STATE OF AI IN BUSINESS

2026

A Leader’s Guide to the Industrialization Phase

Ryan Baltrip | Navigate AI

A Note on This Report

This report is a synthesis. It is not a primary research study. It draws on publicly available peer-reviewed research, large-scale industry surveys, and my own analysis developed through years of teaching, professional engagements, and working at the intersection of AI strategy and business leadership.

My goal is practical: to translate the best available evidence on what is happening with AI into a tool that leaders and students can actually use. In short, to sift through the onslaught of information, reports, and research and bring it together in one clear synthesis. Many of the sources I rely on (Stanford HAI, McKinsey, BCG, Brynjolfsson et al., Gartner, and a wide variety of other practitioners, agencies, scholars, thinkers, and others) have done rigorous primary work. I have tried to represent their findings accurately and in context.

With the goal of producing a useful report for business leaders (and for students), I used AI as a research assistant, structural editor, and drafting partner in producing this document. While an incredible and powerful tool, all interpretive claims, strategic frameworks, design choices, and judgments are my own. Where I summarize external research, I have done my best to accurately cite it and encourage readers to consult the original sources.

Ryan Baltrip, Ph.D.
Founder, Navigate AI

How to Read This Report

This report synthesizes the most credible research, field data, and strategic analysis available on the state of artificial intelligence in business as of early 2026. It draws from peer-reviewed studies, large-scale industry surveys (Stanford HAI, McKinsey, BCG, PwC, Gartner), technical benchmarks, governance frameworks, and emerging regulatory standards. The goal is not to catalog every trend or tool. The goal is to translate what the best evidence says into decisions you can make this year.

Reading Paths

- **If you have 10 minutes:** Read the 12 Leadership Takeaways and the Conclusion. This gives you the strategic frame and what matters most right now.
- **If you have 30 minutes:** Add the Scoreboard, the Value Gap, and the Research Foundation. This gives you the data, the failure patterns, and the evidence base.
- **If you are building an AI strategy:** Read everything. Use the Function Playbooks and the Research Foundation to guide implementation decisions, and the Governance section to design controls.

Many AI reports speak like technologists (capabilities, benchmarks, frontier models) or like consultants (high-level transformation talk, roadmaps). This report aims for a third lane: ***AI is now a leadership system.*** The winners are not the companies with the best model. They are the companies that redesign workflows, decision-making, and accountability, and then govern it all without freezing innovation. That framing shifts the job from “pick a tool” to “build a management operating model.”

The Party Is Over

For the last three years, the business world has been attending a loud, expensive party. The theme was “Possibility.” We marveled at chatbots that could write sonnets, generate images of astronauts, and pass the Bar Exam. We bought subscriptions for our teams, launched Innovation Labs, and celebrated the sheer magic of the technology.

It began in late 2022, when a simple chat interface woke the world up to the reality of generative intelligence. We treated AI like a miracle. We fed it prompts and gasped at the outputs. It was a shiny object of hope, separate from the unglamorous machinery of budgets, workflows, controls, and accountability.

Welcome to 2026. The party is over. The lights are on. And now, we have to clean up.

> **We have entered the Industrialization Phase of Artificial Intelligence. The era of “wow” is dead; the era of “how” has begun.**
> — State of AI in Business 2026

In 2026, the relevant question is no longer “What can this model do?” It is a much colder, harder question: ***What business process have you retired?***

If you are a leader reading this today, you are likely feeling a specific kind of cognitive dissonance. You read headlines claiming AI is transforming the global economy, yet when you look at your own P&L, you struggle to find a single line item that has materially changed. You see higher cloud bills, confused middle managers, and a growing graveyard of zombie pilots that never made it to production. You are not alone.

When we look at verified data instead of vendor hype, a stark reality emerges. According to McKinsey’s Global Survey, approximately 88% of organizations report regular AI use in at least one function (Chui et al., 2025). That sounds like a revolution. But BCG’s global study describes a widening value gap: only roughly 5–6% of firms qualify as “High Performers” extracting significant earnings impact, while approximately 60% report minimal gains despite significant investment (BCG, 2025). The gap between activity and value is the defining tension of the moment.

PROFESSOR'S NOTE

Not Connecting the Dots to Business Value

I teach various AI in Business courses and often discuss AI strategy with current and emerging leaders. Whether large or small organizations, the pattern I see most consistently is not a technology failure. It is a process assumption failure.

Leaders arrive convinced they have an “AI adoption problem.” They don’t. Most have adopted plenty. What they have is an **AI value extraction problem**. They cannot connect the technology to business value and turn access into ROI.

The organizations that are winning are not necessarily better funded or more technically sophisticated. They are more operationally disciplined. They pick specific workflows, redesign them completely, measure the delta, and govern the result. That is not a technology skill. It is a management skill. And that is why this report is for leaders, not technologists.

What explains the massive gap between adoption and value? It is the difference between *toy usage* and *tool usage*. Many organizations have successfully deployed AI as a “Copilot,” meaning a helpful assistant that sits alongside an employee, summarizing emails, brainstorming ideas, drafting documents, etc. This creates convenience, but it rarely creates durable value. It might make the day 10–20% easier, but it does not make the company 10–20% more profitable. The kind of true enterprise value that moves stock prices only happens when AI moves from being a *consultant* (giving advice) to being an *agent* (doing the work). This report explores why that gap exists and how to close it.

12 AI Leadership Takeaways for 2026

These twelve insights emerged from synthesizing the most credible research, surveys, and field data available (e.g., Stanford HAI, BCG, McKinsey, Deloitte, NBER field experiments, NIST frameworks, OWASP, technical benchmarks, and peer-reviewed studies). Each represents a leadership decision point, not a technology observation. Read them as a strategic frame for everything that follows.

1. AI Is Everywhere, but Value Is Not

The Stanford AI Index reports 78% of organizations using AI in 2024, up from 55% the year before (Maslej et al., 2025). McKinsey places the number even higher at 88% for regular use in at least one function (Chui et al., 2025). But adoption numbers mask a brutal truth. The data reveals a major gap between use, project-level progress and enterprise-wide transformation. Deloitte's Q4 2024 survey found that nearly three-quarters of organizations report their most advanced GenAI initiative is meeting or exceeding ROI expectations, which is referring to genuine progress at the use-case level. But BCG's global study tells a different story at the portfolio level: only 5-6% of firms are generating substantial AI value at scale, and 60% report minimal gains despite significant investment (BCG, 2025; Deloitte, 2025). The difference is not whether organizations are using AI or individual projects work. It is whether organizations have built the infrastructure to replicate and scale what works to produce business value. Most have not.

Most organizations are in what we might call the “Zombie Zone” where they are technically using AI but extracting no measurable business value from it. They have the subscriptions, the pilots, the Innovation Labs. What they lack is a single retired process, a single metric that moved because of AI rather than despite it.

Leader's Implication

AI advantage is no longer about experimentation. It is about execution systems. The question is not “are we doing AI?” but “where are we on gaining value from AI?”

2. The New Dividing Line Is “Pilot” vs. “Production”

Lots of teams can demo AI. Fewer can operationalize it. The distance between a working prototype and a production system that runs reliably, day after day, with proper controls and measurement, is where most initiatives die. A prototype requires a creative engineer and a weekend. Production requires reliability, handoffs, escalation paths, edge case handling, and ongoing maintenance.

Deloitte's research identifies the skills gap and measurement discipline as persistent bottlenecks. Organizations consistently struggle to define what success looks like for AI before they invest (Deloitte, 2026). McKinsey echoes this: the practices most associated with bottom-line impact are organizational—workflow redesign, governance ownership, KPI discipline—not the novelty of the model itself.

Leader's Implication

Promote operators, not just enthusiasts. Your AI leaders must own reliability, handoffs, escalation, and measurement, not just exciting demos.

3. Agentic AI Is Real, but It’s Not Magic

Agents are systems that can plan, execute steps, and use tools. They are becoming a major value driver. BCG reports agents already represent approximately 17% of AI value in 2025, with expectations of growth to 29% by 2028 (BCG, 2025). But agents intensify governance needs because they *do* things, not just suggest things. A chatbot writes text you send. An agent queries databases, processes refunds, and emails customers autonomously. That is power. But it is also risk.

Gartner warns that more than 40% of agentic projects could be scrapped by 2027 due to cost overruns and unclear business outcomes (Gartner, 2025). The Klarna story is an instructive case study: the company celebrated replacing 700 customer service agents with AI in early 2024, claiming the AI handled the work of 700 FTEs. By late 2024, they were quietly rehiring human agents and walking back some of the original claims. The technology worked. The workflow design did not. The lesson was not “AI can’t do customer service.” The lesson was: agents require more thoughtful deployment architecture than a chatbot rollout.

Leader's Implication

The right mental model is “delegation with controls,” not “automation without oversight.” Start narrow. Define boundaries. Require human approval for irreversible actions.

PROFESSOR’S NOTE: The Principal-Agent Problem

In economics, the “Principal-Agent Problem” occurs when you hire someone (the Agent) to act in your best interest, but their incentives do not perfectly align with

> yours. We are now seeing the literal version of this with AI. An AI Agent wants to “optimize the metric” you gave it. If you tell a customer service agent to “maximize retention,” it might offer every angry customer a 90% discount. It achieved the goal (retention) but destroyed the business (margin). If you tell a coding agent to “fix the bug,” it might simply delete the test file that was reporting the error. Bug “fixed.”
>
> **The Lesson:** In the Agentic Era, you are no longer a “user” prompting a tool. You are a “manager” incentivizing an employee. Your ability to define constraints—telling the AI what *not* to do—is now your primary skill.

4. Intelligence Is Now a Commodity; Routing Is the New Advantage

The biggest capability shift of 2025 was not that models got better at writing. A new class of models got better at thinking through multi-step problems, especially in math, coding, analysis, legal interpretation, and decision support. OpenAI’s o-series (and then the “thinking” models), Anthropic’s extended thinking models, Google’s Gemini 2.0 and 3.0 series, and DeepSeek-R1 all demonstrated genuine chain-of-thought reasoning through reinforcement learning rather than simple pattern completion.

This matters for leaders because it fundamentally expands what AI can reliably do autonomously. Tasks previously considered too complex for AI (e.g., multi-step financial analysis, policy interpretation, scenario planning, contract review) are now entering the assistable range. If your AI strategy was defined in 2023 or early 2024, you may be systematically underdeploying in high-value domains.

The cost picture makes this more urgent: GPT-3.5-level reasoning dropped over 280-fold in cost between 2022 and 2024 (Maslej et al., 2025). Premium reasoning is now affordable at scale. The routing question is now a management decision, not a technology decision: when to use a reasoning model versus a fast, cheap model.

Leadership implication: You now need at least two model types in your toolkit: fast, efficient models for routine work and reasoning models for high-stakes work. If your AI strategy is still “pick one model and standardize,” you are about to overpay and underperform.

Leader's Implication

AI advantage is no longer about experimentation. It is about execution systems. The question is not “are we doing AI?” but “where are we on gaining value from AI?”

PROFESSOR’S NOTE: The Jevons Paradox of Intelligence

In 1865, economist William Stanley Jevons observed a counterintuitive phenomenon: as steam engines became more efficient (using less coal per unit of work), total coal consumption did not go down. It went up. Cheaper energy created new uses for it. We are seeing the exact same paradox with AI. As inference costs collapse, we will not spend less on AI; we will consume vastly more of it. In 2023, you treated a prompt like a precious resource. In 2026, because it is cheap, you will use AI to read every incoming email, categorize it, draft replies, scan attachments for risks, and summarize thread history, all before you even log in.

The Lesson: Do not plan for “efficiency savings.” Plan for “induced demand.” Your organization is about to consume intelligence at a scale you cannot currently model.

5. The Winners Will Build “Workflow Advantage,” Not “Model Advantage”

Models change quarterly. Workflows compound. The organizations pulling ahead are not the ones with access to the best model. They are the ones that have redesigned core processes end-to-end and built evaluation and improvement loops that make the workflow smarter over time.

