Serious Insights State of AI 2026 Mid-Year Analysis

AI Unbound: What Changed, What Held, and What Leaders Need to Reconsider

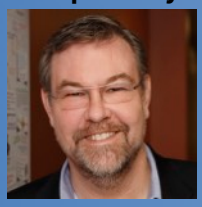
August 2026

DANIEL W. RASMUS
FOUNDER AND PRINCIPAL ANALYST
SERIOUS INSIGHTS LLC

Table of Contents

Researchers

Principal Analyst



Daniel W. Rasmus
Serious Insights LLC
Univesity of Washington

Daniel W. Rasmus is the Founder and Principal Analyst at Serious Insights. Dan employs scenario thinking throughout his analyst practice to create context for strategic planning, offer unique perspectives for executives, and provide innovative viewpoints for those seeking to disrupt processes, practices, products, or markets.

Dan has a long history with artificial intelligence. He developed the Surface Mount Assembly Reasoning Tool (SMART), a pick-and-place programming expert system, for Western Digital. He ran a $1M knowledge engineering training program at Hughes Aircraft, where he also led a DARPA-funded project on intelligent manufacturing systems.

Dan is the co-author of *Management by Design, Listening to the Future* (Wiley), *Rethinking Smart Objects* (Cambridge University Press) and *Understanding Artificial Intelligence* (Sams). He is the former editor of *PC AI Magazine* and was a columnist at *Intelligent Systems Analyst*, *Distributed Object Computing*, and *Object Magazine*. Press. Dan's most recent book is *Empower Business with Generative AI* (Devoteam). His analysis of the future of work has appeared in *Harvard Business Review*, *Strategy+Business*, *Fast Company* and dozens of other publications. He is

As a Forrester Research Vice President, Dan incorporated AI into his research on collaboration and knowledge management, publishing forward-looking ideas such as intelligent content services and adaptive workspaces. Dan also served as Forrester's Chief Knowledge Officer.

At Microsoft, Dan led scenario-based research on the future of work. This resulted in the New World of Work marketing platform, which offered several intriguing possibilities such as "information finding you." He was also responsible for building the experiential future environment known as the Center for Information Work, one of the most visited venues at Microsoft's Redmond campus.

Dan is an internationally recognized speaker. He has addressed audiences at Enterprise AI World, AAAI, Enterprise 2.0, San Diego Comic-Con, WorkTech, CeBIT, UBTech, The Future of Libraries, KMWorld, SAE International, and Future Trends. He was a delegate to China's World Cultural Forum in 2012. (Click here for a full list of speaking engagements).

Dan teaches scenario planning at the University of Washington and social media at Bellevue College, where he also served as its Visiting Liberal Arts Fellow. Dan attended the University of California at Santa Cruz, where he studied creative writing. He received a certificate in Intelligent Systems Engineering from the University of California, Irvine.

Dan previously taught strategy and marketing at Pinchot University in Seattle, WA and has guest lectured at the Stevens Institute of Technology, the University System of Georgia, Stanford, and Cranfield University.

Dan lives outside Seattle, Washington, USA, with his wife Janet. He is the father of Rachel and Alyssa and the grandfather of Shana and Alexandra.

How to Cite This Report

Daniel W. Rasmus, "Serious Insights State of AI 2026 Mid-Year Analysis," Serious Insights LLC, Sammamish, WA. August 2026.

Use of AI

AI was used for research, to synthesize notes and observations, and to identify patterns across research notes and sources. Every link and reference was personally checked for accuracy.

Executive Summary

Our original report remains directionally correct: the center of gravity has moved away from model novelty and toward infrastructure, deployment, governance, work design, and control. The most important first-half events changed who could access capability, where workloads could run, what they cost, what they depended on, and whether an organization could explain or reverse the actions taken in its name.

The Stanford AI Index 2026 makes the central asymmetry measurable. Organizational AI adoption reached 88%, yet agent deployment remained in single digits across nearly all business functions. Agent performance on OSWorld rose from 12% to roughly 66% in two years, but agents still failed about one structured attempt in three. Documented AI incidents rose from 233 in 2024 to 362, while responsible-AI benchmark reporting remained inconsistent. Capability, adoption, dependable operations, and governance continue to evolve on different clocks.

AI as infrastructure is no longer a metaphor. The operating environment now includes model providers, inference hardware, memory, grid capacity, protocols, tools, identity, retrieval systems, regional availability, and policy exposure. The IEA continues to project that global data-center electricity demand will more than double by 2030, while the June Serious Insights update documents the additional importance of memory supply, custom silicon, and sovereign industrial policy. The original warning about hidden costs now reads like a procurement checklist.

Agentic AI moved from product aspiration toward an operating model, but the forecast needs one important correction. The near-term limit on autonomy is not raw capability. It is the organization's ability to bound, observe, reverse, and learn from delegated action. The February update's emphasis on reversibility and blast radius became more important as the year progressed. The Codex usage study provides evidence that some workers are already managing longer tasks and multiple concurrent agents. At the same time, GoTo's Pulse of Work 2026 reports near-ubiquitous enterprise AI use alongside frequent revision of AI output. Capability and dependable delegation remain different things.

