THE AI INSTITUTE / OPERATING INTELLIGENCE

EVIDENCE RETROSPECTIVE · Retrospective 2023 / R01

When the Frontier Was Jagged

What leaders could reasonably know about generative AI in 2023, what the next three years changed, and why the original evidence never supported universal automation.

Boards, strategy leaders, transformation offices and research teams
Research period: January 2023 – December 2023
First published: 5 August 2026
Evidence current to 31 July 2026

PUBLICATION RECORD

An Institute thesis, supported by evidence

The early evidence did not show that generative AI improved all knowledge work. It showed a jagged capability frontier: large gains inside a task-specific boundary, new errors outside it, and a management requirement to evaluate the whole human–AI configuration.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

01	**The strongest 2023 experiment was evidence about a defined task set and population, not a universal productivity coefficient.**
02	**The central strategic error was converting local capability evidence into enterprise-wide transformation claims.**
03	**Subsequent studies strengthened the case for conditional value while weakening the idea that access alone redesigns work.**
04	**Historical evidence is most useful when leaders recover its original boundaries, then test whether those boundaries still apply locally.**

Research method

This evidence retrospective reconstructs the public decision record for January–December 2023, then tests it against research published through 31 July 2026. Contemporaneous evidence includes field experiments, economic working papers and NIST AI RMF 1.0. Subsequent evidence includes representative adoption surveys, peer-reviewed and working-paper field studies, official statistics and later implementation guidance. We extracted the population, task, intervention, outcome and stated limitations for central claims; treated working papers and vendor-involved research according to their disclosed status; and avoided combining incomparable effect sizes. The Institute frameworks are analytical syntheses, not independently validated causal models. Source notes label whether evidence was available during the period or added through hindsight.

First published 5 August 2026 · Updated not revised · Research period January 2023 – December 2023 · Version 1.0 · Suggested citation: The AI Institute, *When the Frontier Was Jagged* (2026).

CONTENTS

Inside this paper

R **References**

How to read the evidence

External research supports the Institute's thesis; it does not substitute for it. Figures retain their original populations and definitions. Institute frameworks and inferences are labelled, and limitations are recorded before the references.

01 HOW THIS RETROSPECTIVE WORKS

Two clocks, not one invented history

This paper uses two clocks. The first is the 2023 decision clock: evidence, standards and arguments that were publicly available during that calendar year. The second is the hindsight clock: evidence published from 2024 to July 2026 that helps test what the early claims did and did not predict. Sources are labelled contemporaneous or subsequent so hindsight is never passed off as prior knowledge.

A retrospective is not a reproduction. This paper was first published in 2026, its PDF metadata and suggested citation remain 2026, and the cover names the period being studied rather than implying an earlier release. The research question is narrower: what could a diligent executive team reasonably have concluded by December 2023, and how should that conclusion be revised now?

The period opened with rapid public access to capable generative systems and closed with the first influential field experiments, a new US risk-management framework and an expanding debate about jobs, skills and governance. Those signals arrived at different levels of evidence. A model demonstration, a controlled experiment, a worker survey and an operating standard answer different questions and should not be collapsed into one narrative. [134]

The Institute reconstructs claims rather than sentiment. For each major proposition, we identify the population, task, intervention, outcome and boundary. Later evidence is used to test transportability, not to make 2023 actors appear prescient.

EXHIBIT / TWO EVIDENCE CLOCKS

2023	2024	2025	2026
Contemporaneous record	**Adoption becomes measurable**	**Field evidence broadens**	**Operating thesis revised**
Experiments, frameworks and economic arguments available during the period.	Representative use surveys begin separating access, frequency and work hours.	Results diverge across customer service, knowledge work and software development.	The unit of analysis shifts from model capability to workflow configuration.

02 THE 2023 RECORD

Capability arrived faster than organisational evidence

By early 2023, leaders could see that general-purpose language systems could produce useful text, analysis and code across many tasks. What they could not see was a mature distribution of enterprise outcomes. Demonstrations established possibility; they did not establish reliability, economics, control effectiveness or the organisational changes required for durable value.

