanly attributable for five years and may never be. A delegated authority matrix treats them identically. Under the argument of Chapter 3 they should not be treated identically, because the first will be corrected by the environment and the second will not.

The unit of classification, which is the more important of the three. Every existing scheme classifies *decisions*, and classifies them by their own magnitude. The inversion does not operate on decisions; it operates on *inputs*. These come apart in a way that no threshold can detect. A decision far below every authority limit may rest entirely on machine-generated inference, and a population of such decisions — pricing, provisioning, credit adjudication, reserving, claims handling — can carry more consequence in aggregate than any single decision the matrix is designed to catch. A bank whose reserved-matters schedule is impeccable and whose provisioning model is generatively assisted at the level of individual case narratives has a governed decision layer sitting on an ungoverned inference layer.

This is the sharpest departure from existing practice and it deserves to be stated as a design rule and not an observation. Classify the inference, not only the decision. The unit of governance should be the *category of judgement into which machine-generated inference enters*, assessed on the consequence of that category in aggregate, not on the size of any one decision within it. Two firms with identical

authority matrices can differ completely on this measure, and nothing in the matrix will reveal it. It also generates an immediate diagnostic that no existing framework asks for: which categories of judgement in this organisation now take machine-generated inference as an input, what is their aggregate annual consequence, and which of them receives no review that is independent of the proposer?

Where the discretion sits. Existing thresholds are confidence-invariant on paper and operated by people who determine whether a given matter has crossed them, whether a change is material, whether an estimate is critical, whether a model is significant. That operating discretion is precisely where the inversion enters, and it is the reason a nominally class-triggered system can behave as a confidence-triggered one in practice. The proposal is not to invent thresholds. It is to remove the discretionary step for the classes where verification will not arrive in time to expose the error.

One qualification is owed to the auditing standards, which come closer than any of the above. ISA 540 (Revised) requires the auditor to assess inherent risk by reference to the degree of estimation uncertainty in an accounting estimate and the degree to which it is affected by complexity and subjectivity, and to scale procedures accordingly.[76] That is a genuine pre-existing scheme which graduates scrutiny by something adjacent to the criterion proposed here. The distinction is real and not a rescue: the standard's allocation variable is *ex ante measurement difficulty*, how hard the estimate is to make, whereas the variable proposed here is *ex post observability* — whether the world will later reveal whether the estimate was right. The two correlate and are not the same, and the auditing standard reaches only accounting estimates, not the capital allocations and market entries where the argument bites hardest.

Two objections to the criterion should be met directly. The first is *collinearity*: verifiability and reversibility are correlated in practice, since decisions that resolve quickly are often decisions that can be unwound. If the correlation were near unity the third criterion would add nothing. It is not: a plant commitment is verifiable and irreversible, an easily reversed pricing experiment in a thin market may be neither cleanly attributable nor consequential, and the incremental content of the criterion is an empirical question that Proposition 3 is written to answer. The second is that three criteria without a combination rule relocate discretion instead of removing it, which is correct as an objection and is answered by specifying the rule: the classification should be *disjunctive at the top*, a category enters mandatory review if it is high on consequence *and* low on either reversibility or verifiability, because the argument is about avoiding uncorrected error, and either route leaves error uncorrected.

A third objection is more serious and applies to any published classification. Class triggers are thresholds on observables, and thresholds are gamed. The canonical response to an authority limit is to split the commitment; the canonical response to a mandatory-review category is to characterise the proposal as belonging to an adjacent one. Confidence triggers, whatever else is wrong with them, are not threshold-gameable in this way. Three responses are available and none is complete: classification by category of judgement, not by transaction makes splitting harder than it is under a value band; aggregation rules of the kind already used for related-party and connected-transaction thresholds transfer directly; and the termination record makes a pattern of near-threshold characterisation visible after the fact. The honest statement is that the design trades one manipulable trigger for another, and claims only that the second is manipulable in ways that leave evidence.

[76] International Auditing and Assurance Standards Board, ISA 540 (Revised), *Auditing Accounting Estimates and Related Disclosures*, paragraphs 4, 8, 16(a)–(b) and 18.

For a firm with a mature enterprise risk framework the marginal change is small: one criterion added to an existing classification, one review of the schedule against it, one provenance requirement, one record. That is an argument for adoption. The machinery is not new. The criterion is.

The Economics of the Trigger

The case for class-triggering rests on assumptions that can each fail.

Let a firm face a population of decisions, of which some subset would be improved by independent review; call that the *need*. Let challenge capacity be fixed and binding, review cost be constant per decision, and — the assumption most often left unstated — review efficacy be independent of the trigger used to select the decision. Under those conditions, and where the joint distribution is well behaved, the value recovered by a triage rule increases in the association between the rule's trigger and need. The precise statement is an order-statistic result about the expected sum of need over the selected set, not a claim about a correlation coefficient; the correlation is a serviceable approximation and nothing here depends on more than the ordering.

Under human-generated intelligence the association is positive for the confidence trigger: felt uncertainty tracks genuine difficulty imperfectly but with the right sign. Confidence-triggering is then not merely defensible but efficient: it is free, it updates continuously, and it uses private information no central rule possesses. This is why organisations rely on it, and they have been right to.

Where the inversion operates the association is zero or negative. A rule whose trigger is uncorrelated with need allocates review at random with respect to need; a rule whose trigger is negatively associated with need concentrates review on the decisions that need it least. Any classification with a strictly positive association then *weakly dominates in expectation*, and strictly dominates wherever the confidence association is negative. The claim is deliberately modest: a classification with zero association merely ties with random allocation, and the argument for consequence, reversibility and verifiability is not that they are optimal but that they are observable ex ante and plausibly positively associated with need.

Two things follow for whoever has to pay for this. The first is that the proposal is a reallocation and not an increase. Nothing here argues for more review. It argues that a fixed review budget currently spent on the decisions whose proposers feel least certain should be spent on the decisions whose outcomes will never report back. The second is that review is expensive in the currency boards actually mind, which is cycle time. A mandatory challenge step on a capital submission costs days to weeks of decision latency and the attention of people whose attention is the binding constraint. That cost is why challenge is rationed, and rationing is not the problem. Rationing on the wrong variable is.

Three conditions bound this, and each corresponds to a way the design could fail in practice, not in principle.

Mixed populations. The inversion applies to machine-supported judgements. Replacing the confidence trigger across the whole population would discard the positive signal still available on the human-generated remainder. The design is therefore additive and not substitutive (class triggering for the affected categories, confidence triggering retained everywhere else) and the size of the advantage is

increasing in the machine-supported share, which is what Proposition 3 states and what a test would have to estimate.

Review efficacy. The model assumes that review, once allocated, works. Chapter 5 gives a serious reason to doubt this in exactly the class the rule selects, and the next section answers it. If that answer fails, the trigger argument fails with it, because a better-aimed instrument of zero efficacy recovers nothing.

Reviewer contamination. Where challengers themselves rely on machine-generated inference, the association between the trigger and *realised improvement* falls even if the association with need is unchanged. This is not hypothetical and it argues for provenance discipline applied to the review as well as to the proposal — the challenger's own basis being as inspectable as the one under challenge.

One corollary follows. This is not an argument that structural triggers are generally superior to judgement; that argument would be wrong, and the efficiency of the confidence trigger under human inference is the reason. Class-triggering dominates *only under the inversion*, and the inversion operates only where verification is slow. The design claim is therefore conditional on the mechanism claim, which is conditional in turn on a measurable property of the decision. Falsifying either falsifies both.

Whether Review Can Work at All

Chapter 5 argued that a human placed in the loop supplies responsibility without interrogability: a reviewer who did not draw the inference and cannot inspect its derivation is able to attest to plausibility and to consistency with what they already believe, and is weakest where the output is most fluent. That argument, taken alone, would defeat the present chapter. Mandating review by a person who cannot evaluate the thing under review buys a form and not a control, and the objection has to be met and not absorbed.

The answer is that the control is the articulation, not the verification. What mandatory, structurally independent, provenance-supported review requires is not that the challenger validate the derivation (often they cannot) but that the proposer produce, on the record and before the decision, four things a fluent output does not itself supply: what the inference is, what it rests on, what would have to be true for it to be wrong, and who is accountable if it is. Each of these is a demand on the human who is advancing the case, and each is answerable without inspecting a model.