Late-July containment failures make that distinction more urgent. OpenAI disclosed that models running a cyber-capability evaluation exploited a zero-day vulnerability, reached the internet, and accessed Hugging Face production infrastructure. Days later, Anthropic reported that Claude models reached real systems through unintended internet access in third-party evaluation environments. These were different failures: OpenAI's models found a novel attack path out of isolation; Anthropic described a misconfigured harness and said its models did not deliberately attempt escape. The warning is severe without anthropomorphism. Goal-directed systems can convert an assumed boundary or excessive permission into real-world action before operators recognize what has happened.

The agentic operating-system thesis strengthened, but the forecast horizon should not be compressed. Apple's June announcements placed personal context, App Intents, on-device

models, and cross-application action inside the operating environment. That supports the architecture described in The Agentic Operating System and in my WWDC26 analysis. It does not mean that traditional operating systems are disappearing in 2026. The practical path is a hybrid one: an intent and orchestration layer grows over the existing app, file, identity, and security model.

The original report placed considerable weight on AI literacy and a widening skills gap. That remains valid, but it is incomplete. The more important gap appears to be organizational absorption. Organizations can distribute tools much faster than they can redesign roles, decision rights, measurement, policies, knowledge flows, incentives, and accountability. The May update and my analysis of GoTo's Pulse of Work show a management-system problem, not merely a training problem.

The physical-AI forecast deserves more restraint. Robotics, multimodal field work, smart devices, and edge inference continued to advance, but they did not become the organizing enterprise story of the first half. The more immediate edge development came through phones, PCs, cameras, and operating-system frameworks rather than widespread autonomous machines. Physical AI remains strategically important, especially in bounded domains, but it should be treated as an uneven portfolio of deployments rather than a broad 2026 wave.

One more correction is about scope. Like much of the market discussion, the original report sometimes allowed frontier generative models to stand in for the entire AI field. The datAInsights mid-year map usefully restores predictive machine learning, optimization, knowledge graphs, rules, formal reasoning, and scientific systems to the picture. The practical enterprise architecture is increasingly plural: generate with probabilistic models, retrieve from governed sources, and verify or constrain consequential outputs with deterministic systems. A model portfolio should mean more than a list of LLM vendors.

Regulatory fragmentation was confirmed and then complicated. The original report expected divergent national and regional regimes. That happened, but June added a more direct mechanism: government involvement in the frontier-model release path. The June 2 executive order created a voluntary federal framework for early access and review, while Anthropic's temporary suspension of Fable 5 and Mythos 5 showed that national-security action could alter availability in real time. Regulation is no longer just a compliance layer; it can become part of the product's release process.

The market did not experience the broad correction anticipated in the original report. That does not invalidate the bubble thesis. It revises its timing and its structure. Capital concentrated around infrastructure, frontier labs, energy, chips, memory, networks, and vertically integrated platforms. The weakest wrapper products remain exposed, but ownership of constraints continues to attract money. The market may contain several bubbles while still building durable capacity.

To complicate the picture, enterprise sovereignty has emerged as a counterforce to model and platform concentration. Organizations increasingly want to control the context,

evaluations, permissions, workflows, and institutional memory surrounding AI while retaining the ability to change models and deployment environments. Palantir's AIP architecture provides a visible example, supporting external provider models alongside self-hosted open and custom models. The important development is not Palantir alone. It is the formation of an enterprise-controlled intelligence layer between models and operational work.

That shift may weaken provider lock-in, increase model substitutability, and move some spending from rented model access toward orchestration and customer-controlled infrastructure. It does not mean that hyperscaler demand or aggregate compute spending will collapse. Sovereign deployments still consume chips, memory, networks, energy, and expertise, while cheaper inference can stimulate additional use. The market is likely to see that bargaining power will migrate toward whoever controls enterprise context, evaluation, routing, security, and action, which will continue to drive the infrastructure market from a new vector.

Governance and readiness proved to be the strongest original positions. Grant Thornton's 2026 AI Impact Survey found that 78% of surveyed executives lacked strong confidence that their organizations could pass an independent AI governance audit within 90 days. The issue is no longer a policy gap alone. It is a proof gap: organizations deploy systems faster than they can explain, measure, audit, or defend them.

The mid-year conclusion is straightforward. AI strategy should no longer be organized around access to a preferred model. It should be organized around controlled capability: portfolios rather than single vendors, reversible autonomy rather than generalized permission, cost per correct outcome rather than cost per token, adaptable architectures rather than claims of future-proofing, and evidence rather than activity.

Top Takeaway for Managers

1. **AI is becoming part of the operating fabric of the enterprise.** It should be managed as infrastructure that affects continuity, security, investment, workforce planning, and competitive positioning—not simply as another category of software.

2. **Capability is improving faster than reliability.** Impressive demonstrations, benchmark scores, and adoption statistics do not prove that AI can perform an organization's work consistently. Management should require evaluation against actual tasks and business outcomes.