NIST released AI RMF 1.0 in January 2023 around four continuous functions—govern, map, measure and manage. Its existence is important historical evidence: the need for contextual mapping, measurement and lifecycle governance was not an objection added after deployment problems appeared. It was part of the available operating record from the start. [3]

Economic analysis during 2023 also warned against treating automation as a one-directional labour story. The Turing Transformation argued that task automation could change skill premiums and expand who can perform complementary work. It was a theoretical mechanism, not a forecast that every organisation would realise those benefits, but it made the outcome focus explicit. [4]

A reasonable 2023 position was therefore neither ‘ignore this’ nor ‘transform everything’. It was to build an evidence portfolio: identify candidate tasks, measure outcomes, map risk, and preserve the ability to learn as capabilities and work practices changed.

03 THE LANDMARK EXPERIMENT

Large gains appeared—inside a defined frontier

The most influential enterprise experiment studied 758 consultants performing 18 realistic tasks. On tasks judged to sit inside the tested system's capability frontier, access increased completion, speed and human-rated quality. The experimental design and preregistration made this far stronger than product anecdotes. [12]

The result was economically meaningful but bounded. The participants were consultants from one firm; the work was a selected set of knowledge tasks; the system and prompting approach reflected a particular point in model development; and the outcome measures were task performance rather than realised client or firm economics. Those are not defects. They define what the study can support.

The study also found different interaction patterns. Some participants divided work between human and system; others integrated AI throughout the task. That variation matters because 'AI access' was not a single treatment in practice. People built different configurations, and those configurations changed how capability entered the work. [1]

The defensible 2023 conclusion was conditional: for some professionally relevant tasks, some workers using a capable system could produce more or better output quickly. It was not evidence that every workflow should be automated, that quality would remain high beyond the task boundary, or that the surrounding organisation would capture the gain.

> **Contemporaneous finding — the experiment established a material task effect. Institute inference — task effects become enterprise value only through workflow, control and capacity decisions.**

EXHIBIT / WHAT THE 2023 EXPERIMENT ESTABLISHED—AND WHAT IT DID NOT

758	18	+25%	Local
Participants	**Tasks**	**Speed signal**	**External validity**
A substantial professional sample in one consulting organisation.	Selected realistic tasks, not the full distribution of consulting work.	Approximate improvement for tasks inside the evaluated frontier.	A task and population result, not a universal firm productivity estimate.
HBS working paper [2]	HBS working paper [2]	HBS summary [1]	

04 BOUNDARY CONDITIONS

Outside the frontier, confidence could amplify error

The same research programme supplied the warning that later headlines often removed. On a task outside the tested capability frontier, participants with AI were less likely to reach the correct answer. A system could make work feel fluent and complete while moving the user away from the right result. [12]

This creates an asymmetric management problem. Gains inside the frontier are visible as faster production. Failures outside it can be hidden by plausible language, especially when reviewers lack the expertise or time to reconstruct the reasoning. The risk is not simply a hallucinated sentence; it is misplaced organisational confidence.

The frontier is also dynamic. Model updates can move it, but changing tools do not remove the need to identify it. A task that works in one context can fail after data, instructions, users, integrations or consequence change. Evaluation therefore belongs to the deployed configuration and must be repeated after material change.

A board could already have asked in 2023: which tasks have been tested, where did performance deteriorate, who can recognise failure, and what happens when the system is uncertain? Those questions remain more useful than a general claim that the next model is better.

05

THE NARRATIVE FAILURE

A task result became a transformation story

Market discussion converted early evidence through four leaps. It moved from selected tasks to whole jobs; from participants to all workers; from output measures to financial value; and from a temporary model configuration to a stable technological rule. Each leap might become true in a particular organisation, but none followed automatically from the experiment.

The confusion was reinforced by the word productivity. In a study it can mean tasks completed, issues resolved or time spent. At enterprise level it may mean value added relative to labour and capital. A faster draft does not establish higher firm productivity if review grows, demand is fixed, saved capacity is not redeployed or errors create downstream cost.

The generalisation error also obscured heterogeneity. Workers differed in baseline experience, task approach and reliance on the system. Later customer-service evidence found gains concentrated among less experienced workers, with far smaller speed gains and small quality deterioration among the most skilled. [11]

The Institute's retrospective judgment is that 2023 evidence was strong enough to justify structured experimentation and too weak to justify universal operating claims. The error was not optimism. It was failing to preserve denominators, task boundaries and outcome definitions while translating research into strategy.

