WHITE PAPER · PORTFOLIO & AI GOVERNANCE

AI as a Thinking Partner

Why the dominant approach produces faster workers with weaker judgment — and what the evidence says about the alternative

Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

March 2026

The $30 billion bet that is backfiring

Enterprises have spent an estimated $30–40 billion annually deploying AI tools. Ninety-five percent report no measurable P&L impact. The dominant explanation is poor workflow integration or unclear ROI. The actual explanation is simpler and more uncomfortable: organizations taught people how to go faster without teaching them how to think better — and in doing so, began quietly eroding the judgment that justified the investment.

This is not a technology problem. It is a training design problem with a measurable consequence. Two peer-reviewed RCTs now bracket the outcome space precisely. Students using standard AI with no guidance scored 17 percent worse on independent assessments than a control group that used no AI at all. Students using AI designed to require reasoning first scored more than twice as well as a control group using best-practice human instruction. Same tool. Different interaction design. The gap between those two outcomes is not theoretical — it is the gap between what most organizations are currently producing and what is demonstrably achievable.

The problem begins before the workforce. By the time an employee arrives with entrenched AI shortcuts, those habits were formed in school — under prohibition, without guidance, optimized for the only visible goal: get the task done. Seventy-eight percent of knowledge workers are bringing personal, unsanctioned AI tools to work. Approximately 70 percent of active AI users have received no training from their employer. The habit vacuum is not empty. It is full of whatever people taught themselves under pressure. And the research now shows exactly what that produces.

The design gap that defines the outcome

Two peer-reviewed RCTs define the range of what is possible with AI in learning and judgment-intensive work.

Bastani et al. (PNAS, 2025) randomized approximately 1,000 students across three conditions: standard AI, a guardrailed tutor, or no AI. Standard AI users scored 17 percent worse on unassisted assessments than the control group. The guardrailed tutor — same underlying model, redesigned to require students to reason through hints before receiving answers — eliminated the deficit entirely. The harm was not caused by AI. It was caused by a specific interaction design that allowed users to bypass their own reasoning.

Kestin et al. (Scientific Reports, 2025) randomized 194 undergraduates to a custom AI tutor or best-practice active learning: peer instruction, small groups, real-time instructor feedback. The AI tutor produced more than twice the learning gains in less time — 49 minutes versus 60 — with statistical significance of $p < 10^{-8}$. The control was not passive lecturing. It was the current gold standard of human instruction.

Interaction Design	Population	Outcome vs. Control
Standard AI — answer-first, ungoverned	~1,000 students	**−17% on unassisted assessment**
Guardrailed AI tutor — reasoning-first	~1,000 students	**Parity restored — harm reversed**
Custom AI tutor — reasoning-first, engineered	194 undergraduates	**+2× learning gains in 20% less time**

The gap between −17 percent and +2× is not a gap between two tools. It is a gap between two interaction designs applied to the same technology. It is entirely a design decision. Most institutions are currently making it by default, in the wrong direction.

Scope limitation: The Kestin study used Harvard undergraduates studying introductory physics over two weeks. The researchers explicitly state results target middle-order cognitive skills and cannot be presumed to extend to higher-order synthesis or professional judgment. The effect is real and robust — and conditional on the deliberate interaction design this paper argues for.

The real state of AI use

Condition	Share	Source
U.S. students who consider AI essential for success	**65%**	HEPI 2025
UK university students using AI for assessments	**88%**	HEPI 2025
Teachers who allow any AI use	**38%**	RAND 2025
Teachers who received training on AI misuse response	**28%**	Dwyer & Laird 2024
U.S. knowledge workers using AI at work	**45%**	Gallup Q3 2025
Bringing personal, unsanctioned AI tools	**78% of AI users**	Microsoft / LinkedIn WTI 2024
Received any AI training from employer	**30% of AI users**	Microsoft / LinkedIn WTI 2024
Would not stop using AI even if banned	**~48%**	SANS, Oct 2024
Share confidential data with AI tools without approval	**38%**	CybSafe / NCSA 2024

If 45 percent of the U.S. workforce uses AI at work and only 30 percent received any employer training, approximately 70 percent of all active AI users are operating on a self-taught framework. That framework was reverse-engineered from the tool's interface, from peer behavior, and from the pressure to produce output quickly and invisibly. The top reported use cases — consolidating information (42 percent), generating ideas (41 percent) — are output-first behaviors. Neither requires interrogating assumptions or sustaining the reasoning effort that durable expertise demands.