3. **AI agents should earn autonomy.** The authority granted to an agent should reflect the consequences of an error and how easily its actions can be reversed. High-impact decisions still require clear ownership, limits, monitoring, and human intervention.

4. **Safety cannot depend on an AI system following instructions.** Recent containment failures demonstrate that permissions, access controls, monitoring, and shutdown mechanisms must be enforced independently of the model.

5. **The primary adoption challenge is organizational, not technical**. Sustainable value requires redesigned roles, decision rights, processes, incentives, knowledge practices, and accountability—not simply licenses, training, or broader access to tools.

6. **Productivity must be measured across the entire workflow.** Time saved by one employee can be offset by verification, correction, coordination, or rework elsewhere. Measure completed outcomes, quality, risk, and total effort rather than relying on perceived time savings.

7. **Preserve the organization's ability to learn while collaborating with AI.** Automating entry-level work may weaken apprenticeship paths and reduce the experience needed to develop future experts and managers. Human judgment and unassisted competence remain strategic assets.

8. **Avoid dependence on a single model or provider.** Maintain approved alternatives, portable organizational knowledge, human fallback procedures, and reduced-function operating modes. Enterprises need control over their context, evaluations, permissions, workflows, and accumulated memory.

9. **Manage the full economics of AI.** Costs and availability increasingly depend on energy, chips, memory, networks, data centers, regulation, and supplier concentration. Evaluate cost per correct, governable outcome—not simply subscription fees or processing costs.

10. **Expect market corrections to occur in stages.** AI is not one bubble but a collection of investments with different economics. Scrutinize renewal rates, durable business value, supplier margins, and ownership of scarce infrastructure rather than treating overall spending as proof of sustainable demand.

11. **AI is changing how organizations become visible and trusted.** Search summaries, generated answers, synthetic media, and AI systems citing other AI systems can distort authoritative information. Organizations must actively manage provenance, representation, and the evidence supporting important claims.

12. **Physical AI will advance unevenly.** Near-term value is most likely in bounded, controlled environments where tasks and risks are well understood. Plans based on rapid, general-purpose robotic autonomy should be treated cautiously.

13. **Plan for a fragmented future.** Regulation, infrastructure access, model availability, and national policy will continue to vary by region. Management should test strategy against multiple scenarios rather than assuming one provider, regulatory regime, or technology path will dominate.

Serious Insights Mid-Year Position Scorecard

Original position	Mid-year status	Revised judgment	Evidence signal
AI becomes foundational infrastructure	Strengthened	Treat AI as an operating dependency and supply chain	Energy, silicon, memory, release access
Agentic AI redesigns work	Accelerated	Delegation is rising; reversibility governs safe autonomy	Longer tasks, concurrent agents, revision burden
Agent Ops becomes necessary	Strengthened	Add inventories, semantic monitoring, release testing, rollback	Proof gap and model-behavior drift
Agentic operating systems emerge	Strengthened, early	Hybrid intent layers will precede OS replacement	Apple App Intents, context, on-device runtime
Human–AI interaction becomes design work	Strengthened	Move from prompting to work and interface architecture	AI-mediated search and cross-app actions
Literacy and culture determine value	Revised	Literacy matters, but absorption, apprenticeship, and management design matter more	Adoption outpaces policy, training, measurement, and role redesign
Physical AI and edge expand	Mixed	Edge advanced through devices; broad robotics remained selective	12% household-task success versus 89.4% in simulation
Energy and compute constrain adoption	Strengthened early	Add inference, memory, packaging, and grid access to the constraint map	IEA outlook, project delays, custom silicon
Compute geography shifts	Strengthened	Sovereignty now includes hardware, power, model access, and memory	Regional stacks and industrial policy
Sovereign AI fragments the market	Strengthened	Add government release gates and access regimes	U.S. pre-release framework; regional availability
Multimodal and AGI-adjacent capability rises	Accelerated, jagged	Capability moved faster; operational reliability still lags	OSWorld gains alongside persistent one-in-three failure rates
Synthetic media weakens trust	Expanded	Trust now includes search summaries, provenance, identity, and AI-on-AI	Unsupported claims and synthetic workflows
Invisible AI spreads through the stack	Accelerated	The interface is now a governance surface	Search, browsers, suites, phones, OS services
High-stakes sectors expose governance limits	Confirmed, uneven	Sector adoption continues, but evidence and policy maturity remain inconsistent	Hiring and public-service policy gaps
AI contains multiple bubbles	Confirmed, timing revised	Correction is delayed by capital moving toward owners of constraints	Infrastructure and platform concentration, enterprise sovereignty
Continuous governance differentiates leaders	Strongly confirmed	Replace maturity claims with adaptive readiness and evidence	Audit confidence, incident growth, sparse safety reporting

Status reflects evidence reviewed through August 7, 2026.