EXHIBIT / FOUR UNSUPPORTED LEAPS

01	02	03	04
Task → job	**Participant → workforce**	**Output → value**	**Snapshot → law**
Selected task performance does not specify how an occupation changes.	One professional population does not represent every role or skill level.	Task metrics do not include adoption cost, rework or capacity allocation.	One model and workflow configuration does not define a permanent frontier.

06 WHAT HAPPENED NEXT

Use spread through people before operating models changed

Representative US surveys later showed how quickly individual use spread. By late 2024, nearly 40% of adults aged 18–64 reported using generative AI, 23% of employed respondents had used it for work in the previous week and 9% used it every workday. The same research estimated only 1–5% of work hours were assisted. [5]

Those figures resolve an apparent contradiction: a technology can diffuse rapidly through workers while remaining shallow in total work. Adoption counted as people touched, weekly users or hours assisted produces different answers. None measures whether a controlled workflow changed or whether the employer realised value.

This was an unmanaged adoption wave because the unit of diffusion was often the individual account. Employees could change drafting, research and analysis behaviour without a use-case register, process baseline, data boundary or shared evaluation. Formal transformation programmes were therefore observing only part of the change.

The historical lesson is that visibility must precede confident scaling. Organisations need ways to discover embedded AI features and employee-built practices without criminalising exploration. Otherwise the official portfolio measures only centrally funded projects while material work changes elsewhere.

EXHIBIT / FAST DIFFUSION, SHALLOW WORK PENETRATION

~40%	23%	9%	1–5%
Adults using GenAI	**Weekly work use**	**Daily work use**	**Hours assisted**
US adults aged 18–64 reporting any use by late 2024.	Employed respondents using GenAI for work in the previous week.	Employed respondents reporting use every workday.	Estimated share of total work hours touched by the tools.
NBER [5]	NBER [5]	NBER [5]	NBER [5]

07 GOVERNANCE IN THE RECORD

The operating disciplines were available early

NIST AI RMF 1.0 framed governance as continuous and cross-cutting, supported by mapping, measurement and management. It emphasised context, affected people, human roles and lifecycle change. That structure undercuts the later idea that governance necessarily arrived as paperwork after innovation. [3]

The framework did not prescribe one control set or prove that implementation would create financial value. It offered a vocabulary and outcome structure. Organisations still needed to translate it into owners, inventories, evaluation methods, monitoring, incident processes and decisions about acceptable risk.

Subsequent guidance made those operational artefacts more explicit. NIST's AI Resource Center and ARIA pilot extended the operational focus toward contextual testing, evaluation, verification and validation; the generative AI profile added technology-specific risks and actions. Australia's Guidance for AI Adoption organised implementation around accountability, impact planning, risk management, transparency, testing and monitoring, and human control. [9101819]

The retrospective conclusion is not that every firm should have implemented a mature control plane in 2023. It is that experimental velocity and governance did not need to be sequential. A lightweight register, named owner, scoped data boundary, task evaluation and stop rule were already compatible with rapid learning.

08 JOBS AND CAPABILITY

Exposure was repeatedly mistaken for realised displacement

Early labour analysis attempted to map which occupations and tasks could be affected by generative systems. Exposure measures are useful for identifying where work may change, but they are not observed job-loss statistics. They generally do not include adoption cost, demand response, organisational redesign, regulation or the creation of complementary tasks. [68]

The Turing Transformation offered a second mechanism: automation of some tasks can raise the value of complementary skills and expand the pool of people able to perform other work. That possibility makes job-level forecasts dependent on how tasks are recombined, how demand changes and whether workers can move into the complementary work. [4]

The ILO's later refined index retained the transformation emphasis. It estimated one in four workers was in an occupation with some exposure, while 3.3% of global employment sat in the highest exposure category. Its method combined almost 30,000 task definitions, worker assessments, expert deliberation and model predictions; it measured technical exposure, not realised displacement. Stanford administrative payroll analysis later found a relative decline among early-career workers in highly exposed occupations, but remained observational and sensitive to timing and controls. [1417]

A responsible workforce strategy should therefore separate four questions: which tasks are technically exposed, which are economically attractive to change, which changes are organisationally feasible, and what actually happens to jobs and people. Collapsing them produces false precision and weak decisions.