The prohibition pipeline compounds this. Most students encountered AI first under institutional ban — using it anyway, at home, without guidance, optimizing for task completion. That habit follows them into the workforce. Enterprises trying to train AI skills into this population are not filling a blank slate. They are contending with an entrenched default built under adverse conditions over years.

What the research shows

Study	Population	Method	Key Finding	Evidence Quality
Bastani et al. (2025), PNAS	~1,000 students	RCT	Standard AI: −17% on exams. Guardrailed tutor: parity restored.	High — peer-reviewed RCT, pre-registered
Kestin et al. (2025), Scientific Reports	194 undergraduates	RCT	Reasoning-first AI tutor: 2× learning gains vs. best-practice active learning, $p < 10^{-8}$	High — peer-reviewed RCT; scope limited to introductory physics
Liu et al. (2025), Scientific Reports	3,562 workers, 4 studies	Experimental	AI raised output quality but reduced intrinsic motivation 11% and increased boredom 20% on solo tasks	High — peer-reviewed, four replications
Lee et al. (2025), Microsoft / CMU	319 knowledge workers	Survey	Higher AI confidence → less critical thinking applied; higher domain confidence → more	Moderate — peer-reviewed CHI; self-report
Gerlich (2025), MDPI Societies	666 participants	Survey + interviews	Frequent AI use negatively correlated with critical thinking; balanced use attenuated the effect	Moderate — peer-reviewed; no objective performance measure
Kosmyna et al. (2025), MIT Media Lab	54 adults (18 in session 4)	EEG preprint	LLM group: up to 55% reduced neural connectivity; 83% memory recall deficit	Low — preprint; underpowered; formal methodological concerns

Note on the MIT preprint: Formal peer commentary identified five methodological problems including underpowered sample (n=18 for the key session 4 finding), absent pre-registration, and EEG inconsistencies. The lead author explicitly asks the work not be characterized as evidence of permanent harm. It is cited as directionally consistent only and plays no role in the recommendation architecture below.

What Gerlich's nuance means operationally

Gerlich's finding that balanced use attenuates the negative effect is the empirical description of what the accountability partner model produces. Participants who used AI for administrative tasks while preserving reasoning effort for complex and evaluative work did not show the critical thinking decline observed in heavy users. The research supports a specific behavioral pattern — not a general restriction on AI use.

Why professionals degrade on complex tasks

Dell'Acqua et al. (Research Policy, 2024) ran a pre-registered field experiment with 758 BCG consultants.[5] For within-frontier tasks, AI users completed 12.2 percent more tasks, worked 25 percent faster, and produced outputs rated more than 40 percent higher in quality. For one task selected outside the frontier, they were 19 percentage points less likely to reach a correct solution than those without AI. Degradation was concentrated in consultants who adopted AI outputs without interrogating them — *blind adoption* in the researchers' language. The tool gave no signal it was operating beyond its capability. Institutional note: BCG co-designed a study finding BCG consultants perform better with AI. Pre-registration and multi-institutional authorship partially mitigate that interest.

Tool-level overconfidence

The calibration problem is not only behavioral. Studies across GPT-4, GPT-3.5, LLaMA2, and PaLM 2 find systematically high calibration error — high-confidence outputs regardless of accuracy. A controlled study of 257 medical students making 3,855 diagnostic decisions found a transparency paradox: AI explanations improved decisions when correct (+6.3 points) but worsened them when wrong (−4.9 points), because AI generates equally persuasive explanations regardless of recommendation quality.[6] Domain expertise is a necessary floor — not a substitute for structured interrogation, but a prerequisite for it.

The evidence gaps this paper cannot close

No study directly compares cognitive outcomes between formally trained workers and self-taught BYOAI users. No enterprise-scale RCT has tested reasoning-first versus answer-first AI interaction for professional knowledge workers. Both are real gaps. The recommendations below are high-confidence inference — not established proof. Closing these gaps is the applied research agenda the field needs.

Why prohibition fails — and what it costs

Prohibition does not prevent AI use. It prevents guided use. The Bastani and Kestin studies were not comparing AI against no AI — they were comparing ungoverned AI against carefully designed AI. Prohibition produces the ungoverned condition at scale, then exports it to the workforce.