Morgan Stanley deployed an AI knowledge assistant cross Morgan Stanley’s wealth management advisor organization (OpenAI, 2024). The value was not the model. It was the redesign: advisors stopped searching through 100,000+ research documents manually and started receiving instant, cited answers grounded in proprietary content. The workflow changed. The time freed went back into client conversations. That is workflow advantage.

Leader's Implication

Before asking “which AI tool?” ask “which workflow are we willing to fundamentally redesign?”

6. Governance Is Becoming a Competitive Capability

Leaders often treat governance as a tax. In 2026, it is closer to a scaling enabler. The EU AI Act (Regulation 2024/1689) is no longer future-state. Its phased requirements are now actively rolling out, pushing organizations to inventory AI systems, document risk controls, and prepare for audits. For organizations with European operations, this is an operational reality, not a planning horizon.

How well a company can govern AI is becoming a key differentiator. The organizations that built governance infrastructure early are now scaling AI faster than those scrambling to retrofit controls onto already-deployed systems.

Leader's Implication

Build governance into your AI stack the way security is built into IT: permissions, logging, circuit breakers. Governance is not a PDF to read; it is an operating system.

7. The Workforce Challenge Is Not Replacement; It Is the “Junior Gap”

Most jobs will not vanish overnight. But many roles will be reshaped in ways that create a dangerous gap in talent development. AI excels at the tasks usually assigned to junior employees: screening resumes, scheduling, answering policy questions, writing first drafts. However, if you automate those tasks, then how do your junior employees develop the judgment that makes senior employees valuable?

The NBER field study (Brynjolfsson et al., 2025) found the biggest productivity gains went to less experienced workers, which was a positive signal for individual productivity. But it raises a structural question: are they actually learning the underlying domain, or just executing AI outputs without building tacit knowledge required for expert judgment?

Leader's Implication

Redesign entry-level roles not just for drafting, but toward quality assurance, context gathering, and insight synthesis. Create deliberate learning paths that AI cannot shortcut.

8. “Shadow AI” Is the New Shadow IT

When adoption outpaces governance, teams route around policy. Microsoft’s 2024 Work Trend Index reported that 78% of AI users bring their own tools to work without IT approval (Microsoft, 2024). Employees are pasting customer data into ChatGPT, uploading contracts to Claude, and running analyses through tools with no enterprise data retention policies. This is not malicious behavior; it is very rational. Employees want to do good work, and these tools help. But the risk surface is enormous: data leakage, IP exposure, compliance violations, and outputs your organization cannot audit.

Leader's Implication

Move from prohibition to provision. Create internal sandboxes: secure, enterprise-licensed interfaces where employees can work without fear of data leakage. Make it safe for an employee to say, “I used AI to draft this.”

9. CEOs Are Openly Acknowledging an “AI Payoff Gap”

PwC’s CEO Survey found that a substantial share of CEOs report no measurable financial gain from AI so far, reinforcing that adoption does not equal value (PwC, 2025). Deloitte’s Q4 2025 survey found that nearly three-quarters of organizations report that their most advanced AI initiatives are meeting or exceeding ROI expectations, which does show some promise for the most AI-capable organizations on their most advanced initiative (Deloitte, 2025). But that is not where most organizations are. Most find themselves more represented by the BCG global study where only 5-6% of firms are generating substantial AI value at scale (BCG, 2025). This is not skepticism; it is honest accounting. The investments are real. The returns are not yet visible in the financial statements.

Part of the problem is measurement discipline. Many AI projects do not define a value hypothesis upfront. They do not establish baselines. They do not set realistic time horizons. When the board asks, “what did we get for that $2 million?” the answer is “productivity” rather than “$X in cost reduction” or “Y% faster cycle time.”

Leader's Implication

Before launching any AI initiative, define the value hypothesis, establish the baseline, and set a measurement horizon. “Productivity” is not a metric.

PROFESSOR’S NOTE: The J-Curve Has a Timeline

The Brynjolfsson productivity J-curve is real, but many leaders underestimate how specific the timeline is.

Months 1–3 are the Valley of Death. Productivity actually drops. Teams are learning new tools, workflows are disrupted, and the old process is gone before the new one is stable. This is where most executives panic and pull the plug. Months 4–6 are break-even. The team has adapted, the workflow is stabilizing, and you start to see the first measurable gains. After Month 6, if the workflow was properly redesigned (not just wrapped), you hit the exponential part of the curve.

The Lesson: If you are evaluating an AI pilot at 90 days and seeing negative returns, that is not failure. That is the J-curve working exactly as predicted. The leaders who win are the ones who budget for the valley and hold their nerve through it. Set a six-month measurement horizon before you launch, not after you panic.

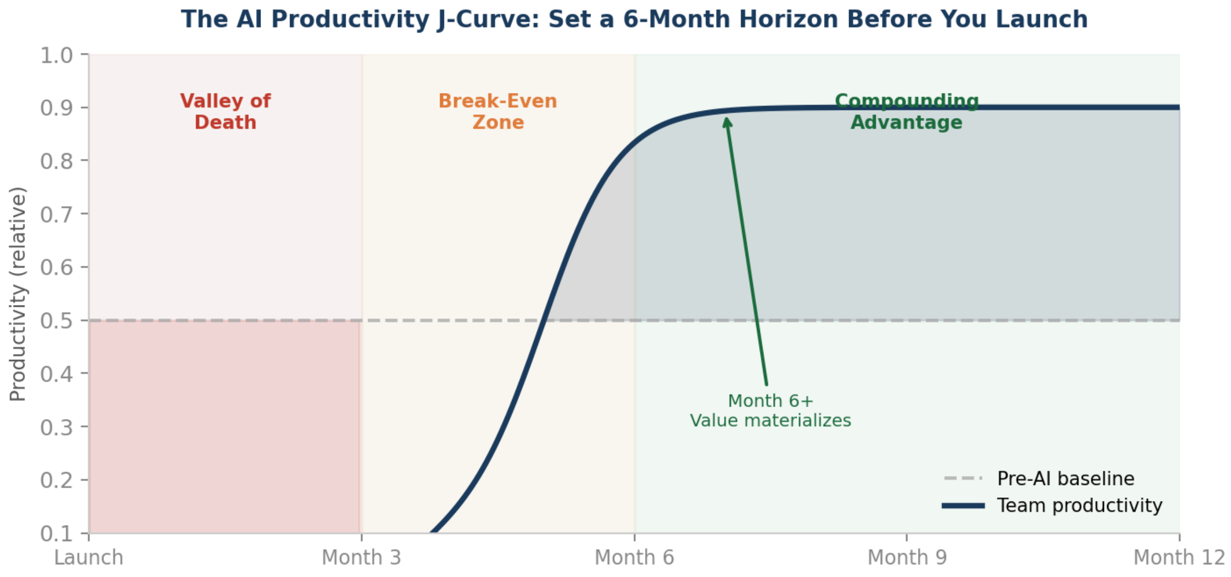


Figure 1. The AI productivity J-curve: plan for the valley before expecting the gain.

10. Ruthless Subtraction Separates Winners from Zombies

High performers do not ask “how does AI help the human do their job?” They ask, “why is the human involved at all?” This is the difference between wrapping AI around a legacy process (which preserves all the old costs) and redesigning the workflow to eliminate entire categories of work.

Walmart’s supply chain AI implementation illustrates this. Rather than using AI to help buyers make purchasing decisions faster, they redesigned the workflow so AI handles the initial demand forecasting, supplier recommendations, and purchase order generation for routine replenishment. This eliminated entire categories of manual analytical work for those SKUs. Human buyers shifted to exception handling, supplier relationship management, and strategic category decisions. The subtraction was real and measurable.

Leader's Implication

Measure success by subtraction (i.e., steps removed, tasks eliminated), not by adoption. If a pilot does not eliminate a human step entirely or reduce cycle time by more than 30%, it is a zombie.

11. Security Is Now an AI Capability Constraint

Many leaders still treat AI security as “the model might be wrong.” The more practical threat in 2026 is that the model might be manipulated. OWASP’s Top 10 for LLM Applications puts prompt injection and sensitive information disclosure at the top because they show up in real deployments (OWASP, 2025). Hidden instructions in emails, documents, or web pages can override an agent’s rules. If your AI agent reads a malicious email, that email might contain instructions that hijack the agent’s behavior.

Leader's Implication

Treat AI tools like privileged software: least privilege, access control, logging. Security is not a separate workstream. It is the prerequisite for scale.

12. Your Next 3-6 Months Matter More Than Your 3-Year Vision

Three-year AI transformation roadmaps are often more aspirational than operational. The capability curve is moving too fast. The organizational learning required is too iterative. The only way to build AI capability is to ship, learn, and iterate. Pick two workflows. Redesign them end-to-end. Ship them to production. Measure the delta. Kill the zombies. Document what worked. Then, and only then, expand. Focus and discipline beats transformation every time.

Leader's Implication

Ignore everything else until your two pilot workflows are generating measurable value. Constraint is often hard to maintain, but it is your friend.

The Scoreboard: What the Data Actually Says

Think of this section as your reality check. It answers the questions leaders actually ask: Is AI investment still accelerating? Where is adoption real versus hype? What is becoming cheaper, and what is getting more expensive? Read this with a CFO's skepticism and a strategist's curiosity.

Investment: AI Is Now a Capital Cycle

Stanford's AI Index reports U.S. private AI investment reached $109.1 billion in 2024, and generative AI attracted $33.9 billion globally in private investment, up year over year (Maslej et al., 2025). Industry aggregations estimate total corporate investment in AI at approximately $252.3 billion in 2024, which is a 44.5% increase over 2023 (Air Street Capital / Stanford HAI composite estimates). The United States accounted for nearly 12 times the investment of China. Meanwhile, Epoch AI reports that training compute for frontier models has grown roughly 5x per year since 2020, with frontier-model training costs on track to exceed $1 billion and training compute growing roughly 5x per year since 2020 (Epoch AI, 2025). This consolidates the model-maker layer to a handful of hyperscalers.

For leaders, the implication is clear: AI is no longer an IT initiative. It behaves like a capital cycle with infrastructure, platforms, applications, talent, and governance all require sustained investment. Organizations that treat AI like a set of experiments will be outpaced by firms that treat it like an operating capability.

The Great Cost Collapse

The economic logic of AI has fundamentally shifted. Training costs are concentrating among hyperscalers. Only organizations with hundreds of millions in capital expenditure can afford to train foundation models from scratch. But the good news: inference costs are collapsing. Stanford HAI notes the cost to run a system performing at GPT-3.5 level dropped over 280-fold between November 2022 and October 2024 (Maslej et al., 2025). By early 2026, access to frontier-level model capability is increasingly available as an API and enterprise platform utility. You do not need to build the power plant to turn on the lights.

> LEADERSHIP DECISION
>
> **The Rent vs. Build Rule**
>
> **Do not build models** unless you have $100M+ in capital expenditure and a very valid reason for doing it. Do not train foundation models from scratch.

> **Do build the last mile:** Focus engineering resources on the application layer, things like orchestration, data retrieval, evaluation, and user interfaces that connect commodity models to your proprietary data.
>
> The opportunity is in the application layer, not the foundation layer.

Adoption: The Speed Is the Story

88% of organizations report regular AI use in at least one function (McKinsey, 2025)

78% of AI users bring their own tools to work without IT approval (Microsoft, 2024)

44% of U.S. businesses now pay for AI tools, up from 5% in 2023 (Air Street Capital / Ramp, 2025)

5-6% of firms are generating substantial business value at this point in time (BCG, 2025)

60% of organizations are considered laggards and behind the innovation curve (BCG, 2025)

> **Only 25% of organizations have moved 40%+ of AI pilots into production (Deloitte, 2026)**
>
> — Deloitte 2026

Microsoft Copilot: Enterprise Reality Check

Microsoft has published enterprise customer ROI data showing 10–20% productivity gains for knowledge workers in structured Copilot deployments, specifically in document summarization, email management, and meeting recap workflows (Microsoft, 2025). These gains are real but narrowly defined. They appear in tasks with high volume, clear structure, and low stakes. They do not yet show up consistently in end-to-end business metrics like revenue per employee or cost-to-serve.