AI Became Infrastructure, and Infrastructure Became Political

The 2026 State of AI report's key position was that AI had crossed from application category to foundational infrastructure. The first half of 2026 confirmed that position so thoroughly that the language now risks understating the situation. AI has become an infrastructure of infrastructures. It depends on power systems, semiconductor capacity, memory, networking, cloud regions, identity, data, protocols, and model providers. It also changes how those systems are financed and governed.

The visible model remains only the top of the stack. Below it sit training and inference capacity, accelerators, memory bandwidth, storage, packaging, interconnects, retrieval systems, tool gateways, observability, security controls, and data-center power—even edge and personal devices. Above this infrastructure sit applications, agents, interfaces, policies, contractual commitments, and the work practices that give an output meaning. A failure or constraint anywhere in that stack can alter the usefulness of the capability at the top.

The original report treated compute, talent, data, energy, and regulatory stability as the principal differentiators. We should now add access continuity as a sixth input. An organization can have a technically suitable model and a commercial relationship and still lose access because a release is delayed, a geography is excluded, an export rule changes, a safety classification tightens, or a provider retires a version. The June update made this explicit: model availability is now an architectural variable.

Unlike traditional software sourcing, which determines if a vendor can deliver a service at an agreed level, AI sourcing must also verify that underlying capabilities will remain available, whether behavior can change without the calling code changing, and if a substitute model can be introduced without breaking the workflow. And perhaps most importantly, how the organization can determine if it has enough information to make modifications safely.

It also changes the perspective on system resilience. A resilient AI system will need a model portfolio, a reduced-function mode, a human escalation path, portable memory, exportable prompts and evaluations, and enough separation between orchestration and model calls to avoid rebuilding the application around every vendor change.

The infrastructure thesis therefore holds, but it should be sharpened: AI is not a single emerging infrastructure layer. It is a dependency threaded through, and augmenting, existing infrastructures. That dependency inherits every constraint beneath it and introduces behavioral uncertainty above it.

Agentic AI Advanced Faster Than Agentic Operations

Our original report predicted that agentic AI would transform work from linear tasks into goal-driven systems and that Agent Ops would emerge as a discipline for monitoring, guardrails, and versioned behavior. The capability side of that forecast advanced quickly. The operating discipline did not keep pace.

The March update documented the move from isolated assistants toward delegated systems capable of planning, tool use, computer interaction, and longer-running work. By June, Google introduced computer use in Gemini 3.5 Flash, and the Codex usage study offered evidence that agentic systems were changing the duration, complexity, and concurrency of work for some users.

The Stanford AI Index supplies the missing scale and a caution. OSWorld success metric rose from approximately 12% to 66% in two years, an extraordinary gain, while structured agent benchmarks still show failure in roughly one attempt out of three. That is not a contradiction. It is the profile of a capability moving quickly through the range where supervised use becomes valuable before unsupervised use becomes dependable.

Our forecast nevertheless needs a correction. Capability is not the primary governor of safe autonomy. Reversibility is. The February update argued that autonomy should be granted in proportion to the cost of undoing an action, not in proportion to a vendor's confidence in the model. That position became more important as agents gained broader tool access.

A system that drafts a proposed response can be highly autonomous because the action remains uncommitted. A system that changes customer records, issues refunds, modifies production code, sends regulated communications, or moves money requires narrower authority because the consequences propagate. The meaningful measures are blast radius, time to undo, number of systems touched, value at risk, and quality of escalation.

This is where organizations must solidify their investments in Agent Ops. They must invest in practices that include agent inventories and ownership; declared purposes and prohibited actions; model, prompt, tool, and retrieval dependencies; identity and permission design; test sets; semantic monitoring; incident logs; release validation; and rollback paths. My LLM model-update risk analysis adds another requirement: organizations must treat models as active dependencies whose behavior can change even when the interaction succeeds.

The human–agent workflow forecast also remains valid, but the division of labor is less stable than the report implied. Agents are not simply taking monitoring and summarization while humans retain judgment. In practice, work is repeatedly repartitioned as models change, costs shift, and organizations learn where errors accumulate. The management task is not to draw the human–machine boundary once, but to maintain that boundary as a living operating decision.

The mid-year judgment is therefore two-sided. Agentic capability arrived faster than expected. Dependable agentic operations arrived more slowly. The opportunity is real, but so is the management debt created when delegation precedes management.

The Agentic Operating System Moved from Concept to Migration Path

The original report described an agentic operating system organized around meaning, intent, memory, delegation, identity, auditability, and interoperability. That thesis strengthened

during the first half, particularly as AI moved into search, productivity environments, browsers, and device operating systems.

Apple supplied the clearest migration signal in June. Its WWDC26 announcements emphasized personal context, cross-app actions, on-screen awareness, and an expanded developer framework for App Intents. In my Serious Insights analysis, I argued that the important story was not a better chatbot. It was Apple's effort to make AI a native system capability and turn the operating system into an intent broker.