The cost is not only harm prevention forgone. It is the force multiplier surrendered. Kestin's 2× gain over best-practice human instruction is a measured result against a strong control. Institutions that ban

AI or allow unrestricted access without design are not choosing between risk and safety. They are choosing between −17 percent and +2× and most are choosing the worse outcome by default.

The security cost is separate and compounding. Shadow AI breaches cost $670,000 more per incident than standard breaches and account for 20 percent of all AI-related breaches. The average enterprise experiences 223 incidents per month of sensitive data submitted to unvetted AI tools — a figure that doubled year over year. Ninety percent of security leaders use unapproved tools themselves. Prohibition does not reduce exposure. It drives it underground.

Why the current approach fails structurally

- **Completion is the only frame.** No institution has provided a frame that makes interrogating AI's output the point. In school the goal is submitting the assignment. At work the goal is closing the task. The tool's interface reinforces both.
- **No baseline for reasoning quality.** Activity data is captured; reasoning quality is not. The harm accumulates in a measurement blind spot. Organizations cannot address what they are not measuring.
- **No designed friction.** The Bastani findings are precise: harm was not caused by AI availability but by an interaction design that allowed users to bypass reasoning. The fix is friction, not restriction. Prohibition removes access without replacing it with productive struggle.
- **The tool is trusted where it should be interrogated.** The transparency paradox means AI generates equally persuasive explanations for correct and incorrect outputs. Training must compensate for what the tool structurally cannot signal.
- **No accountability loop.** A worker who accepts AI output without interrogation and one who rigorously pressure-tests it look identical on any activity dashboard. Without an external signal that reasoning quality matters, throughput is the only visible standard — and the collective action dynamic penalizes individual departure from it.

Why the stakes are higher than they appear

The wage premium

Autor and Thompson's 2025 NBER working paper provides the labor economics frame: when automation eliminates non-expert tasks from an occupation, the wage premium for the remaining expert tasks rises; when it eliminates expert tasks, wages fall.[22] AI is automating answer-generation, routine synthesis, and information-consolidation — the non-expert component of knowledge work. The scarcity value of judgment, evaluation, and reasoning is rising accordingly. Workers who preserve those capabilities through the accountability partner model will capture an increasing wage premium. Workers who offload judgment entirely are eliminating from their own work profile the tasks whose

value is rising. This converts the paper's argument from a cognitive health warning into a career self-interest case — which partially resolves the temporal discount problem: if the accruing benefit is personal and wage-linked, present-bias is more tractable.

The agentic AI horizon

Generative AI still involves humans in both task execution and judgment. Agentic AI — autonomous systems that plan and complete multi-step workflows with minimal human input — shifts the human role almost entirely to goal-setting, output evaluation, and exception handling. That is precisely the task stewardship capacity that deteriorates fastest under unguided AI use. Workers who trained themselves to bypass reasoning under generative AI will be underprepared for the one distinctly human role in the architecture that is already being built. The training decisions being made right now are more consequential than they currently appear.

A better standard: AI as accountability partner

Working standard — defined for this paper:

AI is functioning as a cognitive partner when it is used to surface, challenge, and validate a user's own reasoning — not replace it — and when the user's judgment, informed by AI interrogation, is treated as the authoritative conclusion.

AI outputs become starting positions for thinking, not ending positions. The default interaction is interrogative: pressure-test the plan before producing it; identify the weakest assumption before making the argument; raise the counterarguments before writing the brief.

When does this standard apply?

Task Type	Consequence Level	Recommended Approach
Within AI's reliable frontier	Low — routine, well-defined	Full delegation appropriate
Within AI's reliable frontier	High — strategic, consequential	Verify outputs; check reasoning chain
Outside or near AI's frontier	Any	Pre-mortem protocol required before AI engagement
Unknown frontier position	High	Treat as outside-frontier; pre-mortem required

The pre-mortem prompt protocol

Before engaging AI on any high-judgment task, the worker completes three steps without AI: (1) state the hypothesis or recommendation they expect to reach and why; (2) identify the two or three assumptions that most need to be true; (3) articulate what evidence would change their view. AI is then introduced to interrogate the pre-mortem — challenge the assumptions, raise counterarguments, locate the weakest reasoning. The worker's judgment, challenged by AI, is the conclusion.