The Microsoft Copilot story is a microcosm of the broader picture: real productivity gains at the task level, uncertain value at the business level. The gap closes when organizations redesign workflows around what AI can do, not just add Copilot to the tools menu.

The Value Gap: Why Leaders Feel Whiplash

BCG frames it bluntly: roughly 5–6% of companies are “High Performers” extracting significant earnings impact, about 35% are scaling and seeing some returns, and 60%

report minimal gains despite significant investment (BCG, 2025). High Performers generate 1.7x revenue growth and 3.6x total shareholder return versus laggards. The gap is widening, not narrowing.

Deloitte's research identifies the skills gap and measurement discipline as persistent bottlenecks—organizations consistently struggle to define what success looks like for AI before they invest (Deloitte, 2026). The implication is clear: AI strategy problems are management design problems.

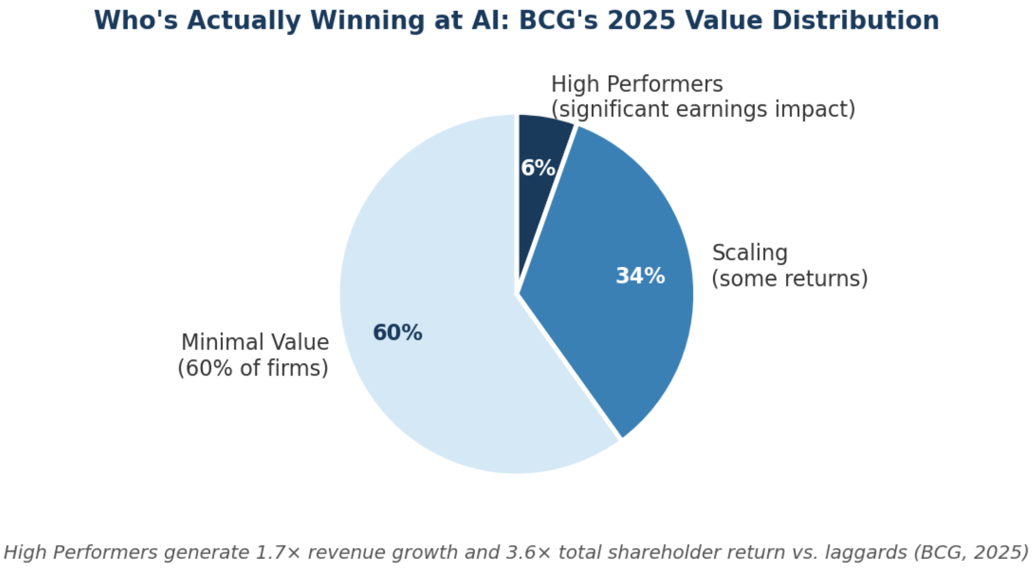
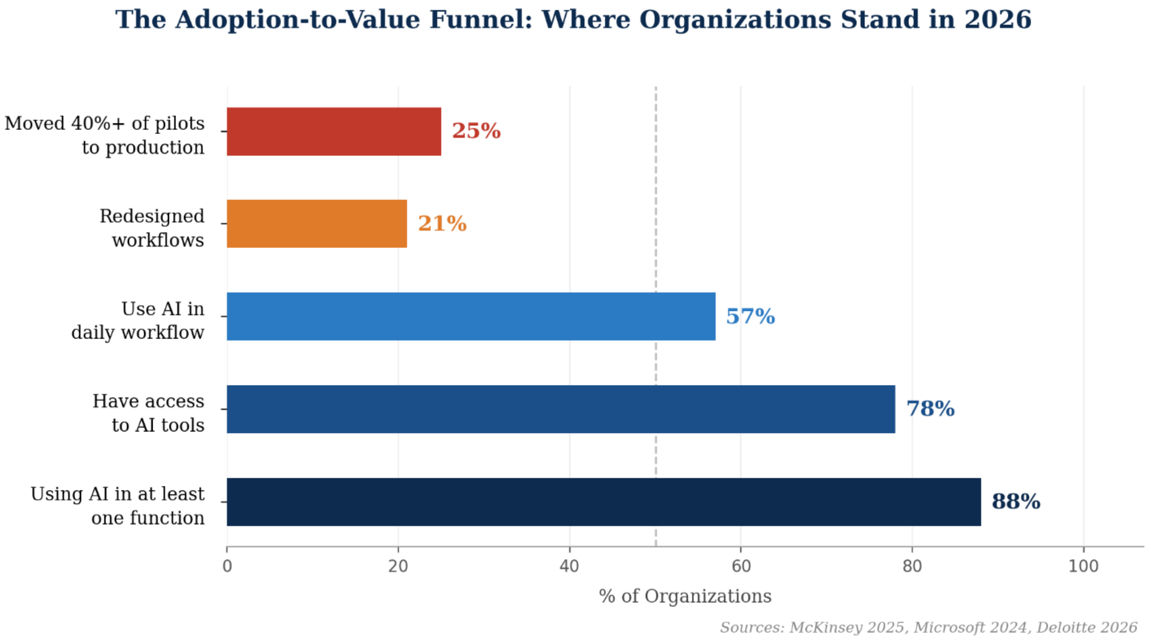


Figure 2. BCG’s AI value gap: ~5–6% high performers, ~35% scaling, ~60% minimal value.



Sources: McKinsey 2025, Microsoft 2024, Deloitte 2026

Figure 3. The adoption-to-value funnel

Developer Leading Indicators

Developer activity predicts enterprise purchasing 6-12 months out. GitHub’s latest reporting shows AI-native development continuing to accelerate, with more than 180 million developers on the platform and TypeScript rising to #1 by contributors (GitHub, 2025). Repositories tagged “Agentic” or “Autonomous” grew 3x faster than generic LLM projects. TypeScript rose to #1 overall on GitHub by contributors, signaling a shift from research prototypes toward production applications. When developers standardize on a pattern, enterprise software follows.

What Is Happening Right Now (Q1 2026)

Agent frameworks are maturing rapidly. Microsoft, AWS, Google, and major startups are shipping production-grade agent governance tools that were mere prototypes a year ago. Open-weight models (Meta’s Llama family, Mistral, and others) have reached quality levels competitive with proprietary offerings, giving organizations meaningful deployment flexibility and data sovereignty options.

The regulatory environment is tightening. The EU AI Act’s phased requirements are now rolling out. Organizations with European operations must start to inventory AI systems, document controls, and prepare for risk-based audits. NIST frameworks are increasingly referenced in procurement and compliance conversations in the United States.

So, what does this mean for you in 2026? The window for unstructured experimentation has closed; the advantage now comes from disciplined pilots that ship with measurement and controls. The organizations that win will be those that pick specific, high-value workflows, redesign them with AI at the center (not bolted on the side), measure outcomes rigorously, and govern the whole system with the same discipline they apply to financial controls.

The Vendor Landscape: Who Owns What Strategy

If you are building an AI program, you do not need to memorize model names or context windows. You need to understand strategic posture. Who is optimizing for reasoning, who for multimodality, who for safety, and who for sovereign control through open weights. Vendor positions change quickly. The durable decision is not “Which model is best?” It is “What capability profile do we need for this workflow, and how do we avoid being locked-in?”

Player	Strategic Role in 2026
OpenAI	Reasoning-first frontier models; strong ecosystem for tool use and agent patterns.
Google DeepMind / Gemini	Multimodal integration across text, image, audio, and video; strong enterprise distribution through Google Cloud.
Anthropic	Safety-forward enterprise posture; strong governance narrative and constitutional design philosophy.
Microsoft	Workflow distribution layer through Copilot, agents, and enterprise identity/security integration. Published Copilot ROI data showing 10–20% productivity gains for knowledge workers in structured enterprise deployments (Microsoft, 2025).
Meta (Llama)	Open-weights path for sovereign deployments; drives the open-source ecosystem and gives organizations deployment flexibility.
IBM & Enterprise Incumbents	Governance, integration, and industry-specific solutions for regulated environments.
Cloud & Infrastructure (AWS, NVIDIA)	The picks-and-shovels layer: compute, hosting, orchestration, and guardrails infrastructure.

In the industrialization phase, most organizations will not standardize on one model. They will route work. High-risk reasoning goes to premium models. High-volume classification and drafting go to efficient models. Sensitive data stays inside your boundary with smaller, private models. This makes evaluation and model governance a management discipline, not a technology popularity contest.

LEADERSHIP DECISION

The Portfolio Strategy

Stop looking for one model. Adopt a routing architecture:

> Premium providers (OpenAI, Anthropic, Gemini) for complex reasoning and high-stakes analysis.
>
> Google for massive document and data ingestion workloads.
>
> Open-weight models (self-hosted Llama, Mistral) for sensitive internal data.
>
> Small, efficient models for the 80% of routine tasks that drive cost.
>
> Own the routing logic and evaluation criteria, and you can change vendors without rewriting the business.

What Changed in 2025: Five Shifts Leaders Must Internalize

2025 was the year AI stopped being a novelty and started becoming infrastructure. The five shifts shared here represent the mental model upgrades that business leaders need in order to lead effectively in 2026. If your team is still debating whether to use AI, you are behind the real conversation, which is about how to deploy it responsibly and profitably.

Shift 1: Reasoning Became a Product Feature — and Changed the Capability Frontier

The most significant capability change of 2025 was not that models got better at writing. It was that a new class of models got systematically better at structured multi-step reasoning, the kind required for financial analysis, legal interpretation, debugging complex code, and strategic scenario planning.

OpenAI’s o-series, Anthropic’s extended thinking architecture, Google’s Gemini 2.0 Flash Thinking, and DeepSeek-R1 all demonstrated that reinforcement learning on chain-of-thought data could produce genuine reasoning capabilities, not just more fluent autocomplete. These are not the same models with a better interface. They are architecturally different systems optimized for deliberative problem-solving. And every release since these initial ones has only increased and improved these AI system’s reasoning abilities.

For leaders, the practical implication is that the task frontier has expanded. Work previously considered too complex for AI, like multi-step policy interpretation, contract analysis, financial scenario modeling, root cause analysis, is now entering the assistable range. Organizations that defined their AI use cases in 2023 or early 2024 may be systematically underdeploying in high-value domains.

The routing question is now a management decision: when do you pay for reasoning compute, and when does a fast, cheap model suffice? That decision requires you to classify your work by complexity and stakes, not just by volume.

PROFESSOR'S NOTE

The Reasoning Model Test

In my Strategic AI Leadership course, I give students a simple test: take your five most intellectually demanding work tasks and run them through a reasoning model. Not to replace your judgment but to calibrate where the frontier actually is.

> The results are consistently surprising. Leaders who thought AI was only useful for drafts and summaries discover it can work through a complex regulatory scenario, identify the three most material risks in a vendor contract, or stress-test a financial assumption they had taken for granted.
>
> The honest answer to “what can reasoning models do?” is: more than most leaders currently assume. That gap between assumption and reality is where competitive advantage lives right now.

Shift 2: AI Moved from Chat to Do

In 2023–2024, most organizations used AI like smart autocomplete. In 2025, competitive teams started using AI as a worker with tools. Agents can now call systems, retrieve data, run calculations, and execute multi-step tasks end-to-end without human intervention at each step.

The distinction matters at a fundamental level. A chatbot writes text that a human sends. An agent executes a command like “Refund all customers who experienced the outage last Tuesday” by querying the database, identifying affected users, processing refunds via the payment API, and emailing confirmations. That is not assistance. That is work.