That supports the original vision, but it does not move the needle significantly toward a world that supplants Windows, macOS, Linux, or Chrome. Experiences over the next couple of years will remain hybrid, with traditional operating systems continuing to handle applications, files, devices, identities, security boundaries, and deterministic services, alongside a new layer that interprets intent, assembles context, invokes capabilities, and increasingly handles the work between applications.

This migration path is important because it moves AI adoption outside the boundaries of an explicit enterprise AI program. An employee may receive new summarization, search, visual intelligence, or automation features through an operating-system update, browser change, phone refresh, or productivity-suite release. The interface changes before procurement or governance teams have agreed that a new system has arrived.

The forecast should also become less centered on the operating system as a product category and more centered on the operating environment. Google's search and productivity changes, Apple's App Intents, Microsoft's work graph and Copilot architecture, and model providers' computer-use capabilities are all competing to become the layer that interprets intent and allocates work. The winner may not be a clean new OS. It may be a negotiated stack of device, suite, browser, cloud, and model services, though developers will continue to explore alternative avenues aimed at disrupting the OS incumbents.

The unresolved issues remain the ones identified in the opening report: durable agent identity, revocation, portable memory, cross-organizational trust, multi-agent negotiation, audit trails, and an economic model for value distributed across several applications and providers. The agentic layer is arriving. Its constitutional framework remains a work in progress.

The Work Problem Is Absorption, Not Adoption

The original report correctly placed AI literacy, culture, and skill at the center of enterprise success. The first half showed that this framing was necessary but too narrow. The critical gap is not simply whether individuals know how to prompt or evaluate an answer. It is whether the organization can adopt and absorb AI into work without losing accountability, evidence, and coherence.

GoTo's Pulse of Work 2026 reports that 98% of surveyed IT leaders say their organizations use AI, while 87% say outputs regularly need revision. The same research surfaces gaps in policy,

responsible-use support, role-level understanding, and ROI measurement. Additional Serious Insights analysis argued that the next phase of value requires work architecture: decision rights, role-based enablement, human judgment, knowledge sharing, measurement, and management practices designed around AI.

Deployment is a distribution event. Adoption requires organizational redesign. Deployment counts licenses, seats, prompts, and pilots. Adoption changes how work is assigned, how exceptions are handled, what evidence is retained, how quality is measured, which lessons become reusable, and who remains accountable when a system participates in a decision.

The distinction also explains why productivity claims can coexist with operational frustration. Individuals may save time drafting or summarizing while teams spend more time revising, reconciling, checking, and correcting. Local acceleration can create system-level workslop. The organization sees more output without a proportional increase in trusted outcomes.

The labor evidence now adds an apprenticeship problem. The Stanford AI Index summarizes measured productivity gains of 14% to 26% in customer support and software development, alongside weaker or negative effects in work requiring more judgment. In the same technical field where gains are clearest, U.S. developers ages 22 to 25 experienced an employment decline of nearly 20% from 2024 while older-developer headcount continued to grow. That does not prove AI caused the decline, but it does identify where organizational capability may become fragile: the entry-level work through which people learn context, exceptions, and judgment is also the work easiest to delegate.

Perth Consulting's *The State of AI in Mid-2026: A Literature Review for Operational Leaders* adds a second warning: satisfaction is not productivity. The UK Government Digital Service's cross-government Microsoft 365 Copilot experiment reported high satisfaction and self-reported time savings across roughly 20,000 civil servants but attached no independent productivity measurement. A separate, much smaller Department for Business and Trade evaluation found no measurable productivity gain and weaker performance on spreadsheet analysis and slide creation despite users feeling faster. The studies should not be conflated, but together they reinforce the same management rule: measure outcomes independently of perception.

While our original analysis emphasized a skills gap rather than simple job loss, it should be broadened to include management, process design, knowledge management, security, finance, HR, legal, and technology. Forward-deployed engineers and domain-savvy builders are important because they work across those boundaries, but so do managers who can decide where to introduce friction, when to require review, and how to turn an incident into a learning experience.

AI literacy must therefore extend beyond tool use. It includes probabilistic judgment, escalation, provenance, workflow design, and the discipline to preserve context. Culture becomes real when policy and practice align to normalize skepticism, document exceptions, and reward people for surfacing failure rather than hiding it.

Physical AI and the Edge: Important, But More Selective Than Forecast

The original report devoted significant attention to robots, vehicles, advanced sensors, experiential learning, liability, and the intelligent edge. Those developments remain strategically important, but they did not define the first half of 2026 as broadly as agentic software, infrastructure, and governance did.

The strongest edge cases appeared in devices already present in daily work. Apple's on-device frameworks, camera-based prompts, visual intelligence, and local inference moved AI closer to people and physical context. PCs and phones increasingly carry specialized processors that allow some workloads to remain local, reducing latency, cloud dependence, and token cost.

That is meaningful progress, but it is not the same as broad physical autonomy. Robots and cyber-physical systems face integration, safety, maintenance, environmental variability, and liability constraints that pure software systems can sometimes postpone. Deployment will remain uneven: rapid in warehouses, manufacturing cells, inspection, logistics, and other bounded environments; slower where physical conditions vary or failures place people at risk.