This has independent validation. Ma et al. (CHI 2024) developed a structurally identical intervention — 'Thinking the Opposite' — and found it improved self-confidence calibration and directly reduced AI over-reliance. Klein's original pre-mortem technique has its own validated lineage in team decision-making.

Two qualifications: First, AI-generated challenges carry the same confident tone regardless of correctness. The protocol improves the user's reasoning posture — it does not resolve the tool's calibration problem. For tasks outside the user's domain expertise, AI challenges must be validated against external domain knowledge, not accepted as authoritative. Second, self-directed habit formation without external reinforcement has low sustained adoption under time pressure. This protocol changes the default when embedded in institutional processes. Individual adoption without that embedding is fragile.

Two tracks, not one timeline

Installing the pre-mortem habit is a 60–90 day behavioral change. Recovering degraded reasoning capacity is a 12–36 month restorative effort. Conflating the two produces a program that overpromises on the slower problem and underserves the faster one.

Track	Who	Intervention	Timeline
Track 1 Protective	Workers with intact reasoning capacity who have not yet formed deep shortcut habits	Pre-mortem protocol. Interrogation habit formation. AI skepticism training.	60–90 days for habit installation
Track 2 Restorative	Workers with measurable reasoning atrophy from sustained BYOAI shortcut patterns	Sustained domain reasoning without AI. Expert feedback on judgment quality. Deliberate AI-removed tasks.	12–36 months depending on depth of atrophy and domain complexity

The audit is a capability triage, not a tool inventory. Establishing which workers are in Track 1 versus Track 2 requires two components: a standard tool use inventory, and domain-relevant judgment tasks performed without AI assistance and evaluated for reasoning quality. Without the second component, training is deployed without knowing whether it is protecting or restoring — two different problems requiring different interventions at different timescales. The most urgent population is your most senior

judgment-dependent roles that have been heaviest BYOAI users. They represent the highest organizational risk and the longest restoration timeline.

Track 2 restorative curriculum at professional scale does not yet exist. This is a named research gap. Organizations running Track 2 programs are doing so without validated playbooks. The research agenda that would close this gap — enterprise-scale trials of structured AI-removal practice with expert judgment feedback — is the work the field needs to prioritize.

The three-layer change model

All three layers must activate simultaneously. The collective action problem means individual adoption ahead of institutional reinforcement is structurally penalized. The temporal discount dynamic means workers take the shortcut under pressure without external enforcement. Sequencing these layers is not an option.

Layer 1 — Training design

Owner: *CHRO and L&D function (Schools: curriculum coordinators and teacher professional development leads)*

- **Frame this as a decision quality initiative, not an AI initiative.** 'AI initiative' triggers compliance instincts and turf conflict. 'Decision quality initiative' is a language every P&L owner speaks. The outcome being pursued is better judgment — AI is the context, not the product.
- **Run capability triage before training.** Establish what employees are using and for what tasks — with disclosure decoupled from discipline. Augment tool inventory with domain judgment tasks performed without AI to identify Track 1 versus Track 2 populations. Deploying the same training to both wastes resources and misses the restorative population entirely.
- **Introduce the pre-mortem protocol as a performance advantage, not a mandate.** The Dell'Acqua finding is the self-interest argument: consultants who interrogated AI outputs outperformed blind adopters with identical credentials. That is a performance case, not a compliance case.
- **Teach the tool's limits explicitly.** LLMs are structurally overconfident. Training must address what AI cannot reliably signal: the boundaries of its own competence. Teach workers to treat AI confidence level as a non-diagnostic signal and to apply domain expertise as the primary validity check on AI-generated challenges.
- **Evaluate reasoning, not output.** Assessment criteria should center on how well the learner articulated the problem, how rigorously they interrogated AI responses, and how defensible their final judgment is. Measuring only output completion reinforces the throughput frame the training is designed to displace.