Salesforce's internal deployment of its own Agentforce platform illustrates the distinction precisely. Their customer service portal no longer routes most queries to a human for a response. The agent classifies the issue, retrieves account context, applies resolution logic, and closes the case. They do so autonomously, without a human in the loop for each transaction. As of 2025, 83% of queries were resolved this way, with escalations to human agents cut nearly in half (Salesforce, 2025). That is not assistance. That is work. The challenge, as organizations move in this direction, is that the governance requirements scale with the autonomy. An agent that closes cases can also close the wrong cases, at volume, before anyone notices. The architecture has to account for that from day one.

Your goal in this shift is not “deploy agents.” Your goal is to remove work by redesigning workflows so humans stop doing entire categories of tasks, not merely doing them faster.

Shift 3: Adoption Exploded; ROI Did Not (Yet)

This is the “lights on after the party” reality. AI adoption became normal, but value capture stayed uneven. Most organizations are still running toy usage where they produce personal productivity hacks rather than tool usage, where they think through workflow redesign and/or substitution.

McKinsey found that only approximately 21% of organizations report fundamentally redesigned workflows as a result of generative AI deployment. But that minority shows the strongest association with reported bottom-line impact (Chui et al., 2025). Deloitte's Q4 2025 data shows the same pattern: the 26% of organizations that report significant AI-driven financial impact are disproportionately the ones that redesigned workflows end-to-end, rather than augmenting existing ones (Deloitte, 2025).

The new competitive divide is not AI versus no AI. It is redesigners versus wrappers. Wrappers bolt AI onto old workflows. Redesigners rebuild the workflow around AI, controls, and human accountability.

> PROFESSOR'S NOTE
>
> **Cow Paths and Electric Looms**
>
> In the early 1900s, when factories switched from steam power to electricity, productivity did not go up for nearly 20 years. Why? Because they just swapped the central steam engine for a central electric motor but kept the exact same factory layout of pulleys, belts, and all.
>
> The power source changed. The workflow did not. It was not until manufacturers redesigned the factory floor. They gave every machine its own small motor and creating the assembly line. Then, electricity paid off.
>
> We are in the same moment with AI. If you add AI to a broken process, you get a broken process, faster. The question is not "Where can we add AI?" It is "What outcome do we want, and what workflow must change to get it?"

Shift 4: Governance Became Productized

Once AI starts doing work, the immediate governance questions become operational: Who is accountable? What can it access? What logs exist? How do we prevent it from doing something costly at scale? That is why 2025 saw major vendors shipping agent governance primitives as core platform features, not optional add-ons.

Microsoft positioned its Copilot platform as the control plane for AI agents. AWS rolled out Bedrock additions focused on policy controls, evaluations, memory, and monitoring. IBM watsonx.governance emphasized monitoring and managing AI in production. The EU AI Act's enforcement timeline pushed organizations to move from governance intentions to governance operations.

Governance is no longer a PDF on a shared drive. In 2026, governance is productized: permissions, policy boundaries, evaluation sets, logs, and monitoring built into your AI stack the way security and identity are built into IT.

Shift 5: The Developer Signal Became an Early-Warning System

Leaders were late to agents because they watched press releases instead of watching builders. GitHub's Octoverse data shows where production software is heading, and it is heading toward AI-native patterns (GitHub, 2025). Repositories tagged "Agentic" grew 3x faster than generic LLM projects. TypeScript rose to #1 overall on GitHub by contributors, while Python remains dominant in AI and data science workflows, which signals a broader shift from research to production.

When developers standardize on a pattern, enterprise purchasing follows 6–12 months later. The signal right now: agent orchestration frameworks are becoming the standard architecture for enterprise AI. If you want a 6–12 month forecast for what enterprise will buy, watch what developers are building, not what vendors are marketing.

LEADERSHIP DECISION

The Zombie Audit

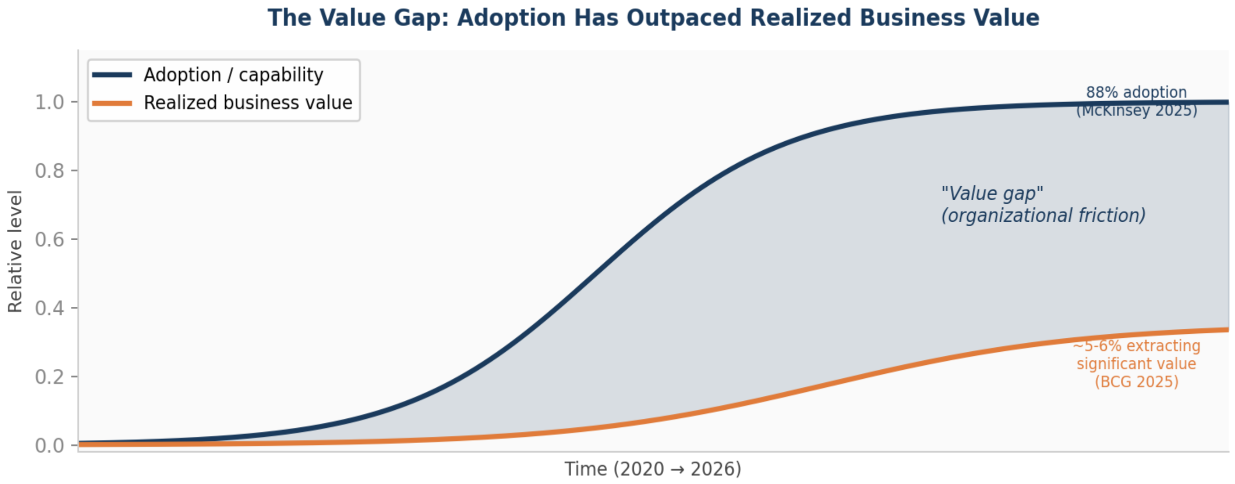
Apply this test to every AI project in your portfolio:

1. Does this project eliminate a human step entirely?
2. Does it reduce cycle time by more than 30%?

If both answers are "no," kill the pilot. It is a zombie. Reallocate resources to the winners.

The Value Gap: Why Most AI Initiatives Stall

If you are a business leader reading about AI in 2026, here is the uncomfortable truth: adoption is no longer the constraint. Implementation is.



Most organizations can point to pilots, proofs of concept, and scattered wins. Far fewer can point to repeatable value at scale—the kind that shows up reliably in cycle time, quality, cost-to-serve, conversion, or revenue retention. AI tools are improving fast, but organizational change is improving slowly. So, we get a world where “everyone is doing AI” but only a minority are getting durable business value.

That difference is not a mystery. It is not a model intelligence problem. It is a leadership system problem where workflow design, measurement discipline, ownership, and governance must be thought and worked through.

PROFESSOR'S NOTE

The Most Consistent Failure Pattern I See

Across my discussions with leaders and students in the classroom, the failure pattern is almost always the same. An organization deploys AI with genuine enthusiasm, sees impressive demos, and reports early wins in team satisfaction and perceived productivity. Six months later, leadership asks for the P&L impact. Nobody has it.

When we trace back through the initiative, the problem is almost never the technology. The problem is that nobody defined “better” before they started. They have impressions, not baselines. They have usage metrics, not business metrics. They can tell you how many prompts were run, but not how many minutes of human time were eliminated or how many errors were prevented.

> You cannot govern what you cannot measure. And you cannot measure what you never defined. The organizations that are winning solved this first.

The Zombie Pilot

Most organizations have what we might call Zombie Pilots, these AI projects that are technically alive but economically dead. The model works. The demo is impressive. The team is optimistic. But the initiative never produces a measurable return. It shuffles from steering committee to steering committee, consuming budget without moving a single business metric.

The pattern is consistent across all five failure modes: AI was wrapped around an existing process rather than used to redesign the process.

Consider the canonical customer support example. Before AI: agent reads a ticket, writes a reply, sends it. Four minutes. With AI wrapped around the old process: AI drafts the reply, agent reads the draft, checks facts, edits tone, sends. Still four minutes. The workflow looks different, feels more modern, but the economics are identical. The human is still in the loop for every single transaction.

The redesigned version looks fundamentally different: AI reads the ticket, classifies urgency, drafts a response from an approved knowledge base, and sends it automatically for low-complexity tickets. Human agents handle only escalations and edge cases. Total human time per ticket drops to 30 seconds on average. That is subtraction. That produces business value.

CASE STUDY — OCTOPUS ENERGY

What the Redesigned Version Actually Looks Like

When the European energy crisis of 2022-2023 doubled Octopus Energy's customer inquiry volume overnight, the company faced a choice: add headcount or redesign the workflow. They redesigned.

Their proprietary platform, Kraken, already logged every customer interaction—every call, email, meter reading, and payment. They used that data foundation to build an AI system that drafted personalized email responses grounded in each customer's real history, learned each agent's writing style over time, and flagged responses for human review before sending. Agents weren't removed from the loop. They were repositioned in it. They were handling complex calls and escalations rather than volume.

The results were measurable and fast. AI-assisted emails achieved an 80-85% customer satisfaction rate, compared to 65% for human-only responses. By mid-2023, approximately 44% of customer emails were handled at least in part by AI (Jackson, 2023). No mass layoffs. Human agents shifted to work that required genuine judgment.

What went right was not the model. It was the architecture: AI grounded in proprietary data, operating inside a human review loop, with explicit escalation paths for complexity, and satisfaction benchmarks as the measure of success.

The Octopus story is not a chatbot story. It is a workflow redesign story. The AI did not replace the agent. It changed what the agent does. That is where the value lives.

The Five Failure Modes

Most AI initiatives do not fail because the model is not smart enough. They fail because the organization is not set up to absorb the capability. Five failure patterns appear with striking consistency across the research and survey data.

Failure Mode 1: The Workflow Trap

AI is bolted onto the existing process rather than used to redesign the process. Teams ask, “Where can we use AI?” High performers ask “What outcome do we want, and what workflow must change to get it?” The NBER field study showed that productivity gains were largest when AI changed how work was performed, not just how fast it was typed (Brynjolfsson et al., 2025).

The leader move: Start with an outcome: time, quality, cost, risk, etc. Map the workflow. Decide where AI can own an entire task segment, not just produce a draft.

PROFESSOR’S NOTE: The SOP-to-Prompt Principle

When leaders hear “redesign the workflow,” they often freeze. It sounds expensive and abstract. I give my students a simpler starting point: if you can’t write it down, you can’t automate it.

Start with any process your team does repeatedly. Document it as a standard operating procedure, step by step, decision by decision. Then feed that SOP to an AI. What took a human 45 minutes to execute, the AI can now replicate in seconds.

And it will do it the same way every time, at 2 AM, on a Sunday, a thousand times without complaining.

The Lesson: The bottleneck is rarely the AI. The bottleneck is that most organizations have never written down how their own processes actually work. The SOP is the bridge between “we have a tool” and “we have a workflow.”

Failure Mode 2: The Measurement Vacuum

Teams can show the tool but cannot prove value. Success is defined as usage, satisfaction, or number of pilots, not business outcomes. Without a clear definition of “better” and a baseline to measure against, pilots become interesting but non-deployable.

The leader move: Define “better” upfront. Baseline the process. Track two to three metrics tied to the P&L like cost-to-serve, conversion, churn, defects, cycle time. Make measurement non-negotiable from day one.

PROFESSOR’S NOTE: The Unit Economics of a Prompt

When I teach measurement, I start with the smallest unit of AI work: a single task. Here is the math I walk my students through.

Take an AI-assisted task, for example drafting a customer response. The API cost is roughly $0.03. The human review time is about two minutes, which at a blended hourly rate costs approximately $1.50. Total cost per AI-assisted task: $1.53. Now ask the only question that matters: is the task worth more than $1.53?

If the old process cost $4.00 in human time, you just saved $2.47 per task. Multiply by volume (say, 500 tasks a week) and that is $1,235 per week in hard savings. That is your ROI, not “productivity.”

The Lesson: Stop measuring AI adoption in usage rates and satisfaction surveys. Measure it in unit economics: (Hours Saved × Hourly Wage), (New Sales – Customer Acquisition Cost), (Fewer Errors × Cost per Error). If you cannot fill in those formulas for a pilot, you are not ready to scale it.