EXHIBIT / EXPOSURE IS THE START OF THE CAUSAL CHAIN

Can	Pays	Fits	Occurs
Technical exposure	**Economic incentive**	**Organisational feasibility**	**Observed outcome**
A system can perform or assist part of a task under test conditions.	Expected benefit exceeds implementation, review, failure and change costs.	Data, roles, systems, controls and demand can support the change.	Employment, task, quality and value effects are measured after adoption.

09 HINDSIGHT TEST

Later studies strengthened conditional value

A customer-support deployment across 5,172 agents later reported a 15% average increase in successful issue resolution per hour. Gains were concentrated among less experienced workers, consistent with AI transferring elements of stronger practice. The setting, system vintage and workflow remained specific, but the operational dataset was far richer than a laboratory task. [11]

A randomised six-month experiment involving 7,137 knowledge workers across 66 firms found active users spent about two fewer hours on email each week. It did not detect broader shifts in task quantity or composition from individual provision. That result supports a distinction between personal efficiency and coordinated workflow change. [12]

An RCT with experienced open-source developers reported that participants took longer when AI tools were available despite expecting the opposite. The sample was small and specialised, and a later methodology update warned that newer-tool estimates were unreliable because of selection and measurement problems. The correct lesson is not that AI slows software work universally; it is that task and worker context can reverse the sign of an effect. [13]

Together these studies strengthen the 2023 frontier thesis. AI can create material gains, but the size and direction depend on task fit, experience, use pattern and surrounding process. The evidence base became broader without becoming a universal coefficient.

EXHIBIT / THREE LATER FIELD RESULTS, THREE DIFFERENT OPERATING QUESTIONS

+15%	−2h	+19%
Customer support	**Email time**	**Completion time**
Issues resolved per hour; largest gains among less experienced agents.	Approximate weekly reduction among active users; no broad task-composition shift.	Experienced developers took longer in a small specialised RCT.
QJE [11]	NBER [12]	METR [13]

10 EXPECTATIONS VERSUS OUTCOMES

The early thesis held in some places and failed in others

The expectation that model capability would improve rapidly held, as later AI Index synthesis documented across multiple capability measures. The expectation that rapid capability would automatically create deep firm adoption did not. Official indicators through 2025 continued to show large gaps by firm size, sector and digital maturity. OECD analysis emphasised connectivity, data, skills and finance as complementary enablers for SMEs. [715]

The expectation that knowledge work would change held, but the pathway was not a clean substitution curve. Field evidence showed augmentation, compressed learning, personal time savings, no detected coordination change, and negative effects in a specialised setting. The distribution mattered more than the average.

The expectation that governance would need to become operational also held. Later standards and guidance converged on named accountability, impact assessment, testing, monitoring and response. The continuing gap was implementation evidence: organisations more often endorsed principles than demonstrated controls in operation. [1819]

Australia's position reinforced the diffusion problem. The Productivity Commission argued in 2024 that near-term opportunity sat downstream, in adapting and implementing general-purpose technology locally. Later ABS evidence showed reported AI use varied sharply by industry and innovation activity. [1620]

EXHIBIT / 2023 EXPECTATION AND 2026 RETROSPECTIVE JUDGMENT

Held	Failed	Mixed	Held
Capability would advance	**Access would equal transformation**	**Work would be broadly faster**	**Governance must operationalise**
Systems improved and useful task coverage expanded rapidly.	Use diffused faster than coordinated workflow and value evidence.	Effects varied materially by task, experience and configuration.	Later guidance converged on evidence-producing lifecycle controls.

11

INSTITUTE SYNTHESIS

The unit of advantage is the configuration

The revised Institute thesis is that model capability is an input; the value-producing unit is a configuration of task, people, model, context, controls and capacity allocation. Two organisations using the same model can produce different outcomes because they select different tasks, supply different context, distribute authority differently and act differently on saved capacity.

This framing preserves the jagged frontier but makes it operational. Task fit asks whether the system performs under representative conditions. Human fit asks who benefits, who is deskilled and who can detect failure. Workflow fit asks how inputs, hand-offs, review and demand change. Control fit asks whether consequence is bounded and visible. Economic fit asks whether realised outcomes exceed full cost.