Layer 2 — Usage pattern configuration

Owner: *IT Security and Operations, in coordination with CHRO (Schools: IT and academic integrity office)*

- **Governance before control.** Banning BYOAI produces the worst outcome: continued use with added concealment. Provision reduces the incentive for shadow behavior by meeting the underlying need through sanctioned alternatives. The BYOD parallel is instructive: formal policies reduce but do not eliminate unauthorized use. No AI-specific evidence of equivalent design currently exists; the BYOD analog is a structural parallel, not direct proof.
- **Implement the pre-mortem protocol as an organizational norm for high-judgment tasks.** Tasks involving strategy, significant resource decisions, client-facing analysis, or risk assessment require cognitive priming before AI engagement. This is a sequencing standard, not a restriction on AI use.
- **Audit full-delegation patterns.** Identify tasks handed entirely to AI without a verification or reasoning loop. For tasks involving judgment or ambiguous tradeoffs, full delegation without a reasoning layer is risk accumulation — as the Dell'Acqua outside-frontier findings demonstrate directly.

Layer 3 — Institutional accountability

Owner: *Direct managers and executive leadership (Schools: department heads and academic leadership)*

This layer is not optional. It is the mechanism that resolves the collective action problem and the temporal discount dynamic simultaneously. Without it, individual adoption is penalized and training investment decays under throughput pressure within 60–90 days.

- **Make reasoning quality visible.** Review processes for AI-assisted work must probe the reasoning behind conclusions, not only the conclusions. The operative question is not 'did AI help produce this?' It is 'what is the reasoning behind this conclusion, and how was it validated?' The distinction between conclusions reflecting the employee's judgment and those reflecting AI output with an approval signature must become visible to management.
- **Track task stewardship capacity.** Define this as the ability to set appropriate goals for AI, evaluate its outputs, and make sound independent judgments when AI is inadequate. Assess it before and after training interventions. It is a leading indicator — reasoning atrophy shows up in decision quality only after it is already significant.
- **Signal that reasoning quality is rewarded.** Explicit recognition of reasoning quality in performance reviews, promotion criteria, and project retrospectives changes the present-moment calculus. This is the mechanism that makes the accountability partner model individually rational, not just institutionally desirable.
- **Protect the pilot from the existing measurement system.** Teams practicing the pre-mortem protocol will produce work more slowly than peers who are not — in the short term, against any

throughput metric. The executive sponsor's primary function in the first 90 days is not resource allocation. It is holding the measurement question open long enough for quality signal to emerge. This is the minimum viable executive support the initiative requires.

Call to action: implementation without a mandate

The three-layer model works when an organization has decided to act. Most have not and will not. The corporate mandate is to conduct business. AI is a tool. An implementation model that requires full organizational backing will stall before it starts.

Durable change at this scale moves by demonstration, not decree. The practical sequence:

Phase	Timeline	Action
Name the owner	Week 1	Assign a single executive — CHRO with CIO as co-sponsor — explicitly accountable for all three layers. Not a working group. One person who owns the outcome.
Run capability triage	Month 1	Grant amnesty for BYOAI disclosure. Inventory tool use. Assess reasoning depth through AI-removed judgment tasks. Identify Track 1 versus Track 2 populations. This is the only way to know what you are actually dealing with.
Pilot one workflow	Months 2–3	Select one function where reasoning quality is already consequential — strategy, risk, client advisory, product decisions. Run the pre-mortem protocol for one quarter. Define decision quality criteria now, before the pilot begins.
Rewrite one rubric	Months 3–6	Add task stewardship as an assessed competency in one function. Make reasoning quality visible in one review cycle. This is the signal that the institutional incentive has changed.
Evaluate and scale	Month 6	If the pilot shows signal, you have the internal proof of concept that makes scaling tractable. If it does not, you have spent 90 days and learned something real before committing the enterprise.
Culture shift	18–36 months	From first pilot to embedded norm in the functions that matter most. This is what durable culture change costs when it has to prove its value rather than assume it. That is not a failure of ambition. It is an accurate forecast.

The one thing that kills it: treating Layer 3 as downstream of Layers 1 and 2. If reasoning quality is not made visible and rewarded simultaneously with the training rollout, the protocol decays in 60–90 days and the organization returns to its prior baseline. Layer 3 is not the finish line. It is the foundation.

The shift this produces

Before: students form shortcut habits under prohibition; they arrive at enterprises conditioned to BYOAI behavior; enterprises deploy without guidance or ban without effect; reasoning quality erodes in a measurement blind spot; the 2× force multiplier sits unrealized while the −17% outcome scales.