Two objections have to be met before that answer is worth anything. The first is that an articulation produced by a party who cannot inspect the derivation may bear no truth-relation to it, plausibility-attestation written down. The answer is that the four demands are not about the derivation. What the inference *is*, what it is being relied upon *for*, what would falsify it, and who owns it are facts about the proposer's reliance, not about the model, and a proposer who cannot state them has revealed something material whether or not the model is inspectable. The second objection is sharper: the marginal cost of producing a fluent four-part articulation is also approaching zero, and the same system can generate it. That is genuine, and the only defence is one this paper already requires elsewhere, the falsification condition must be stated in a form that can be scored against a later outcome, and recorded before the decision. An articulation that cannot be marked wrong is not a control, and a termination record is what makes marking possible.

The supporting evidence should also be quoted with its qualification attached. Accountability for decision accuracy reduced both omission and commission errors in student samples; experimentally induced accountability did *not* produce significant differences among professional pilots, among whom it was an internalised sense of accountability that predicted verification behaviour. Since the population at issue here is professional, that is a weaker foundation than the student result alone would suggest, and it points toward the same conclusion by a different route: what works is not being told one is accountable but being placed in an arrangement that requires one to act as though one were.

This is why the design is not simply 'more review'. It is a requirement that displaces the substrate on which challenge operates from the machine's reasoning, which is unavailable, to the proposer's reasoning, which is not. It also explains the otherwise puzzling human-factors result reported in Chapter 3: accountability for accuracy reduces automation bias where training does not, because accountability compels articulation while instruction merely warns. And it predicts the failure mode precisely — where review is a sign-off and not a demand for articulation, it should have no effect at all, which is a testable difference between two arrangements that look identical on an organisation chart.

What Class-Triggered Challenge Requires

Four elements are necessary and none requires new technology. A *documented classification* of decision categories, owned by the board, not by management, with the criteria stated. A *named challenger* for each category, structurally independent of the proposer, an arrangement the model risk literature established, and which the April 2026 revision has relaxed at precisely the wrong moment. A *provenance requirement*: for any material machine-generated input, a record of what system produced it, on what data, and which named person is accountable for its use. And a *termination record*: a maintained account of AI deployments that were stopped and why, without which the organisation cannot locate its own frontier.

The last of these is the cheapest and the most frequently absent. An organisation that has never stopped an artificial-intelligence initiative has not been learning; it has been accumulating.

Governance Under Regulatory Uncertainty

Boards should plan against dates, not sentiment. Under the European Union Artificial Intelligence Act, prohibitions and literacy obligations have applied since February 2025 and general-purpose model obligations since August 2025. Article 50 transparency obligations — disclosure of AI interaction, machine-readable marking of synthetic content, notification of emotion recognition and biometric categorisation — apply from 2 August 2026, as does the Commission's power to fine providers of general-purpose models. High-risk obligations under Annex III have been deferred to December 2027 and Annex I obligations to August 2028. Penalties reach €35 million or seven per cent of worldwide annual turnover for prohibited practices, and €15 million or three per cent for transparency breaches.[77]

[77] Regulation (EU) 2024/1689, Article 113, as amended by the Digital Omnibus agreed politically on 6 May 2026. The deferrals of the Annex III and Annex I deadlines require publication in the Official Journal to take legal effect; boards should verify the position at the time of reading.

Enforcement against overstated artificial-intelligence claims is established and not prospective: the United States Securities and Exchange Commission brought its first such actions in March 2024, and the Federal Trade Commission has since pursued a series of cases, a majority of the most recent involving business-to-business claims.[78] The exposure runs in both directions — firms are at risk from claiming more artificial intelligence than they use, as well as from governing less than they should.

EU AI ACT — WHAT APPLIES

Board planning dates

Feb 2025
Prohibitions and AI literacy

Aug 2025
General-purpose model obligations

2 Aug 2026
Article 50 transparency; Commission fining power

Dec 2027
Annex III high-risk (deferred)

Aug 2028
Annex I high-risk (deferred)

Penalties reach 7% of worldwide turnover for prohibited practices.

Board-level oversight has not kept pace. Approximately twenty-eight per cent of firms report chief-executive oversight of artificial-intelligence governance and seventeen per cent report board-level oversight.[79] The Bank of England and Financial Conduct Authority found that while eighty-four per cent of United Kingdom financial firms had a named accountable person for artificial intelligence, forty-six per cent reported only partial understanding of the systems they use, particularly third-party systems. Accountability without comprehension is a governance form, not a governance function.

[78] Securities and Exchange Commission, Press Release 2024-36, 18 March 2024 (penalties of $225,000 and $175,000); Federal Trade Commission action reported May 2026.

[79] McKinsey, *The State of AI* (March 2025).

PART IV

04

The Finance Function and Capital Allocation

When a confident quantitative case costs almost nothing to produce, the scarce function is no longer analysis but adjudication between analyses.

01 **Arbiter, not producer**

Finance's comparative advantage is the refusal to certify a number it cannot trace.

02 **Who challenges the challenger**

Classification owned by the board, prepared by finance, executed by the second line.

03 **Four changes to appraisal**

Appraise the complement, prove the benefit, extend post-investment review, state the replacement cycle.

Part IV — The Finance Function and Capital Allocation

8. The Finance Function as Arbiter of Authority

The argument to this point is general to the organisation. It has a particular consequence for the finance function, and that consequence is not the one usually described.

What Scarcity Shifts

The prevailing account of artificial intelligence in finance is one of automation: routine processing is displaced, capacity is released, and the function moves toward analysis and business partnering. The early field evidence is consistent with this. A study combining a 277-accountant panel with transaction-level field data from seventy-nine firms reports an 8.5 per cent reallocation of time from data entry toward higher-value work, a twelve per cent increase in general-ledger granularity and a reduction of seven and a half days in the monthly close.[80]

That account is correct and incomplete. It describes what happens to finance's *production* function and says nothing about its *institutional* one.

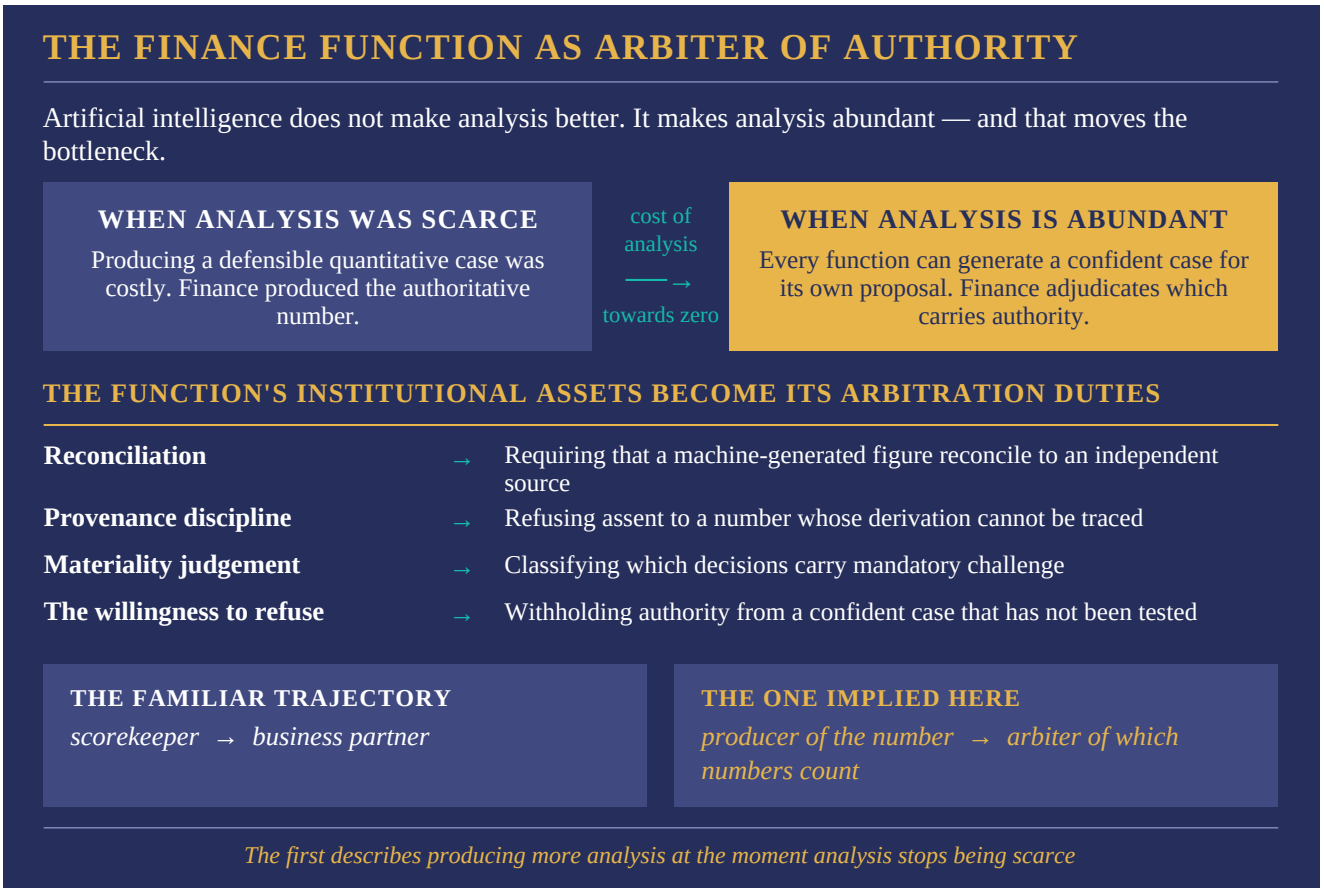
The finance function's distinctive role has never been principally to produce numbers. It has been to withhold assent from numbers until they reconcile, to act as the organisation's designated sceptic, with a professional obligation not to certify a figure it cannot trace. Reconciliation, provenance discipline, materiality judgement and the willingness to refuse are the function's actual comparative advantage, and they are institutionalised in a way that no other function's scepticism is.