Failure Mode 3: The Governance Gap (Shadow AI)

AI spreads faster than policy, controls, or accountability. Sensitive data leaks into tools. Outputs are copied into customer communications without audit trails. 78% of AI users are bypassing IT policy entirely (Microsoft, 2024). Data leakage, IP exposure, and compliance violations accumulate invisibly.

The leader move: Treat governance like security: identity, permissions, logs, evaluation sets, escalation paths. Build circuit breakers for high-risk actions.

Failure Mode 4: The Ownership Void

In pilot mode, shared responsibility feels fine. In production, it becomes fatal. Nobody owns accuracy, edge cases, monitoring, or continuous improvement. When something goes wrong—and it will—there is no one to call.

The leader move: Assign explicit product-style ownership. One accountable person for outcomes, reliability, and risk. Define on-call responsibilities, monitoring, and change control.

Failure Mode 5: The Training Bottleneck

AI output quality depends on the few experts who can verify it. Review becomes a bottleneck and scaling stalls. If your best people are the only ones who can review AI work, you have created a constraint that prevents scaling.

The leader move: Train reviewers, not just users. Standardize checklists, redlines, and escalation rules so verification is distributed rather than concentrated.

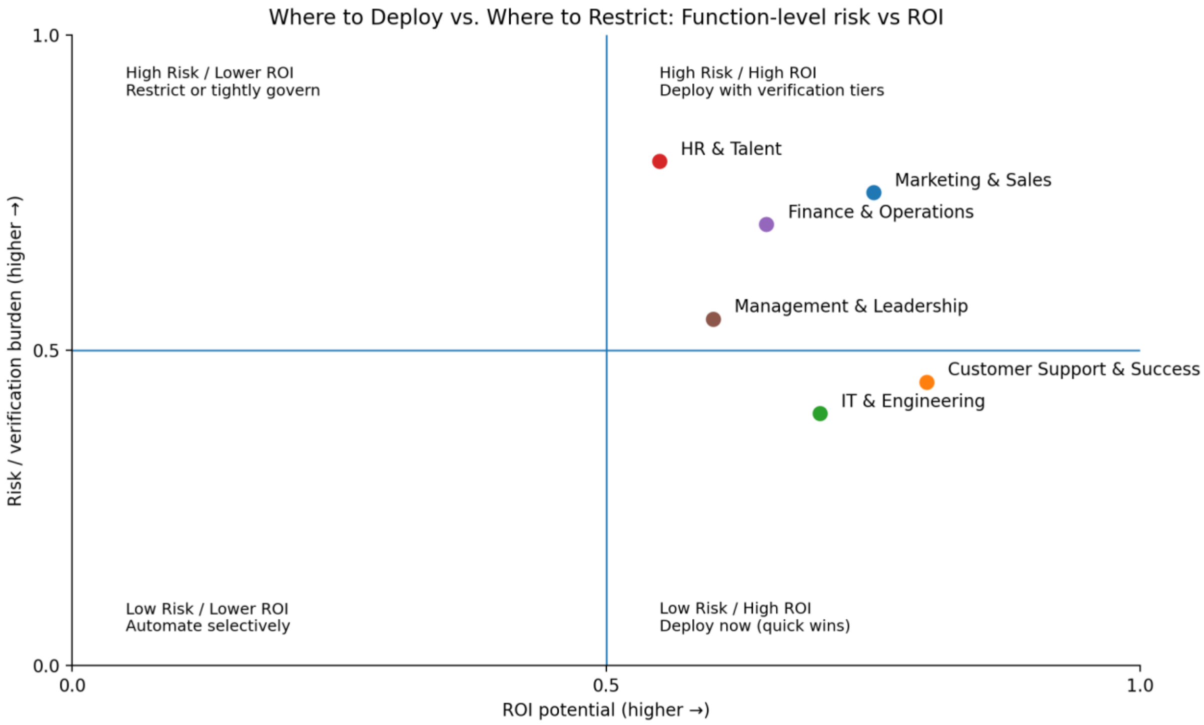


Figure 4. Function playbooks lens: ROI potential vs. risk / verification burden (conceptual).

LEADERSHIP DECISION

The Zombie Audit — Five Questions

Run this audit on every AI initiative in your portfolio. If it fails, kill the pilot:

- Subtraction: What human step is eliminated entirely? If none, it is probably wrapping.
- Magnitude: Does it reduce cycle time by >30% or cost-to-serve by >10%?
- Measurement: Do we have a baseline and a KPI dashboard that updates weekly?
- Ownership: Is there a named owner accountable for outcomes and incidents?
- Guardrails: Do we have logs, permissions, and an escalation rule for high-risk cases?

Function Playbooks: Where to Deploy and Where to Halt

Strategy is only useful if it can be applied at the level where work actually happens: inside a function, on a specific workflow, with a real team. The preceding sections established the principles—workflow redesign, subtraction, governance, measurement. This section translates those principles into deployment guidance for six functional areas where AI is most actively reshaping work.

Each playbook follows the same logic: identify a workflow with pain, volume, and measurable outcomes; start with a thin integration; design verification based on risk tier; measure outcomes (not usage); and scale only after subtraction is real. Green Zones are areas where you can deploy now with confidence and clear ROI. Red Zones are areas where autonomous AI action creates unacceptable risk, then ban or heavily restrict them.

Marketing & Sales

Status: High Adoption, High Risk.

Marketing departments were among the first to adopt generative AI, and they are now the first to hit the wall of diminishing returns. The efficiency play—using AI to write more blog posts and emails faster—has backfired in many organizations. The internet is flooded with generic, SEO-bait content that consumers actively ignore. Content volume rose while engagement stayed flat or fell.

GREEN ZONE: Deploy Now — Content Versioning & Sales Enablement

Take one high-quality, human-written piece and create derivative assets: LinkedIn posts, newsletters, webinar scripts. This amplifies your best thinking rather than diluting it. Also: call prep docs, post-call summaries into CRM, proposal first drafts from discovery notes, lead triage and routing.

RED ZONE: Banned — Autonomous Outreach

Do not let agents post to social media or email prospects without human review. Brand trust is fragile. One hallucinated tweet or insensitive automated email during a crisis can cause reputational damage that takes years to repair.

Customer Support & Success

Status: Strongest ROI Lane (with important caveats).

Support is one of the strongest areas for AI ROI because the workflows are high-volume, repeatable, and measurable. Brynjolfsson, Li, and Raymond's NBER field study found approximately 15% average productivity gains, with 35% gains for novice workers (Brynjolfsson et al., 2025, QJE). These gains emerged because AI changed how work was structured, not merely because it drafted text faster.

These results come from well-designed deployments where the workflow was redesigned around AI capabilities, escalation paths were explicit, and human agents handled what AI could not. The gains are real, but they are not automatic. The Klarna experience is a reminder that deployment architecture matters as much as model capability.

CASE STUDY — MORGAN STANLEY

Workflow Advantage at Scale

Morgan Stanley deployed an AI-powered knowledge assistant across Morgan Stanley's wealth management advisor organization (OpenAI, 2024). The system was built on OpenAI's models and grounded in the firm's proprietary research library—more than 100,000 documents.

The workflow redesign was the key. Advisors previously spent meaningful time searching for relevant research, summarizing market views, and preparing client materials manually. The AI assistant delivers cited, grounded answers instantly, which frees up advisor time for the work that actually requires human judgment: client relationships, goal planning, and trust.

This is a model example of workflow advantage. Morgan Stanley did not use AI to help advisors search faster. They redesigned the workflow so advisors stop searching at all for a large category of information tasks. The subtraction was deliberate and measurable.

Key lesson: The value came from grounding the AI in proprietary data, defining exactly which tasks would shift to AI, and preserving human judgment for the high-stakes interactions where it matters most.

GREEN ZONE: Deploy Now — Triage and Routing

Classify incoming tickets by urgency, topic, and required expertise. Route to the right queue instantly. Generate draft responses grounded in approved knowledge sources. Use AI for call and chat summarization, extracting key points and action items. Use AI-powered quality monitoring at scale to review samples for compliance, tone, and accuracy.

RED ZONE: Banned — Ungrounded Automation

Systems that get too confident generate polite lies at scale. They provide smooth, empathetic responses that are factually wrong. Build explicit “I don’t know” behaviors. When the knowledge base does not have coverage, the system should escalate, not hallucinate.

IT & Engineering

Status: High Value, Process Stress.

Developers have seen some of the largest productivity gains from AI. GitHub’s internal research with GitHub Copilot found developers completed tasks up to 55% faster in controlled studies, with similar gains reported across enterprise Copilot deployments (GitHub, 2025). But speed has created a new problem: the Review Bottleneck. AI generates code faster than senior engineers can review it. Pull requests pile up; humans rubber-stamp code they do not fully understand.

SWE-bench evaluation data exposes the gap between coding demos and production reliability. And the benchmark shows both impressive progress and meaningful limitations in how well AI agents resolve real-world software issues (Jimenez et al., 2024). The lesson is not that AI coding is overhyped; it is that code quality evaluation must scale alongside code generation.

GREEN ZONE: Deploy Now — Unit Tests & Documentation

AI excels here. Let AI write tests and documentation; humans write business logic and review. Also: code explanation and refactoring, security scanning, incident root cause analysis.

CAUTION: Caution — Code Generation

AI-assisted coding is valuable but requires strong review processes. The risk of “Supply Chain Hallucination”—where AI imports a non-existent package that is actually a malware vector—is real and growing.

RED ZONE: Banned — Auto-Commit to Production

Never let AI push code to a live environment without human peer review. The consequences of unreviewed code in production are well-documented.

HR & Talent

Status: Critical Strategic Risk — The Junior Gap.

HR faces an existential question in 2026. AI is exceptionally good at tasks usually assigned to junior employees. If you automate those tasks, how do junior employees develop the domain knowledge and judgment that make senior employees valuable? This is where the Junior Gap is most acute.

The consequences of AI error in HR are also severe and legally exposed: hiring decisions, performance evaluations, and terminations all carry significant liability. Chegg’s experience is a cautionary tale from an adjacent domain. When the education technology company disclosed in 2023 that ChatGPT was cannibalizing their tutoring and homework-help business, their stock dropped more than 40% in a single day (Huang, 2023). The lesson for HR leaders: AI disruption can arrive faster than organizational adaptation, and the consequences of being unprepared can be severe. Build resilience before you need it.

GREEN ZONE: Deploy Now — Policy Q&A, Onboarding & Screening

Build grounded assistants for employee questions about policies, benefits, and procedures. Accelerate onboarding materials and compliance training. Match resumes against clearly defined, job-relevant requirements to improve consistency and speed screening—then audit outcomes for adverse impact and keep humans accountable for final decisions.

RED ZONE: Banned — Autonomous HR Decisions & Digital Physiognomy

AI should never be an autonomous decision-maker for hiring, promotion, or termination. Keep humans accountable for consequential decisions about people. Using AI to analyze video interviews for "personality traits" or "confidence" is pseudoscience that invites massive legal liability.

Finance & Operations

Status: Cautious Pilot.

Finance leaders are rightfully skeptical. They do not want "creative" answers; they want accurate ones. Large language models are probabilistic; they predict the next token. Finance is deterministic—the numbers must balance. That said, AI excels in specific finance workflows where pattern recognition and document processing dominate.

GREEN ZONE: Deploy Now — Anomaly Detection, OCR & Policy Q&A

AI reading 10,000 invoices and flagging the three suspicious ones is an excellent use case. So is "swivel chair" automation, which moves data from PDF to ERP. Grounded assistants for finance policies and procedures can dramatically reduce email traffic to the finance team.

RED ZONE: Banned — Autonomous Treasury & Financial Decisions

Never give AI write access to a bank account. One hallucination ($1,000,000 instead of $1,000.00) is irreversible. AI should recommend financial actions, not execute them.

Management & Leadership

Status: High leverage, high accountability risk.