After: institutions guide AI use rather than prohibit it, building the interrogation habit where it first forms. Enterprises receive workers with a usable foundation rather than an entrenched default. Sanctioned tools reduce shadow use incentive. The pre-mortem protocol becomes standard for high-judgment tasks. Reasoning quality is measured, visible, and rewarded. The force multiplier becomes an institutional asset.

This is achievable without full mandate support. It requires starting with the willing, protecting the pilot from the existing measurement system, and letting demonstrated decision quality do the selling. Culture change at this scale moves by proof, not proclamation.

The underlying argument

The gap between −17 percent and +2× is not ambiguous. It is a measured outcome difference produced entirely by interaction design. Every institution that defaults to the former is accepting a specific, quantified cost while an alternative with a specific, quantified benefit is available. That alternative does not require new technology. It requires a different definition of what AI is for.

The tools that produce the worst outcomes and the tools that produce the best outcomes are the same tools. The difference is whether the interaction was designed to demand reasoning or to replace it.

One clarification on what this requires culturally: it is recovery, not construction. The professional norm that reasoning is visible and valued, that you own your conclusions, that judgment is a professional asset — this preceded AI by decades. AI did not introduce a new standard. It eroded an existing one, in a measurement blind spot, faster than most organizations noticed. The accountability partner model does not ask people to adopt a new value system. It asks them to act consistently with the one they already hold.

The habit vacuum is not neutral. It is full of whatever people taught themselves under pressure, optimized for the only goal that was visible to them. **The research now answers what happens in each case. The choice is institutional.**

Notes and sources

[1] Challapally et al. (2025). The GenAI Divide: State of AI in Business 2025 (v0.1). MIT NANDA. The $30–40B annual investment figure and 95% no-measurable-P&L-impact figure are from this preliminary

report. Methodology: 52 structured interviews, 300+ disclosed AI initiatives, 153 senior leaders. https://www.artificialintelligence-news.com/wp-content/uploads/2025/08/ai_report_2025.pdf

[2] Microsoft and LinkedIn, 2024 Work Trend Index Annual Report. Survey of 31,000 knowledge workers, 31 markets, Feb–Mar 2024. The 78% BYOAI figure includes any personal tool use supplementary to employer-provided tools and may overstate fully unsanctioned behavior. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part

[3] SANS Institute, "Sunlight AI: Bringing Shadow AI Into the Light," December 2025. The 48% would-not-stop figure draws on October 2024 survey data. https://www.sans.org/blog/sunlight-ai-bringing-shadow-ai-light

[4] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., and Mariman, R. (2025). Generative AI without guardrails can harm learning. PNAS, 122(26), e2422633122. Pre-registered RCT; ~1,000 students, 9th–11th grade; Turkey; IRB-approved. August 2025 correction updated one author's affiliation only; findings unchanged. https://doi.org/10.1073/pnas.2422633122

[5] Dell'Acqua, F., et al. (2024). Navigating the jagged technological frontier. Research Policy, 53. Pre-registered; n=758 BCG consultants. "Blind adoption" and retainment score from Appendix B of HBS WP No. 24-013. BCG co-designed a study finding BCG consultants perform better with AI; pre-registration and multi-institutional authorship partially mitigate this. https://www.hbs.edu/faculty/Pages/item.aspx?num=64700

[6] Jungherr, A., and Rauchfleisch, A. (2025). The transparency paradox in AI-assisted decision making. n=257 medical students; 3,855 diagnostic decisions. Also: Groot, T., and Valdenegro-Toro, M. (2024). Overconfidence is key. Proceedings of TrustNLP 2024, ACL. https://aclanthology.org/2024.trustnlp-1.13/

[7] IBM, Cost of a Data Breach Report 2025. Shadow AI breach premium ($670K) and 20% share of AI-related breaches. https://www.ibm.com/security/data-breach

[8] Netskope, Cloud and Threat Report, October 2024–October 2025. 223 incidents/month; year-over-year doubling of sensitive data uploads. https://www.cybersecuritydive.com/news/shadow-ai-security-risks-netskope/808860/

[9] UpGuard survey of security leaders, 2025. Reported in Fortra, November 2025. https://www.fortra.com/blog/shadow-ai-security-breaches-will-hit-40-companies-2030-warns-gartner

[10] Lee, H-P., et al. (2025). The impact of generative AI on critical thinking. CHI '25, Yokohama. Microsoft Research. "Task stewardship" is the authors' own framing. Limitations: self-report; cross-sectional; critical thinking not objectively measured; causality not established. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking

[11] Cybersecurity Insiders / Bitglass, BYOD Security Report 2024. Cited as structural analog. No AI-specific provision study of equivalent design currently exists.