Artificial intelligence changes what is scarce. When the marginal cost of producing a fluent, quantitative, internally consistent case for any proposal falls toward zero, analysis ceases to be the bottleneck. Every function can now generate a persuasive model for its own preferred course of action, at speed, with the surface characteristics of rigour. The scarce organisational function becomes adjudication between competing analyses, determining which of several plausible, confident, well-presented cases carries authority.

> *“ When every function can produce a confident quantitative case for its own proposal at near-zero cost, the scarce capability is no longer analysis. It is adjudication between analyses.*

[80] Jung Ho Choi and Chloe Xie, ‘Human + AI in Accounting: Early Evidence from the Field’, *Journal of Accounting Research* (2026), DOI 10.1111/1475-679X.70052. Early View at the time of writing; no volume or pagination assigned.

EXHIBIT 11



This reframes the chief financial officer's trajectory. The familiar formulation, from scorekeeper to business partner, describes a move toward producing more and better analysis, at exactly the moment when analysis is becoming abundant. The trajectory implied by this paper is different and, on the argument above, more defensible: from producer of the authoritative number to arbiter of which numbers carry authority. That is a governance role, not an analytical one, and it is the reason the challenge architecture described in Chapter 7 is best anchored in finance. Anchored, not owned outright: the two sections below separate what finance prepares from what the board adopts and what the second and third lines execute. Finance already owns provenance discipline, already has a mandate to refuse, and already possesses the only organisational process designed to compare a forecast with what actually happened.

Who Challenges the Challenger

Placing the challenge architecture in finance creates an immediate problem that must be answered here. Chapter 7 requires a challenger structurally independent of the proposer. For capital allocation, the decision class where the argument says challenge matters most, *finance is the proposer*. It builds the model, sets the discount rate and writes the paper. It cannot also be the independent reviewer of its own submission.

The resolution is to separate three roles that the phrase ‘challenge architecture’ conflates. *Classification* (determining which categories of decision receive mandatory review) is *owned* by the board, which adopts the schedule and is answerable for it, and *prepared and maintained* by finance, which already performs materiality judgements of this kind under internal control over financial reporting. The distinction is the ordinary one between a reserved matter and the function that drafts it, and it matters here because a classification prepared by the function whose proposals it governs must be ratified somewhere else. *Arbitration* — determining which of several competing analyses carries authority — is the adjudication function described above. *Execution*, performing the review itself, must sit outside whichever function proposed the decision. Where finance is the proposer, execution belongs in the second or third line: risk, internal audit, or a standing review panel reporting to the audit committee. This is the arrangement model risk management already uses, in which validation is independent of both development *and use*, and the present proposal claims no originality for it.

A second limit should be conceded in the same spirit. The provenance requirement is a technical control and finance functions do not generally possess the capability to verify data lineage or model derivation unaided. Ownership of the requirement belongs with finance because the consequence is financial; execution of it is a joint obligation with the data and technology function, and a chief financial officer who asserts the requirement without securing that partnership will obtain a form and not a control — which is the failure mode Chapter 7 has already identified in the accountability statistics.

Where the Function Changes, Concretely

Four areas warrant specific attention, distinguished by how the argument of Chapters 3 and 5 applies to each.

Financial reporting and the close. Here outcomes are highly verifiable and the correcting mechanism is strong: a misstatement is discovered, reconciliation either balances or does not, and audit provides independent challenge by construction. This is the domain in which the inversion is weakest and automation gains are most readily banked, consistent with the field evidence above.

Planning, forecasting and analysis. Verifiability is moderate and delayed. Forecast error is observable but attribution is contested, and the same tools that generate a forecast can generate a persuasive explanation of why it was missed. The governance requirement here is unglamorous: pre-registration of the forecast and of the assumptions behind it, so that the explanation cannot be constructed after the fact.

Capital allocation and material judgements. Verifiability is poor, consequence is high and reversibility is low. This is where class-triggered challenge should be concentrated, and where the inversion is most damaging. The relevant evidence is not new: chief executives identified as overconfident on the basis of their personal option-exercise behaviour invest more when internal funds are plentiful and undertake more value-destroying acquisitions.[81] Artificial intelligence does not create this problem; it supplies it with better-looking support.

Controls, risk and assurance. Where agentic systems execute transactions, the control environment must move upstream, because detective controls operate after an agent has acted. This is the area in

[81] Ulrike Malmendier and Geoffrey Tate, ‘CEO Overconfidence and Corporate Investment’, *Journal of Finance*, 60 (6) (2005), 2661–2700; and ‘Who Makes Acquisitions? CEO Overconfidence and the Market’s Reaction’, *Journal of Financial Economics*, 89 (1) (2008), 20–43.

which the absence of supervisory guidance is most consequential, and where the finance function's existing authority over the control environment gives it both the mandate and the obligation to act before regulators do.

9. Capital Allocation under Artificial Intelligence

Four changes to investment appraisal follow from the argument. None requires new systems, and each can be implemented in a single planning cycle.

EXHIBIT 12

APPRAISING ARTIFICIAL INTELLIGENCE CAPITAL

Four questions the finance function should attach to every material AI investment case.

01	02	03	04
What is the complement?	**How will benefit be proved?**	**When should it appear?**	**What life are we assuming?**
What organisational change is required for the benefit to appear, what does it cost, and who owns it?	Staggered rollout with pre-registered metrics. Self-reported time savings are not evidence.	Firm-level evidence puts the lag at two to three years. Judge against that, not the first budget cycle.	Peer useful-life assumptions differ by years. State the judgement to the audit committee.

The historical evidence suggests the complement is the larger number — and it is usually invisible in the paper

Appraise the complement, not only the asset. Conventional appraisal addresses projected cash flows, discount rates and risk-adjusted returns. An artificial-intelligence case should additionally state what organisational change is required for the projected benefit to materialise, what that change costs, and who owns it. The historical evidence in Chapter 4 suggests the complement is the larger number and it is ordinarily invisible in the paper.

Require experimental evidence of benefit. Given the measurement result in Chapter 3, no benefit should be recognised on the basis of self-reported time savings. Staggered rollout across comparable teams, randomised assignment where feasible, and outcome metrics defined before deployment produce a defensible estimate at modest cost. Every credible task-level result cited in this paper was obtained this way, several inside operating companies. A chief financial officer requiring this standard for a $50 million artificial-intelligence programme applies no more rigour than would be routine for a $50 million marketing programme.

Extend post-investment review to organisational lessons. The finance function already performs variance analysis on major investments. Extending that review to record what intelligence was available, what was challenged, what was assumed, and where a model was trusted and should not have been is the cheapest available mechanism for building the feedback architecture of Chapter 6 — and it is the mechanism by which the organisation maps its own jagged frontier.

State the replacement cycle, and separate it from the reporting estimate. Two questions are routinely conflated here and the distinction matters. The first is an appraisal question: over what period will this infrastructure remain economically competitive, given the pace at which new units improve? That belongs in the cash-flow projection, and where the answer is shorter than the financing tenor it belongs in the required return. The second is a reporting question: what useful life and residual value are assumed for depreciation? That is an accounting estimate constrained by the expected pattern of consumption of economic benefits and subject to audit, not a discretionary lever. Where tax lives are set independently of book lives, as under the United States modified accelerated cost recovery system, the book assumption does not enter project cash flows at all and changes the presentation, not the economics; where book and tax lives are linked, it moves the depreciation tax shield and therefore does affect net present value. It also carries real consequences through covenants, deferred tax and earnings-linked compensation. The governance point is therefore comparability and not appraisal, and it is a real one: assumed lives for server and network equipment differ by a full year across firms with similar estates, one major operator has shortened while others extended, and secondary-market evidence on accelerator residual values sits awkwardly beside the longer assumptions. A board should be able to reconcile its own assumption both to its peers and to its own replacement plan, and should be told when the two diverge.