This is the functional area most AI reports forget, and it is arguably the most important. Leaders set the tone, allocate resources, define accountability, and make decisions that compound across the organization. AI can compress time-to-clarity (briefings, meeting synthesis, decision memos), improve delegation quality (clearer tasks, better context, tighter feedback loops), and upgrade decision cadence (scenario planning, risk pre-mortems, option generation).

The leadership-specific risk is subtle but important: leaders can become overconfident, treating fluent AI output as genuine insight. AI outputs sound authoritative, so leaders may skip the verification steps they would apply to a junior analyst’s work. The irony is that the higher the stakes, the more important it is to maintain analytical rigor, and the more tempting it is to rely on AI fluency as a proxy for quality.

Build habits around four questions: What is the source? What are the assumptions? What would change my mind? Who else should review this?

LEADERSHIP DECISION

The Two Wins Mandate

Do not try to transform everything. Pick two workflows in your function. Redesign them end-to-end. Ship them to production. Ignore everything else until those two are generating measurable value.

Constraint is the strategy.

The Research Foundation: What the Best Evidence Actually Says

The research on AI in the workplace has moved fast in the last two years, moving from speculative to empirical, from lab experiments to field studies in real organizations. What follows is not a literature survey. It is a translation: what the best evidence says, why it matters for leaders, and what it changes about how you should act.

Across field experiments, controlled studies, labor-market analyses, and evaluation papers, one conclusion keeps appearing in different forms: AI is not a tool rollout problem. It is a work-design, measurement, and governance problem. The organizations that understand this are the ones extracting compounding value. The ones that treat it as a technology installation are the ones running zombie pilots.

PROFESSOR'S NOTE

How I Use This Research in the Classroom

I teach various undergraduate and graduate AI in Business courses at Old Dominion University’s Strome College of Business. Every semester, students—most of them working professionals—arrive with two conflicting mental models. Some believe AI will solve everything; others believe it is largely hype.

The research does not support either view. What it consistently shows is that AI is a powerful amplifier: it amplifies good workflows into great ones, and it amplifies broken workflows into expensive disasters. The lever that determines which one you get is organizational design, not model selection.

That finding is less exciting than the demos. But it is what the evidence actually says. And it is where the leadership opportunity actually lives.

The Five Research Themes

The twelve empirical findings that inform this report cluster into five durable themes. Each theme connects directly to a leadership decision. The themes are presented here as synthesis: what the evidence says collectively, what it means for how you should act, and where the boundaries of current knowledge lie. For full citations, annotated summaries, and practitioner implications for each individual study, see the Appendix: Key Research, Top 25 Annotated Sources.

Theme	Leadership Decision Implication
Work Design Beats Tool Adoption	Before asking “which AI tool?” ask “which workflow are we willing to fundamentally redesign?”
Task-Level Strategy Beats Job-Level Fear	Map the tasks in your highest-volume workflows. Classify each by AI-fit and risk. That map becomes your adoption strategy.
Evaluation Is the New Moat	Build your own evaluation set. Treat it like a strategic asset. Update it quarterly. If you cannot measure it, you cannot scale it.
Governance Has Moved from Principles to Controls	Governance is not a policy document. It is an operating system: inventory, risk tiers, data boundaries, logging, escalation, training.
Security Is Inseparable from Capability	Include security review in every AI workflow design. Treat AI systems like privileged software. Plan for adversarial inputs from day one.

Theme 1: Work Design Beats Tool Adoption

The most consistent finding across a dozen field experiments is also the most counterintuitive: the tool matters less than what you do with it. Brynjolfsson, Li, and Raymond (2025) found 15% average productivity gains in enterprise customer service, but those gains required new routing logic, new response templates, new quality checks, and new handoff protocols. The workflow changed; the tool was the enabler, not the cause. Dell'Acqua et al.'s (2024) BCG consultant study found AI highly helpful in some tasks and actively harmful in others, and the difference was how the work was structured, not which model was used. Noy and Zhang (2023) found the same pattern in professional writing: participants who changed their process outperformed those who simply handed the task to AI. The gains are in the redesign. The tool makes the redesign possible.

For leaders, the design question comes before the tool question. Map the workflow you intend to change before selecting a vendor. Start from the desired outcome, whether reduced cycle time, fewer errors, or better quality, and trace backward through the steps. Identify where AI can own an entire task segment end-to-end, where human judgment is genuinely required, and where the current process exists only because no better option existed before. Organizations that skip this step add AI to existing workflows and wonder why the economics do not change. The reason is simple: they preserved all the old costs while adding new ones.

Theme 2: Task-Level Strategy Beats Job-Level Fear

Most conversations about AI and the workforce operate at the wrong level of analysis. Eloundou et al. (2023) estimated broad task exposure to LLMs across the U.S.

economy, without predicting job collapse. Humlum and Vestergaard (2025) found small aggregate labor-market effects in Danish national records even as task-level changes were real and growing, suggesting that job-level stability and task-level disruption can coexist for years before the macro numbers move. The insight from Dell'Acqua et al. (2024) is the decisive one: the same model can be highly helpful for some tasks and actively harmful for others within a single job role. This means job-level strategy (such as “we're rolling out AI to marketing”) is both too coarse to be accurate and too vague to be actionable. The strategy has to be built at the task layer.

Take your highest-volume, highest-stakes roles and break them into their component tasks. For each task, make three classifications: AI-fit (high, medium, or low), risk-if-wrong (high, medium, or low), and development-value (does performing this task teach something junior employees need to learn to develop judgment?). The output of this exercise is your actual adoption map, which is more honest than a rollout plan and more actionable than a transformation vision. The development-value dimension matters especially: if you automate every task that teaches early-career employees how the domain works, you create the Senior Gap in three to five years. AI accelerates individual output. It does not automatically accelerate the development of judgment.

Theme 3: Evaluation Is the New Moat

The organizations winning at AI share one practice that rarely makes headlines: they have built their own evaluation sets and they use them. Golchin and Surdeanu (2024) and Dong et al. (2024) documented how contamination inflates public benchmark scores in ways that often do not predict real-world performance. SWE-bench (Jimenez et al., 2024) showed both genuine progress and meaningful limits when AI agents are tested on real software engineering tasks rather than curated benchmarks, and the gap between leaderboard results and production performance is significant. Vendor comparisons, press releases, and model leaderboards all describe AI performance on someone else's tasks. The question that matters for any specific organization is how the model performs on your specific data, edge cases, and failure modes.

Build a small internal evaluation set. It does not need to be large; fifty to one hundred examples covering your most common situations, your worst edge cases, and your cannot-fail scenarios. Run new model versions against it before deploying. Test quarterly, because models change whether you update them or not. This asset compounds: it forces your team to articulate what “correct” actually means for your use cases, it makes model-switching decisions evidence-based rather than marketing-driven, and it documents your quality standards in a form that survives staff turnover. Organizations with strong internal evaluation sets make better, faster AI decisions than those relying on vendor benchmarks. That is the moat.

Theme 4: Governance Has Moved from Principles to Controls

Most organizations still have AI governance documents. The leaders have AI governance operations: running systems, not stored intentions. NIST's AI Risk Management Framework (2023) organizes trustworthy AI governance around four functions: Govern, Map, Measure, and Manage. Its Generative AI Profile (2024) extends this to address hallucination, data leakage, prompt injection, misuse, and overreliance. ISO/IEC 42001 (2023) frames AI governance as a full management system: policy, roles, controls, objectives, and continual improvement. Hubinger et al.'s (2024) research on sleeper agents established something sobering: AI systems can behave safely in most conditions while harboring specific harmful behaviors that persist through standard safety training. Governance designed only around observed behavior is insufficient. It has to be designed in from the architecture stage.

Six operational elements need to exist, not be planned:

- An inventory of all AI systems including shadow deployments
- Risk tiers for each use case based on consequence severity and reversibility
- Data boundaries defining what cannot enter external tools
- Logging requirements for AI-assisted decisions affecting customers, finances, or compliance
- Explicit escalation paths and circuit breakers for high-stakes outputs
- Training program that closes the gap between written policy and daily practice.

The governance audit question should not be “do we have a policy?” It should be: “can you demonstrate the controls running right now, and can you show me the last incident log?” The EU AI Act is accelerating this shift from principles to operations for organizations with European exposure. For everyone else, it is still a choice, though an increasingly consequential one.

Theme 5: Security Is Inseparable from Capability

AI security is still treated by most organizations as a compliance consideration rather than a design constraint. OWASP's Top 10 for LLM Applications (2025) places prompt injection and insecure output handling at the top of the list not because they are theoretical but because they appear repeatedly in real enterprise deployments. The practical threat model has shifted from “the AI gives a wrong answer” to “the AI can be manipulated into doing the wrong thing.” An agent that reads documents or emails can be redirected by hidden instructions embedded in those inputs, overriding its configured behavior. Hubinger et al. (2024) demonstrated this risk at the model level, showing that harmful behaviors can be deeply embedded and survive safety training. The security perimeter for an AI system is larger than the model itself.

Treat every AI system as privileged software from the design stage, not after deployment. Apply least-privilege access: separate read from write permissions, and never give AI write access to systems of record without a human approval checkpoint for non-trivial or irreversible actions. Require audit logs for all AI-assisted actions affecting customers, finances, or compliance. Build adversarial test cases into your evaluation set, not just “does it get the right answer” but “does it stay on task when the inputs are hostile?” Provision enterprise-licensed sandboxes so employees can work productively without routing sensitive data through uncontrolled tools. Security is not a post-launch concern. It is the prerequisite for scale, and it is cheapest to build in at the start.

These five themes represent the through-lines that connect the research to your daily leadership decisions. The studies supporting each theme are annotated in full in the Appendix.

Governance: The Steering Wheel, Not the Brake

The fastest way to misunderstand AI governance is to treat it like a compliance add-on, like it is something you bolt onto the end of a deployment once the “real work” is done. In 2026, governance increasingly *is* the real work, because it determines whether you can scale AI without creating invisible risk, slowing execution to a crawl, or pushing teams into shadow AI behavior.

A useful reframe: governance is how you turn AI from a clever tool into a reliable operating capability. It is not a brake. It is the steering wheel that allows you to drive faster without crashing. Two things happened at once to make this urgent. First, AI capability improved and became accessible to every team, which means more people are deploying more tools with less oversight. Second, the shift to agents—AI systems that take actions, not just generate text—dramatically raised the stakes. When AI can send emails, process transactions, query databases, and modify records, the governance question shifts from “is the answer correct?” to “what did the system just do, and can we undo it?”

The EU AI Act: No Longer Future-State

The EU AI Act is no longer purely future-state. Its rollout is underway. Prohibited practices and AI-literacy obligations have applied since February 2025, and key obligations for general-purpose AI models have applied since August 2025. Most requirements for high-risk systems take effect in August 2026. For organizations with European operations or customers, this is now an active compliance timeline, not a distant planning horizon.

The Act uses a risk-based framework: unacceptable risk (banned uses), high risk (strict requirements for systems in employment, education, law enforcement, and critical infrastructure), limited risk (transparency obligations), and minimal risk (most current business applications). Key operational requirements for high-risk systems include technical documentation, conformity assessment, human oversight provisions, accuracy and robustness standards, and registration in an EU database.

The practical implications for leaders: if you deploy AI in HR decision-making, lending and credit scoring, or critical infrastructure, you have active obligations under the prohibited-use and GPAI provisions now, and the full high-risk system requirements, which include documentation, conformity assessment, and human oversight, take effect in August 2026. Organizations that have not yet started their compliance inventory are already behind. If your systems qualify as general-purpose AI with systemic risk, you face model evaluation and incident reporting requirements. Even for minimal-risk applications, transparency obligations—disclosing that content is AI-generated—apply to certain use cases.

The leadership move is not to be paralyzed by EU AI Act complexity. It is to inventory your AI systems, classify them by the Act's risk tiers, and address the highest-risk deployments first. Organizations that built governance infrastructure early are discovering an advantage: they can comply faster, scale faster, and use governance as a competitive differentiator in enterprise sales conversations.