The configuration should be versioned. A model update, new data source, changed prompt, expanded tool right, different worker population or new consequence can invalidate earlier evidence. Approval should therefore attach to a defined configuration and evidence window rather than to an abstract product name.

This is a stricter interpretation of the 2023 evidence, not a rejection of it. The early experiment remains valuable precisely because it showed both gains and a boundary. Later studies make the boundary more important, not less.

> **Revised Institute thesis — Do not ask whether AI works. Ask whether this human–AI configuration improves this outcome, for this population, under these controls, at full cost.**

EXHIBIT / THE SIX-PART CONFIGURATION

01	02	03
Task Representative work, exceptions, consequence and success criteria.	**People** Experience, learning, review capability and affected stakeholders.	**System** Model, prompts, data, retrieval, tools and integration version.
04	**05**	**06**
Workflow Inputs, hand-offs, demand, rework and saved-capacity destination.	**Control** Authority, testing, monitoring, incident response and recovery.	**Economics** Baseline, full cost, quality guardrail and realised outcome.

12 EXECUTIVE APPLICATION

Re-run the 2023 decision with a stronger protocol

Start with a named outcome and a representative work sample. Record the baseline distribution, including cycle time, quality, exceptions, rework and consequence—not only the mean. Predefine the minimum improvement, quality floor and stop conditions before teams see the result.

Test at least three configurations where practical: current work, assisted work and redesigned workflow. Segment results by experience and task type. Observe how people actually use the system, because a licence assignment does not specify the human–AI treatment.

Carry the result into operations with a time-limited authority boundary. Monitor the same outcome and quality measures, plus incidents, review burden and capacity destination. Reassess after model, data, prompt, integration, population or workflow change.

For portfolio decisions, distinguish explore, embed, expand, hold and retire. Exploration needs learning value. Embedding needs repeatable workflow evidence. Expansion needs control and outcome evidence. Holding is appropriate when uncertainty remains material. Retirement is a positive decision when economics or consequence do not justify further work.

13 BOARD AGENDA

Ten questions that survive the hype cycle

Ask which outcome is expected to change, which population and work are in scope, and what current performance distribution forms the baseline. Ask what evidence shows the task sits inside the tested frontier and what conditions place it outside.

Ask who benefits, who reviews, who can recognise failure and how learning pathways change. Ask what data, model, prompts, tools and authority define the approved configuration, and which changes trigger re-evaluation.

Ask where saved capacity goes, how full cost is measured and which quality or risk limits cannot be traded away. Ask what the strongest counterevidence says and whether the team tested the conditions under which the result may reverse.

Finally, ask what would cause the organisation to stop. A portfolio without hold and retire decisions is accumulating belief, not evidence. The enduring lesson from 2023 is that speed of capability raises the value of disciplined learning.

RESEARCH RECORD

Method and limitations

Method

This evidence retrospective reconstructs the public decision record for January–December 2023, then tests it against research published through 31 July 2026. Contemporaneous evidence includes field experiments, economic working papers and NIST AI RMF 1.0. Subsequent evidence includes representative adoption surveys, peer-reviewed and working-paper field studies, official statistics and later implementation guidance. We extracted the population, task, intervention, outcome and stated limitations for central claims; treated working papers and vendor-involved research according to their disclosed status; and avoided combining incomparable effect sizes. The Institute frameworks are analytical syntheses, not independently validated causal models. Source notes label whether evidence was available during the period or added through hindsight.

Limitations

This is not a systematic review or meta-analysis. The 2023 public record was shaped by English-language research, well-resourced organisations and rapidly changing systems; null results and operational failures may be under-published. Later studies use different tools, populations, tasks and outcome definitions, so they test the direction and boundary of the early thesis rather than reproducing one treatment. Retrospective analysis risks hindsight bias even with labelled evidence clocks. The report does not estimate an economy-wide productivity or employment effect and should not be read as legal, workforce or investment advice.

Evidence conventions

Source facts are cited to the original publication. Institute analysis reconciles definitions or combines evidence. Institute inference extends that analysis into a management proposition. Frameworks are decision aids, not claims of universal causality. Every important statistic should be read with its population, date, definition and caveat.