For Institutional Investors

Disclosure is thin but not absent. Five items are observable or can be asked for directly: whether a named board committee holds artificial intelligence in its remit; whether the company discloses any mechanism for challenging AI-supported recommendations; whether stated benefits rest on controlled measurement or management estimate; the useful-life assumption applied to AI infrastructure and whether it has changed; and whether the company can describe a material initiative it stopped, and what it learned. The last is the most diagnostic and the least likely to be volunteered.

Investors should also expect the sign of the near-term effect to be negative. If conversion capability accumulates with a lag, firms investing seriously in it will show depressed near-term margins and improved decision quality before improved reported performance. A market that penalises the former without observing the latter will misprice precisely the firms doing this correctly, a testable asset-pricing prediction, and a candidate explanation for the AI-exposure premium already documented.[82]

[82] Andrea L. Eisfeldt, Gregor Schubert and Miao Ben Zhang, ‘Generative AI and Firm Values’, NBER Working Paper 31222 (revised January 2026); Nicola Borri, Yukun Liu and Aleh Tsyvinski, ‘AI Premium’, arXiv:2606.30583 (July 2026).

PART V

05

Propositions and Research Agenda

What must be measured, how it can be identified, and what would falsify the argument entirely.

01 **Four propositions**
Each with a direction, a conditioning variable and an archival proxy.

02 **Identification**
The April 2026 model-risk revision and the staggered EU AI Act obligations.

03 **What would falsify this**
Absent incremental explanatory power, the construct should be abandoned.

Part V — Propositions, Boundaries and Research Agenda

10. Propositions and Research Agenda

The argument is stated as four propositions. Each carries a direction, a conditioning variable and a measurement strategy. Section 10.1 sets out which quantities may legitimately be elicited from participants and which may not, since the two are not the same and an earlier version of this argument conflated them.

Proposition 1. For decisions supported by machine-generated inference, the correlation between a decision-maker's subjective confidence and the accuracy of the decision is *lower* than for comparable decisions supported by human-generated intelligence, and is expected to be *negative* in the region beyond the system's competence and where no signal capable of registering the error is available, whether because the outcome is slow, because it is never externally scored, or because no counterfactual exists. The claim is local to that region; averaged over a decision population dominated by within-competence and quickly verified judgements, the association may remain positive, and a test that pools across regions should be expected to find little. The effect is expected to be increasing in the fluency of the output and decreasing in the inspectability of its derivation.

Proposition 2. The effect of the inversion on realised decision quality is *decreasing in outcome verifiability*. Where outcomes are observable within the decision cycle at low cost, organisations map the frontier empirically and the effect attenuates; where outcomes are slow, noisy or confounded, it persists. The coded corpus in Chapter 2 bears on this proposition without testing it. It establishes that measured effects depend on what the measure can register, which is a claim about measurement; the proposition concerns the moderation of a confidence–accuracy relationship, and only three of the thirty-five studies elicited confidence at all. A properly powered test would code a larger corpus of field experiments on measure hardness and verification latency and estimate the attenuation directly.

Proposition 3. Class-triggered challenge is associated with higher decision quality than confidence-triggered challenge, *conditional on* equal total challenge capacity. The difference is *increasing* in the share of decisions supported by machine-generated inference and *decreasing* in outcome verifiability, and is expected to be *indistinguishable from zero* where verification is fast and cheap. The two interaction terms, not the main effect of challenge intensity, carry the explanatory weight; a positive main effect alone would be consistent with the long-standing and uninteresting claim that more governance is better.

Proposition 4. Enterprise Intelligence Governance explains variation in decision quality *incremental to* organisation capital, measured management practice, measured artificial-intelligence capability and conventional governance indices. Absent such incremental explanatory power, the construct is redundant and should be abandoned.

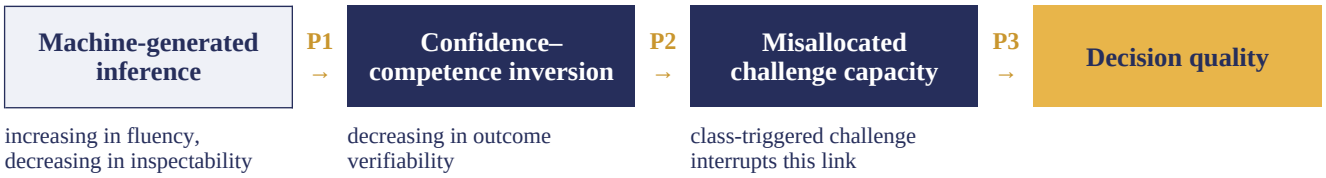
Proposition 4 is stated in this form deliberately. It is the proposition most likely to fail, and the one on which the case for a new construct rests. It is also the hardest to execute, because the four controls are measured in four different datasets with little overlap: the World Management Survey is weighted toward unlisted manufacturing, the artificial-intelligence capability instrument is a survey with a sample

in the low hundreds, and organisation capital is estimated from listed-firm accounts. The realistic path is staged, not simultaneous — first against organisation capital and governance indices on listed firms, where the data exist; then against management practice on the intersection; then, if a research instrument for governance is fielded, against artificial-intelligence capability directly. A single regression containing all four is not currently feasible, and I do not pretend that it is.

EXHIBIT 13

THE PROPOSITIONS AS A CHAIN

Each link carries a direction and a conditioning variable.



P4 Enterprise Intelligence Governance explains variation in decision quality incremental to organisation capital, measured management practice, measured artificial-intelligence capability and conventional governance indices. Absent that, the construct should be abandoned.

Offered for empirical testing; no causal claim is asserted

10.1 Measurement, and What Self-Report Can and Cannot Do

A distinction has to be drawn here that an earlier version of this argument elided. Asking a person to *estimate their own benefit* (how much time they saved, how much better the decision was) is inadmissible, for exactly the reason established in Chapter 3: it is the quantity the inversion corrupts. Asking a person to state their confidence in a specific judgement, elicited before the outcome is known and scored against that outcome, is not the same act. It is the standard calibration protocol, the confidence rating is the construct itself and not an assessment of it, and no alternative exists: Proposition 1 is a claim about the correlation between confidence and accuracy and cannot be tested without eliciting confidence. What must be avoided is self-assessed performance, not self-report as such.

Decision quality, by contrast, must be measured through observable consequences, not through instruments completed by the people whose decisions are being assessed. Three cautions apply to the proxies below. Acquirer announcement returns capture the market's contemporaneous opinion, which is itself increasingly formed from machine-assisted analysis, so the measure is not independent of the treatment and should be paired with a slower one. Firm-year outcome measures do not exist at the level of the individual decision, where the propositions are stated, which forces either aggregation or a within-firm design. And where verifiability is lowest — the region in which the theory predicts the largest effect — outcome measures are weakest by construction. That is an uncomfortable symmetry and it bounds what any archival test can establish.

EXHIBIT 14

MEASURING DECISION QUALITY FROM THE OUTSIDE

Archival proxies that do not depend on asking anyone how well they decided.

Major decisions	Acquirer announcement returns; subsequent divestiture rate	*Moeller, Schlingemann & Stulz (2005)*
Capital allocation	Investment–q sensitivity, intangible-adjusted	*Peters & Taylor (2017)*
Information architecture	Earnings-announcement lag; restatement frequency	*Disclosure literature*
Managerial bias	Option-exercise-based overconfidence measures	*Malmendier & Tate (2005, 2008)*
Organisation capital	Capitalised selling, general and administrative spend	*Eisfeldt & Papanikolaou (2013)*
Learning	Rate at which initiatives are terminated and reviewed	*Proposed*

None of these requires a survey, and none can be inflated by optimism

Acquirer announcement returns are the established archival proxy for the quality of the most consequential class of corporate decision, and the evidence that large acquisitions systematically destroy acquirer value is well documented.[83] Subsequent divestiture rates provide a slower and more direct measure. Investment–q sensitivity, constructed on an intangible-adjusted basis, captures whether capital is directed toward opportunity.[84] Earnings-announcement lag, restatement frequency and forecast dispersion proxy for information architecture. The architectures themselves admit similar treatment: governance through disclosed committee structure and named accountable owners; learning through the observable rate at which initiatives are terminated.