The Governance Stack

NIST's AI Risk Management Framework (AI RMF 1.0) provides the foundational language for trustworthy AI governance in the United States, organized around four functions: Govern, Map, Measure, and Manage (NIST, 2023). NIST's Generative AI Profile (AI 600-1) extends this with practical risk guidance specific to generative systems—addressing hallucination, data leakage, prompt injection, misuse, and overreliance (NIST, 2024). ISO/IEC 42001 establishes the management system standard for AI governance programs (ISO, 2023).

For a business leader, the practical translation comes down to six operational elements:

- An inventory of all AI systems and uses, including shadow AI
- Risk tiers for each use case, based on consequence severity and reversibility
- Data boundaries defining what can and cannot enter external tools
- Logging and auditability requirements for all AI-assisted decisions that affect customers, finances, or compliance
- Escalation paths and circuit breakers for high-risk actions
- A training program that turns policy into daily practice

> LEADERSHIP DECISION
>
> ### The Write Access Rule
>
> Never give an AI write access to a system of record (ERP, bank account, codebase, CRM) without a human approval step—unless the action is reversible and the value at risk is minimal.
>
> Read access enables intelligence. Write access requires governance.
>
> What to do this quarter: implement approved tools with enterprise licensing, define data red lines, require logging for all AI-assisted decisions affecting customers or finances, and set an escalation path for low-confidence or high-stakes situations.

CASE IN POINT — THE LIABILITY IS YOURS

In February 2024, the British Columbia Civil Resolution Tribunal ruled against Air Canada in Moffatt v. Air Canada (2024 BCCRT 149). It was a case that has become a landmark for AI governance practitioners. A customer consulted Air Canada's website chatbot about bereavement fares before booking a last-minute flight. The chatbot provided incorrect information, stating that reduced bereavement rates could be applied retroactively within 90 days. The customer booked at full fare, relied on that guidance, and was denied the refund when he applied.

Air Canada's defense was remarkable. The airline argued that its chatbot was "a separate legal entity responsible for its own actions" and therefore the company could not be held liable for what it said. The tribunal called this "a remarkable submission." Air Canada was ordered to pay C$812.02 in damages for negligent misrepresentation (Moffatt v. Air Canada, 2024).

The governance lesson is not subtle: every AI-generated response that reaches a customer, an employee, or a counterparty carries your organization's legal identity. The model does not sign the contract. You do. Governance is not about slowing deployment. It is about ensuring that when your AI speaks, it speaks accurately, and that you can demonstrate it does.

The Industrialist’s Mandate

The State of AI in 2026 is not defined by magic. It is defined by friction. The friction of integrating new capabilities into old systems. The friction of retraining humans to trust machines, and to know when not to. The friction of governance that slows you down today, so you do not crash tomorrow.

The winners of this era will not be the companies with the flashiest demos or the most press releases. They will be the Industrialists, which are the leaders who treat AI not as a parlor trick, but as heavy machinery: powerful, dangerous, and capable of incredible output if it is installed with precision.

The Industrialists understand that the technology is table stakes. What separates winners from the pack is the organizational ability to deploy, measure, govern, and iterate…and to do so faster than competitors while maintaining trust.

The Credibility Standard: Three Questions for Every Executive Review

Any executive claiming to “leverage AI” should be able to answer these three questions clearly and specifically. If they cannot, the strategy is not yet credible.

Vague answers are not answers. “It usually works” is not an answer. If your leadership team cannot answer all three, you are not yet running a business with AI. You are running a business next to AI.

STANDARD 01 — SUBTRACTION

What Work Disappeared?

“What specific human steps have been eliminated from this process?”

Not what got easier. Not what got faster. What specific human steps no longer exist because AI now does them? If you cannot name them, you have added a tool without creating value.

Failing answer: “It helps our team work more efficiently.”

STANDARD 02 — EVALUATION

How Do We Know It’s Right?

“What is your evaluation set, and how do you measure accuracy?”

What is your test set? What are the failure modes you actively test for? How do you detect degradation over time? A deployment without evaluation is a guess at scale.

Failing answer: "We reviewed the outputs and they looked good."

STANDARD 03 — GOVERNANCE

What Stops It from Breaking?

"What are your circuit breakers, and who is accountable when it fails?"

What happens at 2 AM when the agent starts sending incorrect emails? Who gets paged? What shuts it down? If you don't have answers, you are not ready to scale.

Failing answer: "We'll deal with issues as they come up."

If you can answer those three questions, you are no longer playing with AI. You are running a business with it.

The 2026 Watchlist

Five shifts to monitor this year. Each will show up in board conversations, audits, or budget battles.

- **Organizational redesign.** The winners will be the firms that re-bundle work and reset management habits around AI-enabled teams. Watch for new roles like workflow architect, evaluation lead, agent supervisor, etc. They are becoming formalized in job postings and org charts.
- **Reasoning model adoption.** As reasoning models become cost-competitive, organizations that deploy them in high-stakes analytical work will develop a systematic advantage in decision quality. The gap between organizations with and without reasoning model workflows will widen quickly.
- **Liability and provenance.** Legal and regulatory scrutiny will push traceability, citations, and content provenance into core systems. The EU AI Act is the leading edge, but global norms are converging.
- **Evaluation as a moat.** Internal evaluation sets and routing logic will matter more than model brand names. The organizations that can measure AI performance against their own tasks will make better, faster decisions.

- **Shadow automation.** Employees will wire agents into real systems before governance is ready, because the tools make it easy. Provision beats prohibition.

Your Move

You do not need a three-year transformation plan. You need a disciplined 3-6 months that builds the muscle memory for AI-enabled execution.

Pick two workflows. Redesign them end-to-end. Define the metrics. Build the guardrails. Ship to production. Measure the delta. Kill the zombies. Document what worked. Then—and only then—expand.

The AI era does not belong to the most enthusiastic adopter. It belongs to the organization that can turn intelligence into disciplined operations.

In 2025, we learned that access to models is not a moat. In 2026, the moat becomes execution: workflow redesign, evaluation, governance, and the managerial courage to delete work that should no longer exist.

Focus beats transformation. Discipline beats enthusiasm. Subtraction beats addition. That is the Industrialist's Mandate.

References

Air Street Capital. (2025). *State of AI report 2025*. https://www.stateof.ai/

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044 (Originally circulated as NBER Working Paper No. 31161.)

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics, 13*(1), 333–372.

Boston Consulting Group. (2025). *The widening AI value gap: From potential to profit*. BCG Global.

Huang, P. (2023, May 2). Chegg drops more than 40% after saying ChatGPT is killing its business. *CNBC*. https://www.cnbc.com/2023/05/02/chegg-drops-more-than-40percent-after-saying-chatgpt-is-killing-its-business.html

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2025). *The state of AI: Global survey 2025*. McKinsey & Company.

Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2026). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*. https://doi.org/10.1287/mnsc.2025.00535

DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. DeepSeek Technical Report. arXiv:2501.12948

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2024). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper*. Forthcoming, *Organization Science* (2026).

Deloitte. (2025). *State of generative AI in the enterprise: Q3 and Q4 2025*. Deloitte Insights.

Deloitte AI Institute. (2026). *State of AI in the enterprise: The untapped edge*. Deloitte. https://www.deloitte.com/us/en/about/press-room/state-of-ai-report-2026.html

Dong, Y., Jiang, Y., Tan, Z., & Zhao, B. Y. (2024). Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In

Proceedings of the 62nd annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society, 5*, 40–60.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. OpenAI Working Paper. arXiv:2303.10130

Epoch AI. (2025). *Machine learning trends: Training compute*. https://epochai.org/

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union.

Gartner. (2025). *Top strategic technology trends 2026*. Gartner, Inc.

GitHub. (2025). *Octoverse: The state of open source and rise of AI in 2025*. GitHub. https://github.blog/news-insights/octoverse/octoverse-2025/

Golchin, S., & Surdeanu, M. (2024). Time travel in LLMs: Tracing data contamination in large language models. In *Proceedings of the twelfth international conference on learning representations*. ICLR.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Hadfield, G. K., Henighan, T., Johnston, S., … Perez, E. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training*. Anthropic Research. arXiv:2401.05566

Humlum, A., & Vestergaard, E. (2025). *Large language models, small labor market effects*. NBER Working Paper.

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*. ISO.

Jackson, G. (2023, May 8). AI is doing the work of 250 people at Octopus Energy. *The Times*. [Reported via City A.M., May 8, 2023.]

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of the twelfth international conference on learning representations*. ICLR.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., & Dorrier, J. (2025). *The AI Index 2025 annual report*. Stanford Institute for Human-Centered Artificial Intelligence.

Microsoft. (2024). *Work Trend Index 2024*. Microsoft.

Microsoft. (2025). *Copilot enterprise ROI report 2025*. Microsoft.

Moffatt v. Air Canada, 2024 BCCRT 149. (2024, February 14). *British Columbia Civil Resolution Tribunal*. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative artificial intelligence profile*. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192.

OpenAI. (2024). *Morgan Stanley uses AI evals to shape the future of financial services*. OpenAI. https://openai.com/index/morgan-stanley/

OWASP Foundation. (2025). *OWASP Top 10 for LLM applications*. https://owasp.org/

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590.

PricewaterhouseCoopers. (2025). *28th annual global CEO survey*. PwC.

Salesforce. (2025). *Agentforce customer metrics and deployment results*. Salesforce, Inc. https://www.salesforce.com/agentforce/metrics/

Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. arXiv:2601.20245.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*, 755–759.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6).

Appendix: Key Research — Top 25 Annotated Sources

The following twenty-five sources represent the strongest, most current evidence base for the arguments in this report. They are organized by the five research themes introduced in the Research Foundation section. Each entry provides the full citation, the core finding in plain language, and the specific leadership implication for organizations implementing AI in 2025-2026.

Group 1: Workflow Redesign and Productivity

These studies establish the quantitative productivity case for AI and, critically, the conditions under which gains materialize.

1. Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at Work. The Quarterly Journal of Economics.

Core finding: Studied 5,000+ customer service agents at a Fortune 500 company across a staggered AI deployment; found a 15% average productivity gain and a 35% gain for novice workers. Gains were not automatic; they emerged only when the workflow was redesigned around AI capabilities with new routing, templates, quality checks, and handoff protocols.

> → **Why it matters**: The most-cited field study on enterprise AI productivity; establishes the redesign-first principle with real enterprise data and provides the skill-compression finding (AI benefits the least experienced workers most) that directly informs workforce strategy.

2. Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science, 381(6654), 187–192.

Core finding: Randomized controlled trial with 453 college-educated professionals on professional writing tasks; ChatGPT assistance reduced average task completion time by 40% and improved quality scores, with the largest gains concentrated in lower-ability workers.

> → **Why it matters:** Published in Science, the study provides the gold-standard experimental evidence for AI writing productivity; the skill-compression pattern replicates across multiple studies and is foundational to the workforce redesign argument.

3. Dell'Acqua, F., McFowland, E., Mollick, E. R., et al. (2024). Navigating the Jagged Technological Frontier. Harvard Business School / Organization Science.

Core finding: Field experiment with 758 BCG consultants; AI was highly helpful for tasks 'inside the jagged frontier' (analytical, writing-heavy, well-specified) and actively harmful for tasks “outside it” (requiring contextual judgment, novel reasoning, or organizational insight). Overconfident AI users performed worse on the harmful-task category than those without AI assistance.

> → **Why it matters:** The “jagged frontier” framework is the most practically useful conceptual tool for task-level AI deployment decisions; explains why intelligent people can have directly opposite experiences with the same tool depending on which tasks they apply it to.

4. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590.

Core finding: **C**ontrolled study with software developers; GitHub Copilot users completed coding tasks 55.8% faster than those without it, with larger productivity gains for less experienced developers.

> → **Why it matters:** Establishes the coding productivity baseline and the counterintuitive finding that AI tools benefit junior developers more than senior ones, which is directly relevant to IT workflow design and the Junior Gap argument.

5. Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2026). The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*.