Research cutoff: 31 July 2026. Review cycle: six months, or earlier after a material change in evidence, standards or policy.

REFERENCES

Evidence behind the thesis

01 **Harvard Business School AI Institute, Navigating the Jagged Technological Frontier**

CONTEMPORANEOUS EVIDENCE
Contemporaneous summary of the 758-consultant experiment and its capability-boundary finding.

02 **Dell'Acqua et al., Navigating the Jagged Technological Frontier, HBS Working Paper 24-013**

CONTEMPORANEOUS EVIDENCE
Full preregistered experimental design, tasks, outcomes and limitations.

03 **NIST, Artificial Intelligence Risk Management Framework 1.0**

CONTEMPORANEOUS EVIDENCE
January 2023 voluntary framework organised around govern, map, measure and manage.

04 **Agrawal, Gans and Goldfarb, The Turing Transformation**

CONTEMPORANEOUS EVIDENCE
2023 working paper on automation, augmentation and skill-premium mechanisms.

05 **Bick, Blandin and Deming, The Rapid Adoption of Generative AI**

SUBSEQUENT EVIDENCE
Representative US surveys separating any use, weekly work use, daily use and work-hour exposure.

06 **ILO, Generative AI and Jobs: A Global Analysis of Potential Effects**

CONTEMPORANEOUS EVIDENCE
Early task-based occupational exposure analysis; exposure is not observed displacement.

07 **Stanford HAI, 2024 AI Index Report**

SUBSEQUENT EVIDENCE
Subsequent synthesis of capability, adoption, investment and policy evidence.

08 **OECD, Employment Outlook 2023: Artificial Intelligence and the Labour Market**

CONTEMPORANEOUS EVIDENCE
Contemporaneous evidence and analysis on AI, tasks, skills and job quality.

09 **NIST AI Resource Center**

SUBSEQUENT EVIDENCE
Subsequent operational resources for testing, evaluation, verification and validation.

10 **NIST, ARIA Pilot Evaluation Report**

SUBSEQUENT EVIDENCE
Subsequent evidence on evaluating AI systems in context rather than relying on abstract capability alone.

11 **Brynjolfsson, Li and Raymond, Generative AI at Work**

SUBSEQUENT EVIDENCE
Peer-reviewed customer-support field study covering 5,172 agents and heterogeneous effects.

12 **Dillon et al., Shifting Work Patterns with Generative AI**

SUBSEQUENT EVIDENCE
Randomised six-month experiment across 66 firms and 7,137 knowledge workers.

13 **METR, Early-2025 AI Experienced Open-Source Developer Study**

SUBSEQUENT EVIDENCE
Small specialised RCT reporting longer task completion with AI and strong participant misprediction.

14 **ILO, Generative AI and Jobs: A Refined Global Index of Occupational Exposure**

SUBSEQUENT EVIDENCE
2025 task-level exposure index combining worker assessments, experts and model predictions.

15 **OECD, AI Adoption by Small and Medium-Sized Enterprises**

SUBSEQUENT EVIDENCE
2025 analysis of diffusion gaps and complementary enablers for SME adoption.

16 **Australian Bureau of Statistics, Characteristics of Australian Business 2024–25**

SUBSEQUENT EVIDENCE
Official Australian statistics, including the explicit limitation that AI-use intensity was not measured.

17 **Stanford Digital Economy Lab, Canaries in the Coal Mine?**

SUBSEQUENT EVIDENCE
Administrative payroll evidence on early-career employment, with observational limitations.

18 **National AI Centre, Guidance for AI Adoption**

SUBSEQUENT EVIDENCE
Australian operating guidance covering accountability, impact, risk, transparency, testing and human control.

19 **NIST AI 600-1, Generative AI Profile**

SUBSEQUENT EVIDENCE
Subsequent generative-AI profile extending the AI RMF with technology-specific risks and actions.

20 **Australian Productivity Commission, Making the Most of the AI Opportunity**

SUBSEQUENT EVIDENCE
2024 Australian analysis emphasising downstream adaptation, implementation, skills and digital infrastructure.

THE AI INSTITUTE

Evidence should change what you do next.

Research · Readiness · Capability

theai.institute