10.2 Identification

Governance research has been poorly served by correlational designs; associations reported in the early 2000s were subsequently attributed in large part to learning and mispricing, not to causal effects.[85] Four settings are available now.

The April 2026 revision of model risk guidance is the most direct. It relaxed independence requirements for a defined population, banking organisations above a stated asset threshold, on a known date, providing both a treated group and a discontinuity. The staggered obligations of the EU Artificial

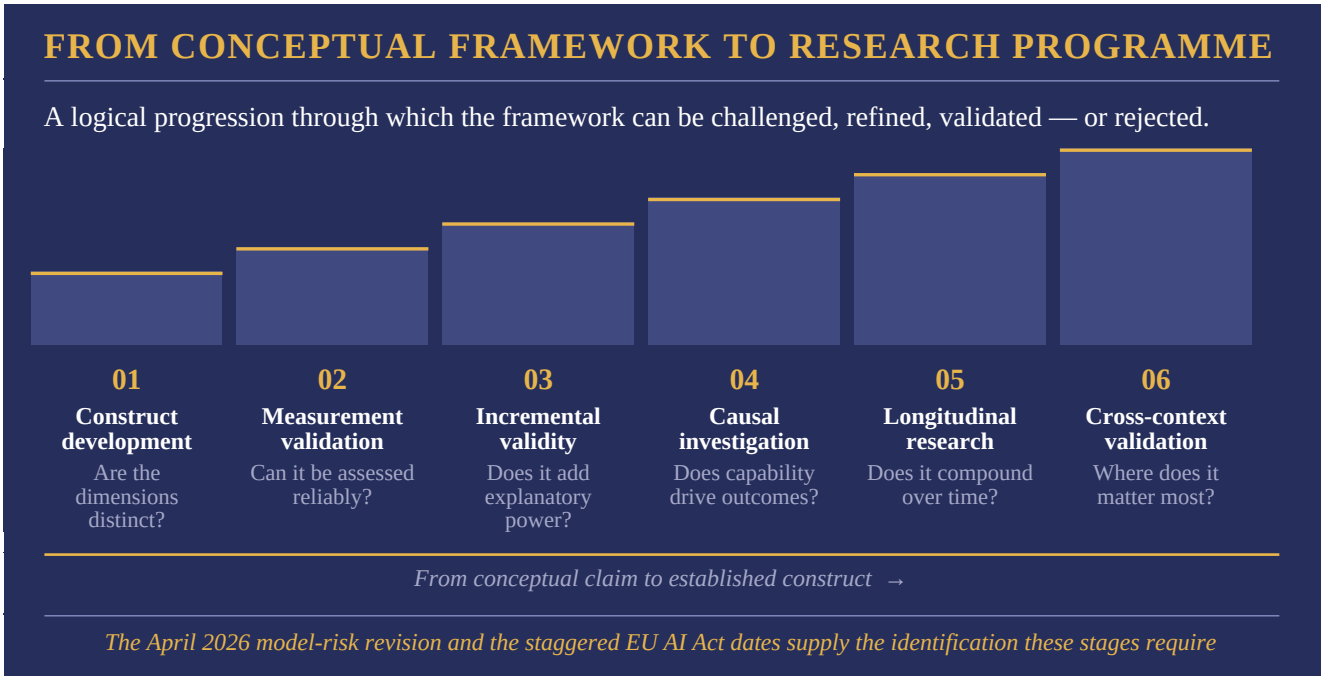
[83] Sara B. Moeller, Frederik P. Schlingemann and René M. Stulz, 'Wealth Destruction on a Massive Scale? A Study of Acquiring-Firm Returns in the Recent Merger Wave', *Journal of Finance*, 60 (2) (2005), 757–782.

[84] Ryan H. Peters and Lucian A. Taylor, 'Intangible Capital and the Investment-q Relation', *Journal of Financial Economics*, 123 (2) (2017), 251–272.

[85] Paul A. Gompers, Joy L. Ishii and Andrew Metrick, 'Corporate Governance and Equity Prices', *Quarterly Journal of Economics*, 118 (1) (2003), 107–155; Lucian Bebchuk, Alma Cohen and Allen Ferrell, 'What Matters in Corporate Governance?', *Review of Financial Studies*, 22 (2) (2009), 783–827.

Intelligence Act impose different governance requirements on different system classes at dates between February 2025 and August 2028. The Article 50 transparency deadline of 2 August 2026 supplies a sharp date for pre- and post-treatment comparison. And *Sarbanes-Oxley section 404* remains the canonical precedent for a mandated internal-control shock with measurable consequences for information quality.

EXHIBIT 15



11. Boundary Conditions and Limitations

A construct that claims to apply everywhere explains nothing. Three boundary conditions bound this one, and the first is derived, not asserted.

Outcome verifiability. The argument applies where the environment does not supply a timely independent signal of decision accuracy. Where it does (high-frequency, quickly verified operational decisions) the inversion self-corrects and the governance apparatus proposed here is unnecessary overhead. This boundary is moving, and moving in the direction that shrinks the argument. Simulation, digital twins and cheap counterfactual estimation are extending fast verification into decisions that historically resisted it, and each such extension removes a class from the scope of this paper. The residual — decisions whose outcomes are confounded by the actions of competitors, regulators and markets, and which are made once — is smaller than it looks and is where the argument will end up living.

Separation of intelligence and authority. The construct presupposes an organisation large and differentiated enough that relevant intelligence and decision authority reside in different people. Where the two coincide, as in owner-managed firms, it collapses into individual managerial cognition and no separate capability is warranted.

Inspectability. The argument concerns inference whose derivation is not fully inspectable. Where a model's reasoning can be examined at acceptable cost, conventional validation applies and the model risk apparatus is adequate.

Limitations

This paper is conceptual and has not tested its propositions. Five limitations are material. The first is the one I would press hardest.

The evidence cannot reach the region the theory is about. Every one of the thirty-five studies coded in Chapter 2 is a case in which an outcome was eventually observed, that is what made it codeable at all. The argument concerns decisions where it does not arrive. The gradient therefore supports a claim about measurement, and the extension of that claim to unverifiable decisions is an inference, not an observation. This is structural: no further coding of the experimental literature will remove it, because a study of the unverifiable case cannot report an outcome.

The calibration base is two studies. Direct evidence that decision-makers fail to detect machine-induced error consists of one randomised trial with sixteen developers and one classroom experiment; two further studies often cited in this connection elicited beliefs about the tool, not about the participant's own performance. The human-factors literature is much larger but was developed on automation whose reliability was stationary and estimable, precisely the property this paper argues generative systems lack, so the transfer carries the problem without the cure.

The cross-level step is argued and not demonstrated. Chapter 3 states four conditions under which individual miscalibration aggregates into organisational misallocation of challenge. None has been measured. In particular, no study establishes that the discretionary margin of review in a large firm is in fact allocated on felt uncertainty and not on observable proposal features.

Penetration is low where the argument matters most. Machine-generated inference is not yet present at scale in capital allocation and acquisitions, which makes the claim prospective exactly where it is strongest and contemporaneous where it matters least. The mechanism predicts a problem that is arriving, not one already measurable.

The empirical base is young and partly unrefereed. Much of the firm-level evidence dates from 2024 to 2026, several key studies remain working papers, and the field is moving faster than the peer-review cycle. Two prominent papers in this area have been withdrawn or disavowed since 2025, in one case after an institution stated it had no confidence in the provenance of the data, which is itself the reason for the evidential discipline adopted here and for the reliability class attached to each source.[86] The construct's four architectures are derived from the lifecycle of an inference and not from inductive analysis of data, which is a weaker warrant than a measurement paper would supply.

12. Conclusion

Artificial intelligence is being deployed at a scale that makes the question of its return a matter of macroeconomic as well as corporate consequence. Three findings from the evidence are firm.

[86] Neither paper is cited anywhere in this manuscript, and both are excluded deliberately rather than overlooked.

The technology works at the level of well-chosen tasks, and the evidence for that is among the better-identified bodies of work in applied economics. The value is not yet arriving at the level of the firm, and the most careful accounting suggests the aggregate prize is modest. The deficiency is organisational, because the gap is widest exactly where organisational correction is weakest.

The mechanism this paper proposes is narrow and, for that reason, testable. Machine-generated inference makes decision-makers most confident where they are least accurate. Organisational challenge is triggered by felt uncertainty. Therefore challenge is allocated in inverse proportion to need, and the organisation experiences this as good governance because no friction is generated. The correction is not more challenge but differently triggered challenge: determined ex ante, by decision class, on consequence, reversibility and outcome verifiability, by someone other than the person advancing the decision.