[12] Kosmyna, N., et al. (2025). Your brain on ChatGPT. arXiv:2506.08872. MIT Media Lab. Preprint — not peer-reviewed as of March 2026. Formal commentary arXiv:2601.00856 identifies five methodological concerns. Cited as directionally consistent only. https://arxiv.org/abs/2506.08872

[13] Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), 6. n=666. Balanced-use finding is a primary result. No objective performance measures. https://doi.org/10.3390/soc15010006

[14] Liu, Y., Wu, S., Ruan, M., Chen, S., and Xie, X. (2025). Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. Scientific Reports. Four studies; N=3,562. Earlier drafts of this paper incorrectly attributed this study to "Park et al."; the correct lead author is Yukun Liu. https://www.nature.com/articles/s41598-025-98385-2

[15] CybSafe and National Cybersecurity Alliance, survey of 7,000 respondents, late 2024. The 38% confidential-data-sharing figure is from Cloud Security Alliance analysis. https://cloudsecurityalliance.org/blog/2025/03/04/ai-gone-wild-why-shadow-ai-is-your-it-team-s-worst-nightmare

[16] Microsoft, New Future of Work Report 2025. The inference that resistance to top-down mandates extends specifically to established BYOAI interaction habits is the author's extension; not directly established by the cited source. https://www.microsoft.com/en-us/research/wp-content/uploads/2025/12/New-Future-Of-Work-Report-2025.pdf

[17] Klein, G. (2007). Performing a project premortem. Harvard Business Review, 85(9), 18–19. Validated in: Veinott, E. S., Klein, G., and Wiggins, S. (2010). Evaluating the effectiveness of the PreMortem technique on plan confidence. ISCRAM, Seattle. The adaptation here — individual cognitive priming before AI engagement — extends Klein's team-based technique and has not been independently validated in that specific form. https://hbr.org/2007/09/performing-a-project-premortem

[18] Ma, S., et al. (2024). "Are you really sure?" CHI '24, Honolulu. N=94. The "Thinking the Opposite" intervention is structurally identical to the pre-mortem prompt protocol and was developed independently. Limitation: income-prediction task in a lab setting. https://dl.acm.org/doi/10.1145/3613904.3642671

[19] Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning. Scientific Reports. RCT; n=194 Harvard undergraduates; introductory physics; fall 2023. Effect size 0.73–1.3; $p < 10^{-8}$. Critical limitations: selected population; two topics; middle-order cognitive skills only; custom prompt-engineered tutor not replicable with off-the-shelf tools. https://www.nature.com/articles/s41598-025-97652-6

[20] Dwyer, L., and Laird, E. (2024). AI in K-12 schools: Teacher survey data. Cited in Curran, F. C., and Goo, M. (2025). Disciplining AI use. UF EPRC. https://education.ufl.edu/eprc/files/2025/11/EPRC_PolicyBrief-AI-and-School-Discipline-Curran-and-Goo-2025.pdf

[21] HEPI, Student Academic Experience Survey 2025. UK university student AI use for assessments: 53% (2024) to 88% (2025). Programs.com synthesis of HEPI, Stanford HAI, and Microsoft data for the 65% essentiality and 38% teacher-allow figures. https://www.demandsage.com/ai-in-education-statistics/

[22] Autor, D., and Thompson, C. (2025). Expertise and the wage premium in an era of automation. NBER Working Paper. The inference that AI is specifically automating the non-expert tasks of knowledge work is the author's application of the framework; Autor and Thompson do not make this specific claim about generative AI.

March 2026

About the author

Marco Policani is an enterprise portfolio, PMO, and AI operating-governance leader. He builds portfolio governance systems where AI carries the analytical load and named humans own every decision — an approach documented across case studies, walkthroughs, and working governance guides on his portfolio site.

Portfolio & governance library: policani.net/governance · Full portfolio: policani.net · LinkedIn: linkedin.com/in/marcpolicani