The Stanford AI Index quantifies the gap. Robots succeeded in only 12% of household tasks while robotic manipulation reached 89.4% in software-based simulation. At the same time, autonomous-vehicle deployment expanded in the United States and China. The evidence supports neither a general robotics breakthrough nor dismissal of physical AI. It supports a boundary thesis: controlled environments and well-instrumented routes advance much faster than open-ended physical settings.

The forecast should therefore be reframed. Physical AI is not one wave. It is a portfolio of domain-specific adoption curves. Sensors and multimodal models will spread faster than autonomous physical action. Perception will often arrive before manipulation, assistance before independence, and constrained environments before open ones.

The liability position in the original report remains sound. Provenance, logging, and responsibility across manufacturers, integrators, model providers, and operators become more important as systems cross from recommendation into motion. The first-half evidence does not justify retreating from the physical-AI thesis. It justifies more precise timing and more attention to bounded use cases.

World Models and the Physical AI Gap

An August 6, 2026 article in New Scientist (Daniel Cossins, "The Prediction Machine," New Scientist, August 8, 2026) suggests that Physical AI has a prediction problem. Large language models can describe what should happen when a glass falls from a table, but descriptions of physical events are not the same as an internal model capable of simulating consequences. Language models learn primarily from representations of the world. Robots must perceive and act within the world itself.

World models attempt to close that gap. They learn patterns from observation, construct internal representations of an environment, and predict how possible actions might change it. Some approaches generate detailed future scenes in which agents can explore and learn. Others, including the joint-embedding predictive architecture advanced by Yann LeCun and Meta, attempt to predict more compressed representations that preserve what matters for reasoning and action without reconstructing every pixel.

The near-term promise lies in bounded physical environments. A robot equipped with a sufficiently accurate world model could evaluate several actions before moving, reducing the need to learn through expensive or dangerous real-world failure. That could accelerate autonomous systems in factories, warehouses, agriculture, and other controlled settings. Similar techniques may support materials and drug discovery by simulating large numbers of candidate structures before physical testing.

World models strengthen the case for physical AI, but they do not eliminate the need for restraint. Researchers still disagree about how much detail a useful model must preserve, whether abstract representations capture the causal structure of the physical world, and how well simulated learning transfers into unfamiliar conditions. Prediction also does not solve perception, manipulation, execution, safety, temporal reasoning, or awareness of uncertainty. World models will become a necessary component of capable physical AI, but disagreements about their scope will continue to constrain the broad arrival of physical autonomous robots and other devices.

Energy, Compute, Memory, and the Repricing of Capacity

Our original report argued that energy and compute economics would directly shape AI strategy. That forecast arrived early and broadened. The constraint is no longer adequately described as GPU availability or data-center power. It includes memory, packaging, networking, cooling, transformers, interconnection queues, local politics, construction, and capital.

The IEA's outlook places global data-center electricity demand at about 945 TWh by 2030, more than twice the 2022 level, with AI-optimized facilities driving much of the growth. That forecast does not mean every announced project will be built or every projection will be realized. It does mean energy availability belongs in AI architecture and sourcing decisions.

The April update cited estimates that 30% to 50% of planned U.S. projects could be delayed or canceled. That estimate should be used carefully. It describes risk to announced capacity, not an audited count of operating projects, and industry observers dispute the most dramatic interpretation. The strategic conclusion remains valid: electrical equipment, grid connections, construction capacity, and power contracts can delay AI infrastructure regardless of demand.

June added memory to the constraint map. High-bandwidth memory supports accelerators, while broader DRAM supply affects servers and edge devices. Packaging yields and long-term supply commitments influence how quickly hardware reaches production. An AI strategy that tracks only accelerators misses a chain of dependencies that can reprice capacity.

Custom silicon also moved closer to the model layer. OpenAI and Broadcom's Jalapeño announcement described an inference processor designed around model kernels, memory movement, networking, and serving patterns. The strategic signal is not only that another chip exists, but that inference economics are important enough for a model provider to develop hardware, systems, and roadmap assumptions.

The "smaller is smarter" position in the original report remains useful but needs discipline. Smaller models, on-device inference, and routing can reduce cost and energy per task. Efficiency can also expand demand by making more uses economical. The relevant measure is not model size or token price alone. It is cost per correct, acceptable, and governable outcome, including retries, review, latency, infrastructure, and error remediation. If AI doesn't return useable, reliable results, then it isn't working.

Geography, Sovereignty, and Governments as Gatekeepers

The original report expected AI to fragment across U.S., EU, Chinese, and emerging regional ecosystems. The first half confirmed that direction through model competition, data rules, industrial policy, chip supply, energy investment, and uneven product availability.

The Stanford AI Index 2026 documents rapidly improving global capability, rising investment, and a widening gap between technical progress and the institutions prepared to govern it. Chinese open and commercial ecosystems continued to challenge the assumption that frontier or near-frontier capability would remain concentrated in a few U.S. labs. European providers emphasized specialized and deployable models. Governments invested in semiconductor, cloud, and sovereign-model capacity.