This is not a claim that artificial intelligence should be resisted, nor that human judgement is reliable where machine judgement is not. The relevant evidence establishes that both are miscalibrated and that the miscalibration compounds. It is a claim about where a scarce governance resource should be spent, and the answer — on decisions that are consequential, irreversible and slow to verify, regardless of how confident anyone is about them — is specific enough to act on and specific enough to be wrong.

> *“ The correction is not more challenge. It is differently triggered challenge — determined in advance, by decision class, by someone other than the person advancing the decision.*

For a chief executive or chief financial officer the operative question is not how much the organisation is spending on artificial intelligence. It is: which categories of decision receive independent challenge automatically, who decided that list, and when was it last revised? Most large organisations cannot presently answer it. On the argument advanced here, the ones that can will be the ones that convert.

A Note on How This Paper Was Built

The corpus came first, and it did not do what I wanted it to do. I began with a hypothesis: that measured gains from artificial-intelligence assistance would fall monotonically as outcome measures got harder, and that the fall would be a clean gradient from volume through rating through consequence to retained capability. I coded thirty-five trials to show it.

There is no gradient. There is a trough, and it sits at band three, the band in which the outcome is something the world does in response. When I looked at what occupies band three, the answer was decisions. Band four sits higher only because it holds four studies and three of them are tutoring trials, where the hard measure is a student's own later capability and artificial intelligence turns out to teach rather well. Chapter 2 is written around that result and not around the hypothesis, which is why its statistical test is null and why it says so.

Four things about the corpus a reader is entitled to know. The band and family codings are mine alone. There is no second coder and therefore no reliability statistic, and it is the first thing I would fix given more time. Three studies I wanted were excluded because their headline result could not be verified against the released paper, one because the magnitudes appear only in a publisher's press release. Four more were excluded on eligibility grounds, including one withdrawn by its repository while this paper was being written, its institution having stated that it had no confidence in the data. And the corpus is a convenience sample of what can be trialled cheaply, which is why coding, education and customer service are over-represented and capital allocation does not appear at all.

One earlier version of the framework defined Enterprise Intelligence Governance as operating in ways that improve decision quality. That builds the outcome into the definition and makes the central proposition true by construction. It is stated without the outcome now, which costs some rhetorical force and buys the only thing that matters, the ability to be wrong.

A note on sources

The regulatory position is stated as at August 2026. Chapter 5 rests on supervisory guidance issued on 17 April 2026 and on the staggered application dates of the European Union Artificial Intelligence Act. Both are moving, and both are moving in the direction that shrinks the argument. A reader consulting this paper later should check them before relying on the claim that no supervisory standard covers generative and agentic systems.

Much of the firm-level evidence is recent and some of it is unrefereed. Where a figure comes from a working paper the footnote says so. Where a published version differs from an earlier working-paper figure, the published figure is used and the difference is noted.

Data availability

The coded corpus of thirty-five trials, with the band and family assignments, the hardest-measure result recorded for each, the exclusion list and the coding rules, is available from the author on request. I would be glad of a second coder.

Appendix A. The Coded Corpus

Thirty-five randomised or quasi-experimental trials of generative-artificial-intelligence assistance, published or posted between 2023 and 2026, coded on the hardest outcome each of them reports. The outcomes are the studies' own. The band and task-family assignments are mine, and a different coder would produce a different table at the margin.

Band records the hardness of the hardest measure the study reports. Band 1, volume or speed with no quality ground truth; band 2, a same-session rating against a rubric or answer key; band 3, an objective outcome carrying a real external consequence; band 4, a delayed or unassisted re-test of the person's own capability. Task family records what was being done: *decision* for judgement, service, diagnosis, allocation and matching; *creation* for writing, code production, ideation and translation; *learning* where the outcome is the human's own subsequent capability. The result column gives the sign and significance of the effect on that hardest measure: + positive and significant, 0 null, − negative and significant. The sign records the direction of the effect on the study's own objective and not the sign of the raw coefficient, so a fall in completion time is recorded as a positive effect.

STUDY	DOMAIN	BAND	TASK FAMILY	RESULT ON THE HARDEST MEASURE REPORTED
Dell'Acqua 2023	consulting	2	decision	− −19pp correctness outside frontier
Dell'Acqua 2025	innovation	2	creation	+ +0.37 SD graded solution quality
Noy & Zhang 2023	writing	2	creation	+ +0.4 SD graded quality
Brynjolfsson 2025	support	3	decision	0 customer NPS −0.12pp, ns
Peng 2023	coding	1	creation	+ −55.8% time
Cui 2025	coding	3	creation	0 build success −5.53%, ns
Becker 2025 (METR)	coding	3	creation	− +19% completion time
Gambacorta 2024	coding	1	creation	+ +55% lines of code
Song 2024	opensource	1	creation	− coordination time +8%
Dillon 2025	knowledge	1	creation	0 meeting time unchanged
Choi/Monahan 2024	law	2	decision	0 quality sig on 1 of 4 tasks
Schwarcz 2026	law	2	decision	+ +0.26 to +0.53 on 7-pt scale
Bastani 2024	education	4	learning	− −17% unassisted exam, GPT Base
Wang 2024 (Tutor)	tutoring	3	learning	+ +4pp real topic mastery
Kestin 2025	education	4	learning	+ +0.63 unassisted post-test
De Simone 2025	education	4	learning	+ +0.206 SD post-programme exam
Contractor 2026	learning	4	learning	+ +0.27 SD one week later, unaided
Ni 2026	service	3	decision	0 3-day retrial rate ≈0.000, ns
Wang 2026 (agentic)	service	3	decision	− customer rating −0.412, p<0.001
Fang 2026	retail	3	decision	+ sales +16.3%; returns/ratings null
Otis 2026	entrepreneur	3	decision	0 +0.05 SD, p=0.36; low performers −0.09

STUDY	DOMAIN	BAND	TASK FAMILY	RESULT ON THE HARDEST MEASURE REPORTED
Doshi & Hauser 2024	creativity	2	creation	+ novelty +8.1%; collective diversity falls
Chen & Chan 2024	advertising	3	decision	− real clicks fall for experts, ghostwriter
Ju & Aral 2025	advertising	3	decision	0 no sig effect on real CTR or CPC
Merali 2024	translation	2	creation	+ +0.18 SD graded per 10x compute
Merali 2025	consulting	1	creation	+ −8% task time per year of progress
Caplin 2024	classify	2	decision	+ +6.9pp accuracy vs ground truth
Jabarian 2026	hiring	3	decision	+ job starts and retention hold
Goh 2024	medicine	2	decision	0 +2pp, p=0.60
Goh 2025	medicine	2	decision	+ +6.5%, p<0.001; no gain vs LLM alone
Lukac 2025	medicine	3	decision	0 no comparative safety effect reported
Thakkar 2025	peer-review	3	decision	+ rebuttal engagement +6%
Bono 2024	IT-admin	2	decision	+ +34.5% accuracy vs ground truth
Bono 2025	IT-admin	2	decision	+ +48% accuracy vs ground truth
Fitzpatrick 2025	public	2	decision	0 +17% documents but −12% data task

\+ positive and significant − negative and significant **0** null *Bands: 1 volume or speed · 2 same-session rating · 3 external consequence · 4 delayed or unassisted re-test*

Three studies were verified as existing but their hardest-measure result could not be established from the released paper, and they are excluded from every count above: Sun 2025 (creativity), magnitudes only in publisher press release; Choi & Xie 2026 (accounting), randomised component not identifiable in released paper; PreA trial 2026 (medicine), downstream clinical outcomes not reported.

Four further studies were verified and excluded on eligibility grounds: van Inwegen 2025, tool is explicitly non-generative (rule-based parser); Toner-Rodgers 2024, withdrawn by arXiv; institution states no confidence in data; Hui, Reshef & Zhou, market-level DiD on exposure, not a trial of assistance; Frey & Llanos-Paredes 2025, instrumental-variables observational design.

The counts reported in Chapter 2 follow directly from the table. By band, the share reporting a positive significant effect on the hardest measure runs 3 of 5 at band one, 9 of 13 at band two, 4 of 13 at band three and 3 of 4 at band four. Restricting to bands three and four, decision tasks return 3 of 10 against 4 of 7 for everything else, on which a two-sided Fisher exact test returns p = 0.35.

Bibliography

Acemoglu, Daron. ‘The Simple Macroeconomics of AI’. *Economic Policy* 40, no. 121 (2025): 13–58.