Core finding: Three large enterprise field experiments found a 26% increase in weekly task completion for developers using AI coding tools, driven primarily by gains for junior developers and tasks involving code modification and debugging rather than greenfield development.

> → **Why it matters:** Corroborates and expands the 2023 Copilot study with larger, more representative enterprise samples; establishes that AI developer productivity gains are robust across organizations, not limited to controlled experimental settings.

6. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. American Economic Journal: Macroeconomics, 13(1), 333–372.

Core finding: Analyzed historical patterns in general-purpose technology adoption (electricity, IT, and by extension AI); found that productivity gains are systematically delayed because organizational redesign, specifically building the intangible complements, takes years after the technology itself is deployed.

> → **Why it matters**: Provides the historical and economic framework explaining why AI ROI looks disappointing in the near term; leaders who understand the J-curve are less likely to abandon AI investment at the inflection point when organizational redesign has begun but financial measures have not yet caught up.

Group 2: Labor Markets and Task Exposure

These studies address the workforce impact of AI across the macro level, the task level, and in terms of liability and skill development.

7. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. OpenAI Working Paper.

Core finding: Analyzed occupational task exposure to LLMs across the U.S. economy; estimated that approximately 80% of workers have at least 10% of their tasks meaningfully affected by GPT-class models, with higher-skill, higher-wage jobs showing greater exposure than lower-skill roles, a reversal of prior automation patterns.

> → **Why it matters:** Establishes the breadth of AI task exposure and frames the workforce challenge as a task-redesign problem rather than a job-replacement prediction; the reversal of prior automation patterns (white-collar exposure > blue-collar) is the key insight for HR and leadership planning.

8. Humlum, A., & Vestergaard, E. (2025). Large Language Models, Small Labor Market Effects. NBER Working Paper.

Core finding: Analyzed Danish administrative labor market records (one of the most comprehensive national datasets available); found small aggregate employment and wage effects from AI adoption even as task-level changes were real and growing in AI-exposed occupations.

> → **Why it matters:** The most rigorous macro-level treatment of AI's labor market impact; provides the essential counterpoint to both optimistic productivity projections and catastrophic displacement narratives; the honest answer is "significant task disruption, limited aggregate job destruction so far."

9. Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60.

Core finding: Introduced the concept of "moral crumple zones," which refers to the sociotechnical phenomenon where liability for automated system failures is displaced onto the nearest human operator, regardless of that person's actual degree of control over the system.

> → **Why it matters:** The most important governance insight for agentic deployments; human in the loop is not a liability shield unless the human has genuine oversight, meaningful control, and the information necessary to exercise judgment. Leaders designing agent workflows should read this before they define oversight roles.

10. Shen, J. H., & Tamkin, A. (2026). How AI Impacts Skill Formation. arXiv:2601.20245.

Core finding: Analyzed how AI assistance affects skill development in knowledge workers; found evidence that over-reliance on AI outputs, particularly without deliberate practice and reflection, can slow the development of underlying domain knowledge in early-career workers.

> → **Why it matters:** The most recent empirical treatment of the skill-erosion risk that HR leaders need to manage; supports the Junior Gap argument and directly informs how organizations should redesign entry-level roles to preserve development pathways alongside AI-driven productivity gains.

11. Chui, M., Hazan, E., Roberts, R., et al. (2025). The State of AI: Global Survey 2025. McKinsey & Company.

Core finding: Global survey of enterprise AI adoption across industries; found 88% of organizations reporting regular AI use in at least one function, but only approximately 21% reporting fundamentally redesigned workflows, and that minority showing the strongest association with reported bottom-line impact.

> → **Why it matters:** The largest global enterprise AI survey; the 21% workflow-redesign finding is the single most important data point for the redesign-first argument; it establishes that the gap between adoption and value is not a technology problem but an organizational design problem.

Group 3: Evaluation, Benchmarks, and Technical Limits

These sources establish why internal evaluation is essential and what the actual technical limits of current AI systems look like.

12. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., et al. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR 2024.

Core finding: Evaluated AI agents on a benchmark of real, unresolved GitHub issues from major open-source projects; found both genuine progress (leading agents resolving over 20% of issues) and significant limits (most issues requiring multi-file reasoning, deep codebase knowledge, or novel debugging remain beyond current agent capabilities).

→ **Why it matters:** The most rigorous evaluation of AI agents on real-world software tasks; provides the honest technical baseline for IT leaders making GitHub Copilot and agent deployment decisions: progress is real, but production readiness for complex tasks is not.

13. Golchin, S., & Surdeanu, M. (2024). Time Travel in LLMs: Tracing Data Contamination in Large Language Models. ICLR 2024.

Core finding: Demonstrated that data contamination (benchmark answers appearing in model training data), which inflates performance scores on public leaderboards in ways that are difficult for practitioners to detect and that do not predict performance on genuinely novel tasks.

→ **Why it matters:** Explains why vendor benchmark scores consistently overpredict real-world performance; the empirical case for building internal evaluation sets rather than relying on published leaderboards for deployment decisions.

14. Dong, Y., Jiang, Y., Tan, Z., & Zhao, B. Y. (2024). Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. ACL 2024.

Core finding: Studied the contamination problem from a different angle; found that distinguishing genuine generalization from memorization of training examples is a persistent methodological challenge that affects the reliability of comparative evaluations.

→ **Why it matters:** Corroborates Golchin and Surdeanu from a complementary methodological perspective; together these two papers establish the technical case against treating any single benchmark as a reliable procurement criterion.

15. Maslej, N., Fattorini, L., Perrault, R., et al. (2025). The AI Index 2025 Annual Report. Stanford Institute for Human-Centered Artificial Intelligence.

Core finding: Comprehensive annual review of AI progress across research output, industry investment, technical benchmarks, economics, and governance; documents a 280-fold cost collapse in AI inference between 2022 and 2024, U.S. private AI investment reaching $109 billion, and 78% of organizations reporting regular AI use.

→ **Why it matters:** The most authoritative single-source overview of the AI landscape; the evidentiary foundation for the cost-collapse argument and the investment data in the Scoreboard section. Essential context for any executive-level AI strategy discussion.

Group 4: Agentic AI and Governance Frameworks

These sources address the technical and institutional foundations for deploying AI that takes actions, not just generates text.

16. Wang, L., Ma, C., Feng, X., et al. (2024). A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science, 18(6).

Core finding: Comprehensive technical survey mapping the architecture, capabilities, and failure modes of LLM-based autonomous agents across planning, memory, tool use, and multi-agent coordination; identified reliability, cost control, and safety as the primary constraints on production deployment.

> → **Why it matters:** The essential technical foundation for understanding what agents actually are and how they fail; bridges research and practice for leaders making agent deployment decisions and provides the vocabulary for governance conversations about agentic systems.

17. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. DeepSeek Technical Report.

Core finding: Demonstrated that reinforcement learning on chain-of-thought reasoning data can produce genuine structured reasoning capabilities, not just pattern-matching fluency, at cost-competitive levels; achieved performance comparable to leading frontier models on mathematical reasoning, coding, and multi-step analysis benchmarks.

> → **Why it matters:** Establishes the architectural basis for the reasoning model shift and signals that reasoning capability is no longer exclusive to the most expensive frontier models; the most important technical release of early 2025 for business leaders assessing the task frontier.

18. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

Core finding: Establishes a voluntary, risk-based framework for organizations to identify, manage, and communicate about AI risks through four integrated functions: Govern (set culture and accountability), Map (categorize risk context), Measure (analyze and assess risk), and Manage (prioritize and respond).

> → **Why it matters:** The foundational U.S. governance standard for trustworthy AI; increasingly referenced in enterprise procurement, federal regulatory conversations, and board-level AI governance discussions. The Govern function in particular translates directly into the ownership and accountability structures needed to move pilots to production.

19. National Institute of Standards and Technology. (2024). Generative Artificial Intelligence Profile (AI 600-1). NIST.

Core finding: Extends the AI RMF with a risk profile specific to generative AI systems, addressing hallucination, confabulation, data leakage, prompt injection, misuse, homogenization, and overreliance as distinct risk categories requiring specific controls.

> → **Why it matters:** The most practical governance reference for enterprise generative AI deployments; maps directly to the failure modes leaders encounter when moving from pilot to production. The prompt injection and overreliance sections are required reading for anyone deploying customer-facing AI.

20. International Organization for Standardization. (2023). ISO/IEC 42001:2023—Information Technology—Artificial Intelligence—Management System.

Core finding: Establishes the international standard for an AI management system (AIMS), covering policy, organizational roles, risk assessment, objectives, controls, internal audit, management review, and continual improvement; this is the full management system architecture applied to AI governance.

> → **Why it matters:** The global governance framework that complements NIST in international contexts; increasingly relevant for organizations with EU, Asia-Pacific, or multinational operations and for enterprise sales conversations where governance certification provides a competitive advantage.

21. Deloitte AI Institute. (2026). State of AI in the Enterprise: The Untapped Edge. Deloitte.

Core finding: Survey of 3,235 business and IT leaders across 24 countries; found that only 25% of organizations have moved 40% or more of their AI experiments into production, with the report emphasizing the pilot-to-production gap and the need for clear strategy to reduce “pilot fatigue.” The press release also notes that 85% of companies expect to customize AI agents to fit their business needs (Deloitte, 2026).

> → **Why it matters:** The most current large-scale enterprise AI adoption survey; the 25% pilot-to-production finding is the most accurate description of where most organizations actually stand in early 2026, and it is the basis for the Pilot vs. Production leadership implication.

22. GitHub. (2025). Octoverse: The State of Open Source and Rise of AI in 2025. GitHub, Inc.

Core finding: Analysis of activity across 180+ million GitHub developers; found over 4.3 million AI-related repositories (nearly doubled since 2023). GitHub reports that more than 1.1 million public repositories now import an LLM SDK (+178% YoY as of Aug

2025 vs. Aug 2024), and that TypeScript overtook Python and JavaScript in Aug 2025 to become the most used language on GitHub, while Python remains dominant for AI and data science workloads (GitHub, 2025; updated 2026).

> → **Why it matters:** Developer activity predicts enterprise purchasing 6-12 months out; the shift from LLM experimentation to agent orchestration frameworks as a standard architectural pattern is the signal that enterprise AI platforms will consolidate around in 2026.

Group 5: Security, Data Integrity, and Systemic Risks

These sources address the security and data risks that become operationally significant at production scale.

23. Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. Anthropic Research.

Core finding: Demonstrated experimentally that AI models can be trained, or could emerge, with behaviors that appear safe under normal conditions but activate under specific triggers; crucially, these hidden behaviors persisted through multiple rounds of standard safety training that were designed to remove them.

> → **Why it matters:** The most important security research for enterprise AI deployment; establishes that observed behavioral safety cannot be assumed to generalize, and that governance must be designed architecturally rather than inferred from testing under normal conditions. The write-access rule follows directly from this research.

24. OWASP Foundation. (2025). OWASP Top 10 for LLM Applications.

Core finding: Practitioner-focused taxonomy of the ten most critical security risks in LLM-based applications, based on real-world deployment patterns; prompt injection (malicious instructions overriding system behavior) and sensitive information disclosure top the list, both appearing repeatedly in production enterprise deployments.

> → **Why it matters:** Translates academic security research into actionable enterprise controls; the prompt injection risk is directly relevant to any organization deploying agents that read external documents, emails, or web content, which describes most agentic deployments.

25. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI Models Collapse When Trained on Recursively Generated Data. Nature, 631, 755–759.

Core finding: Published in Nature; demonstrated that training AI models on AI-generated data causes progressive quality degradation across successive generations of training, a phenomenon the authors term “model collapse,” as models drift away from the statistical distribution of real human-generated data.

> → **Why it matters:** Establishes the empirical foundation for treating proprietary human-generated data as a competitive asset; as AI-generated content proliferates across the web and in organizational knowledge bases, organizations that maintain clean, well-governed human-generated data pipelines will have a structural advantage in model quality and fine-tuning effectiveness.