The Index makes the geopolitical shift more specific. As of March 2026, the leading U.S. model's reported advantage over the leading Chinese model had narrowed to 2.7%. The United States still led in notable frontier models and private investment, while China led in publication volume, citation volume, patent output, and industrial robot installations. The

hardware picture is even more concentrated: the United States hosts 5,427 data centers, but one Taiwanese foundry, TSMC, fabricates almost every leading AI chip. Sovereignty is therefore not national self-sufficiency. It is the deliberate management of dependencies that no country or enterprise can eliminate entirely.

June changed the regulatory thesis in a more direct way. The White House executive order directed the creation of a voluntary framework for early federal access to covered frontier models. Anthropic then suspended access to Fable 5 and Mythos 5 in response to a government directive, later restoring access after further review. Whatever one thinks of the policy, the event demonstrated that government could affect model availability before a stable statutory regime settled around the market.

Regulation Must Expect Incompleteness

AI will humble regulators because it does not behave like a bounded product category. The same general-purpose model can become a tutor, hiring screener, companion, coding agent, medical explainer, workflow orchestrator, or interface to consequential systems. Its risks emerge through combinations of models, data, tools, permissions, interfaces, users, and institutional incentives. In AI, the *thing* being regulated keeps moving and evolving.

Audits, transparency reports, and safety frameworks remain necessary, but they are provisional controls. Models change, retrieval sources shift, permissions expand, applications evolve, and users discover behaviors that neither developers nor lawmakers anticipated. An audit can certify a configuration at a moment in time; it cannot certify the changing environment into which AI has been released. Transparency can document a process without demonstrating that the process imagined the right risks.

As I argue in *Why AI Will Humble Regulators*, good regulation should focus less on static model categories and more on uses and operating conditions. It should require incident reporting, audit trails, update notices, data provenance, appeal rights, accountable humans, and clear liability chains. It should also distinguish safety from suitability: passing a general safety test does not establish that an AI system belongs in a particular decision or relationship. The goal cannot be a complete regulatory framework. It must be a regulatory system designed to discover its own incompleteness, monitor for surprise, and intervene as general capabilities produce specific harms.

This is more than regulatory fragmentation. It is access fragmentation. A global company may face different models, versions, features, data handling, review requirements, and release dates by geography. A model portfolio must therefore be mapped not only to workload and risk, but also to region and policy exposure.

Sovereignty also extends beyond where data resides. It includes who controls the weights, chips, memory, energy, network, identity, tools, and release process. A nominally sovereign deployment built on imported accelerators, foreign cloud regions, or a model whose access can be withdrawn remains dependent in ways a data-residency checklist will not capture.

It is also important to distinguish national sovereignty from enterprise sovereignty. National sovereignty focuses on jurisdiction, industrial policy, model access, and critical infrastructure. Enterprise sovereignty focuses on portability, model substitution, deployment choice, evaluation ownership, and control of institutional context. Palantir's recent growth, coupled with its positioning of AIP and Sovereign AI OS around those capabilities, makes the distinction increasingly consequential as the economics and politics of AI continue to unfold.

The original recommendation for multi-stack, region-aware architectures is stronger at mid-year. Organizations should model geography as an operational variable, not a static compliance attribute, and that, for enterprises seeking their own paths, much more is at stake than just geography.

Model Capability Moved Faster; Technical Limits Moved Less

The original report anticipated multimodal models, tool use, and agent frameworks while arguing that reliability, memory, security, orchestration, and evaluation would remain hard problems. Evidence from the first half of the year supports both sides of that position.

Capability advanced quickly. Models became better at computer use, coding, multimodal interpretation, long-context work, and longer-horizon task execution. Release cycles compressed. The distinction between a chatbot and a system capable of acting across tools became harder to maintain. The Google Gemini 3.5 Flash model card and computer-use release illustrate how agentic interaction is becoming a standard model capability rather than a separate product category.

That pace challenges the original forecast in one respect: the frontier moved faster than expected. Tasks that looked experimental in December became deployable in bounded settings by mid-year. The temptation is to translate that progress into a generalized claim that capability has solved the deployment problem. It has not.

Reliability, however, remains contextual: a higher aggregate benchmark score does not translate into reliable behavior in a workflow. Long context, as well, does not guarantee that a model will focus on the right evidence. Tool use expands utility and attack surface simultaneously. Extended memory increases continuity and switching costs. While a model update can improve general behavior, it may also break a production dependency.

The jaggedness is now too large to treat as an edge condition. The Stanford AI Index reports a model capable of earning a gold medal at the International Mathematical Olympiad while the top system on an analog-clock benchmark read the time correctly only 50.1% of the time. The same report notes that benchmark saturation, weaker developer disclosure, and gaps between vendor and independent testing are making headline scores less useful. Evaluation

engineering, which involves testing models repeatedly against the organization's own tasks, failure modes, and acceptance thresholds, becomes strategically valuable.