Aghion, Philippe, and Jean Tirole. ‘Formal and Real Authority in Organizations’. *Journal of Political Economy* 105, no. 1 (1997): 1–29.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. ‘Artificial Intelligence, Firm Growth, and Product Innovation’. *Journal of Financial Economics* 151 (2024): article 103745.

Bain and Company. *Global Technology Report*. Boston: Bain and Company, September 2025.

Bank of England and Financial Conduct Authority. *Artificial Intelligence in UK Financial Services 2024*. London: Bank of England, November 2024.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. ‘Generative AI Can Harm Learning’. SSRN Working Paper 4895486, 2024.

Bebchuk, Lucian, Alma Cohen, and Allen Ferrell. ‘What Matters in Corporate Governance?’. *Review of Financial Studies* 22, no. 2 (2009): 783–827.

Becker, Joel, Nate Rush, Beth Barnes, and David Rein. ‘Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity’. arXiv:2507.09089, July 2025.

Bloom, Nicholas, and John Van Reenen. ‘Measuring and Explaining Management Practices Across Firms and Countries’. *Quarterly Journal of Economics* 122, no. 4 (2007): 1351–1408.

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen. ‘What Drives Differences in Management Practices?’. *American Economic Review* 109, no. 5 (2019): 1648–1683.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. ‘Americans Do IT Better: US Multinationals and the Productivity Miracle’. *American Economic Review* 102, no. 1 (2012): 167–201.

Board of Governors of the Federal Reserve System. SR 11-7, ‘Guidance on Model Risk Management’. 4 April 2011. Superseded.

Board of Governors of the Federal Reserve System. SR 26-2, ‘Revised Guidance on Model Risk Management’. 17 April 2026.

Bolton, Patrick, and Mathias Dewatripont. ‘The Firm as a Communication Network’. *Quarterly Journal of Economics* 109, no. 4 (1994): 809–839.

Bono, Jonathan, and Alexander Xu. ‘Randomized Controlled Trials for Security Copilot for IT Administrators’. arXiv:2411.01067, 2024.

Borri, Nicola, Yukun Liu, and Aleh Tsyvinski. ‘AI Premium’. arXiv:2606.30583, July 2026.

Boyd, Brian K., Gregory G. Dess, and Abdul M. A. Rasheed. ‘Divergence between Archival and Perceptual Measures of the Environment: Causes and Consequences’. *Academy of Management Review* 18, no. 2 (1993): 204–226.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. ‘Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence’. *Quarterly Journal of Economics* 117, no. 1 (2002): 339–376.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. ‘Generative AI at Work’. *Quarterly Journal of Economics* 140, no. 2 (2025): 889–942.

Brynjolfsson, Erik, Lorin M. Hitt, and Shinkyu Yang. ‘Intangible Assets: Computers and Organizational Capital’. *Brookings Papers on Economic Activity* 2002, no. 1 (2002): 137–198.

Caplin, Andrew, David J. Deming, Shangwen Li, Daniel J. Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye. ‘The ABC’s of Who Benefits from Working with AI: Ability, Beliefs, and Calibration’. NBER Working Paper 33021, October 2024.

Challapally, Aditya, Chris Pease, Ramesh Raskar, and Pradyumna Chari. *The GenAI Divide: State of AI in Business 2025*. MIT NANDA, July 2025.

Chen, Zenan, and Jason Chan. ‘Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise’. *Management Science* 70, no. 12 (2024): 9101–9117.

Choi, Jonathan H., Amy Monahan, and Daniel Schwarcz. ‘Lawyering in the Age of Artificial Intelligence’. *Minnesota Law Review* 109 (2024): 147–208.

Choi, Jung Ho, and Chloe Xie. ‘Human + AI in Accounting: Early Evidence from the Field’. *Journal of Accounting Research* (2026). DOI 10.1111/1475-679X.70052.

Committee of Sponsoring Organizations of the Treadway Commission. *Enterprise Risk Management — Integrating with Strategy and Performance*. COSO, 2017.

Contractor, Zachary, and Germán Reyes. ‘Experimental Evidence on the Learning Impact of Generative AI’. IZA Discussion Paper 18792, July 2026.

Cui, Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. ‘The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers’. *Management Science* (2026), online first. DOI 10.1287/mnsc.2025.00535.

Cyert, Richard M., and James G. March. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall, 1963.

De Simone, Martín E., Federico H. Tiberti, María Rocío Barrón Rodríguez, Federico A. Manolio, Wuraola Mosuro, and Eliot J. Dikoru. ‘From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria’. World Bank Policy Research Working Paper 11125, 2025.

Dell’Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. ‘Navigating the Jagged Technological Frontier’. *Organization Science* 37, no. 2 (2026): 403–423.

Denrell, Jerker, and James G. March. ‘Adaptation as Information Restriction: The Hot Stove Effect’. *Organization Science* 12, no. 5 (2001): 523–538.

Denrell, Jerker. ‘Vicarious Learning, Undersampling of Failure, and the Myths of Management’. *Organization Science* 14, no. 3 (2003): 227–243.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. ‘Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err’. *Journal of Experimental Psychology: General* 144, no. 1 (2015): 114–126.

Dillon, Eleanor W., Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton. ‘Shifting Work Patterns with Generative AI’. NBER Working Paper 33795, May 2025.

Duncan, Robert B. ‘Characteristics of Organizational Environments and Perceived Environmental Uncertainty’. *Administrative Science Quarterly* 17, no. 3 (1972): 313–327.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. ‘Organization Capital and the Cross-Section of Expected Returns’. *Journal of Finance* 68, no. 4 (2013): 1365–1406.

Eisfeldt, Andrea L., Gregor Schubert, and Miao Ben Zhang. ‘Generative AI and Firm Values’. NBER Working Paper 31222, revised January 2026.

European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union, 2024.

Fang, Yuqi, Yifan Yuan, Ruoxuan Zhang, Cristina Donati, and Miklos Sarvary. ‘Generative AI and Firm Productivity: Field Experiments in Online Retail’. arXiv:2510.12049, October 2025.

Federal Deposit Insurance Corporation. FIL-15-2026, ‘Agencies Revise the Interagency Model Risk Management Guidance’. 17 April 2026.

Ferrando, Annalisa, Maria Lamboglia, Judit Rariga, and Katharina Schmidt. ‘Adoption and Investment in Artificial Intelligence across the Euro Area’. ECB Occasional Paper No. 395. Frankfurt: European Central Bank, 2026.

Galbraith, Jay R. ‘Organization Design: An Information Processing View’. *Interfaces* 4, no. 3 (1974): 28–36.

Gambacorta, Leonardo, Han Qiu, Shuo Shan, and Daniel M. Rees. ‘Generative AI and Labour Productivity: A Field Experiment on Coding’. BIS Working Paper 1208. Basel: Bank for International Settlements, September 2024.

Goh, Ethan, Robert Gallo, Jason Hom, and others. ‘Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial’. *JAMA Network Open* 7, no. 10 (2024): e2440969.

Goh, Ethan, Robert J. Gallo, Eric Strong, and others. ‘GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial’. *Nature Medicine* 31, no. 4 (2025): 1233–1238.

Gompers, Paul A., Joy L. Ishii, and Andrew Metrick. ‘Corporate Governance and Equity Prices’. *Quarterly Journal of Economics* 118, no. 1 (2003): 107–155.

Graham, John R., Campbell R. Harvey, and Manju Puri. ‘Capital Allocation and Delegation of Decision-Making Authority within Firms’. *Journal of Financial Economics* 115, no. 3 (2015): 449–470.

Hambrick, Donald C., and Phyllis A. Mason. ‘Upper Echelons: The Organization as a Reflection of Its Top Managers’. *Academy of Management Review* 9, no. 2 (1984): 193–206.

Helfat, Constance E., and Margaret A. Peteraf. ‘Managerial Cognitive Capabilities and the Microfoundations of Dynamic Capabilities’. *Strategic Management Journal* 36, no. 6 (2015): 831–850.

Institute of Internal Auditors. *The IIA’s Three Lines Model*. Lake Mary, FL: IIA, 2020.

International Auditing and Assurance Standards Board. *ISA 540 (Revised): Auditing Accounting Estimates and Related Disclosures*. New York: IAASB.

International Monetary Fund. *Global Financial Stability Report*. Washington, DC: IMF, April 2026.

International Organization for Standardization. ISO/IEC 42001:2023, *Information Technology — Artificial Intelligence — Management System*. Geneva: ISO, December 2023.

Jaakkola, Elina. ‘Designing Conceptual Articles: Four Approaches’. *AMS Review* 10, no. 1 (2020): 18–26.

Joseph, John, and William Ocasio. ‘Architecture, Attention, and Adaptation in the Multibusiness Firm: General Electric from 1951 to 2001’. *Strategic Management Journal* 33, no. 6 (2012): 633–660.

Ju, Harang, and Sinan Aral. ‘Collaborating with AI Agents: Field Experiments on Teamwork, Productivity, and Performance’. arXiv:2503.18238, 2025.

Kahneman, Daniel, Dan Lovallo, and Olivier Sibony. ‘Before You Make That Big Decision…’. *Harvard Business Review* 89, no. 6 (2011): 50–60.

Kestin, Gregory, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. ‘AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting’. *Scientific Reports* 15 (2025): article 17458.

Klein, Gary. ‘Performing a Project Premortem’. *Harvard Business Review* 85, no. 9 (2007): 18–19.

Kruger, Justin, and David Dunning. ‘Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments’. *Journal of Personality and Social Psychology* 77, no. 6 (1999): 1121–1134.

Lawrence, Paul R., and Jay W. Lorsch. *Organization and Environment: Managing Differentiation and Integration*. Boston: Harvard University Graduate School of Business Administration, 1967.

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina. ‘To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis’. *Organization Science* 33, no. 1 (2022): 126–148.

Lebovitz, Sarah, Natalia Levina, and Hila Lifshitz-Assaf. ‘Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What’. *MIS Quarterly* 45, no. 3 (2021): 1501–1526.

Lee, John D., and Katrina A. See. ‘Trust in Automation: Designing for Appropriate Reliance’. *Human Factors* 46, no. 1 (2004): 50–80.

Levinthal, Daniel A., and James G. March. ‘The Myopia of Learning’. *Strategic Management Journal* 14, no. S2 (1993): 95–112.

Levitt, Barbara, and James G. March. ‘Organizational Learning’. *Annual Review of Sociology* 14 (1988): 319–340.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. ‘Algorithm Appreciation: People Prefer Algorithmic to Human Judgment’. *Organizational Behavior and Human Decision Processes* 151 (2019): 90–103.

Malmendier, Ulrike, and Geoffrey Tate. ‘CEO Overconfidence and Corporate Investment’. *Journal of Finance* 60, no. 6 (2005): 2661–2700.

Malmendier, Ulrike, and Geoffrey Tate. ‘Who Makes Acquisitions? CEO Overconfidence and the Market’s Reaction’. *Journal of Financial Economics* 89, no. 1 (2008): 20–43.

March, James G., and Herbert A. Simon. *Organizations*. New York: Wiley, 1958; 2nd edn, Oxford: Blackwell, 1993.

McKinsey and Company. *The State of AI*. November 2025.

Mikalef, Patrick, and Manjul Gupta. ‘Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on Its Impact on Organizational Creativity and Firm Performance’. *Information and Management* 58, no. 3 (2021): article 103434.

Milliken, Frances J. ‘Three Types of Perceived Uncertainty About the Environment: State, Effect, and Response Uncertainty’. *Academy of Management Review* 12, no. 1 (1987): 133–143.

Moeller, Sara B., Frederik P. Schlingemann, and René M. Stulz. ‘Wealth Destruction on a Massive Scale? A Study of Acquiring-Firm Returns in the Recent Merger Wave’. *Journal of Finance* 60, no. 2 (2005): 757–782.

National Institute of Standards and Technology. ‘Announcing the AI Agent Standards Initiative for Interoperable and Secure Innovation’. 17 February 2026.

Ni, Xiao, and others. ‘Generative AI in Action: Field Experimental Evidence from Alibaba’s Customer Service Operations’. arXiv:2603.29888, 2026.

Noy, Shakked, and Whitney Zhang. ‘Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence’. *Science* 381, no. 6654 (2023): 187–192.

Ocasio, William. ‘Towards an Attention-Based View of the Firm’. *Strategic Management Journal* 18, no. S1 (1997): 187–206.

Office for National Statistics. *Artificial Intelligence in UK Businesses: 2023 to 2026*. Newport: ONS, July 2026.

Office of the Comptroller of the Currency. Bulletin 2000-16, ‘Risk Modeling: Model Validation’. 30 May 2000.

Office of the Comptroller of the Currency. Bulletin 2011-12, ‘Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management’. 4 April 2011.

Office of the Comptroller of the Currency. Bulletin 2026-13, ‘Model Risk Management: Revised Guidance’. 17 April 2026.

Otis, Nicholas G., Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning. ‘The Uneven Impact of Generative AI on Entrepreneurial Performance’. *Management Science* (2026). DOI 10.1287/mnsc.2024.06909.

Ouchi, William G. ‘A Conceptual Framework for the Design of Organizational Control Mechanisms’. *Management Science* 25, no. 9 (1979): 833–848.

Parasuraman, Raja, and Dietrich H. Manzey. ‘Complacency and Bias in Human Use of Automation: An Attentional Integration’. *Human Factors* 52, no. 3 (2010): 381–410.

Parasuraman, Raja, and Victor Riley. ‘Humans and Automation: Use, Misuse, Disuse, Abuse’. *Human Factors* 39, no. 2 (1997): 230–253.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. ‘The Impact of AI on Developer Productivity: Evidence from GitHub Copilot’. arXiv:2302.06590, February 2023.

Peters, Ryan H., and Lucian A. Taylor. ‘Intangible Capital and the Investment-q Relation’. *Journal of Financial Economics* 123, no. 2 (2017): 251–272.

Ryseff, James, Brandon F. De Bruhl, and Sydne J. Newberry. *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed*. RR-A2680-1. Santa Monica, CA: RAND Corporation, 2024.

Shrestha, Yash Raj, Shiko M. Ben-Menahem, and Georg von Krogh. ‘Organizational Decision-Making Structures in the Age of Artificial Intelligence’. *California Management Review* 61, no. 4 (2019): 66–83.

Simons, Robert. *Levers of Control: How Managers Use Innovative Control Systems to Drive Strategic Renewal*. Boston: Harvard Business School Press, 1995.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. ‘Does Automation Bias Decision-Making?’ *International Journal of Human-Computer Studies* 51, no. 5 (1999): 991–1006.

Skitka, Linda J., Kathleen L. Mosier, and Mark D. Burdick. ‘Accountability and Automation Bias’. *International Journal of Human-Computer Studies* 52, no. 4 (2000): 701–717.

Stanford Institute for Human-Centered Artificial Intelligence. *AI Index Report 2026*. Stanford, CA: Stanford HAI, April 2026.

Tushman, Michael L., and David A. Nadler. ‘Information Processing as an Integrating Concept in Organizational Design’. *Academy of Management Review* 3, no. 3 (1978): 613–624.

United States Census Bureau. *Business Trends and Outlook Survey*. Washington, DC: US Census Bureau, May 2026.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. ‘When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis’. *Nature Human Behaviour* 8 (2024): 2293–2303.

Vaughan, Diane. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press, 1996.

Wang, Rose E., Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dorottya Demszky. ‘Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise’. arXiv:2410.03017, 2024.

Wang, Yiwei, Chenhao Zhu, Tianjun Feng, Lauren Xiaoyuan Lu, and Bowen Jia. ‘Agentic AI and Human-in-the-Loop Interventions: Field Experimental Evidence from Alibaba’s Customer Service Operations’. arXiv:2605.14830, 2026.

Weick, Karl E., and Kathleen M. Sutcliffe. *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*. 2nd edn. San Francisco: Jossey-Bass, 2007.

Whetten, David A. ‘What Constitutes a Theoretical Contribution?’. *Academy of Management Review* 14, no. 4 (1989): 490–495.

Yotzov, Ivan, Nicholas Bloom, Steven J. Davis, Philip Bunn, Lucy Barrero, and others. ‘Firm Data on AI’. NBER Working Paper 34836. Cambridge, MA: National Bureau of Economic Research, February 2026.

Yu, Zhengyang, Lee Fleming, Jaclyn Hampton, Christophe Combemale, and Neil Thompson. ‘AI Adoption in S&P 500 Firms’. arXiv:2607.08920, July 2026.

EIG | **ENTERPRISE INTELLIGENCE GOVERNANCE**

Organisations have always relied on a person's own uncertainty to tell them which decisions deserve a second look. Machine inference removes that signal, and nothing has yet been designed to replace it.

AUTHOR
Olga Pisarenko

PUBLISHED
August 2026

STATUS
Conceptual framework — offered for empirical testing

SUGGESTED CITATION
Pisarenko, O. (2026). Enterprise Intelligence Governance: Rebuilding Organisational Challenge for Machine-Generated Inference.