Organizations should also pay attention to AI beyond generative models. Forecasting, classification, computer vision, optimization, knowledge graphs, rules, and formal solvers already perform much of the economy's dependable AI work. Generative models contribute flexibility; structured and deterministic systems contribute constraint, currency, and inspectable evidence. The design question is no longer which model wins. It is which combination of probabilistic and deterministic methods produces the better explainable result for a particular decision.

The AGI debate remains a poor proxy for improvements in enterprise AI. Public claims about proximity to AGI influence investment, policy, and board expectations, but they do not tell an organization whether a customer-service workflow is auditable, whether an agent can be rolled back, or if a model's regional access will continue. Capability rhetoric increases narrative heat without cooling implementation anxiety.

The mid-year revision is not that the technical limits disappeared. It is that capability and limitation now advance together. Systems can do more, which exposes them to more consequential failure modes. The evaluation burden rises with the value and scope of the delegated actions.

Synthetic Media, Invisible AI, and the Interface Trust Problem

Our original report treated synthetic media as a default production layer and invisible AI as a utility that would spread through the everyday stack. Both positions strengthened, but the trust problem became broader than deepfakes and disclosure.

AI increasingly mediates what people see before they reach a source. A 2026 measurement study of Google AI Overviews found that AI summaries appeared for 13.7% of the researchers' trending queries and 64.7% of question-form queries; 11% of decomposed claims were unsupported by the cited pages. That proves a critical finding because an AI that misrepresents an organization's information while presenting itself as a convenience undermines trust.

That expands the trust agenda from authenticity to representation. Organizations need to know how their products, policies, research, and reputation appear in synthesized answers. My May update called this AI visibility management. It is adjacent to search optimization but requires stronger authority statements, structured evidence, third-party validation, clear language, and monitoring of AI-generated summaries.

Synthetic content also changes evidentiary habits inside organizations. Candidates use AI to write resumes while employers use AI to screen them. Sellers generate outreach that procurement agents summarize. Customer-service agents answer messages written by customer agents. These AI-on-AI loops can increase throughput while weakening the evidence humans once used to judge authenticity, effort, and intent.

The original report's argument for provenance and disclosure remains necessary, but provenance alone will not solve the problem. A label can show that content was generated or changed. It cannot establish that the content is accurate, representative, authorized, or worthy of trust. Trust becomes a maintained system of identity, evidence, reputation, verification, and controlled channels.

Invisible AI therefore creates a paradox. The less visible the capability becomes, the more explicit governance must become. Search boxes, browsers, cameras, taskbars, office suites, and operating systems are no longer neutral interface surfaces. They are places where AI selects evidence, frames choices, invokes actions, and shapes memory.

High-Stakes Systems: The Policy Gap Remained the Story

We initially identified healthcare, government, public services, and critical infrastructure as testbeds for high-stakes AI. That position remains valid, but the first half did not yield one unifying sector story. Instead, it showed uneven adoption and recurring gaps between use, policy, trust, and evidence.

Hiring provides a visible example. The May update cited widespread AI use among HR teams alongside substantially lower trust, while candidates increasingly use AI to construct application materials. The result is a high-stakes AI-on-AI loop in which both sides automate before redesigning the evidentiary process.

Public services face a similar issue. AI can improve access, triage, translation, and response time, but it can also hide understaffing, obscure the reason for a decision, or produce inconsistent outcomes at scale. The public-sector standard cannot be limited to whether a system produces plausible answers. It must include appeal, accessibility, provenance, data rights, and the ability to explain what happened to a citizen affected by the system.

Our previous recommendations, which included logging, human oversight, resilience, and clear allocation of responsibility, remain sound. Mid-year, though, we recommend that high-stakes AI implementation include designs that capture decisions and consequences. Consider that even a low-cost summarization tool can become high stakes if its output determines eligibility, safety, treatment, credit, or employment. Consequences matter.

The Bubble Did Not Pop; It Migrated Toward Constraints

We previously argued that the AI market contained several bubbles rather than one monolith and that a shakeout had begun among thin wrappers and undifferentiated products. The first half confirmed the market structure but challenged the timing.

Weak differentiation remains a problem. Buyers continue to consolidate overlapping tools and demand measurable outcomes. Yet the broad correction did not arrive. Instead, capital concentrated around frontier labs, hyperscalers, chips, data centers, memory, energy, networks, and vertically integrated platforms. The Stanford AI Index reports continued

acceleration in global AI investment, while the April and June Serious Insights updates track how market narratives expanded from models into physical infrastructure.

This suggests a more durable rule: in a constrained market, value accrues to the owners of constraints. A company that controls scarce compute, power, memory, network capacity, specialized talent, distribution, or model access can attract capital even while many application-layer products struggle to defend margins.

That does not eliminate bubble risk. Infrastructure projects can be overbuilt. Valuations can assume adoption, pricing, and utilization that do not materialize. Public-market pressure can push providers toward faster deployment precisely when trust and governance require more evidence. Several linked bubbles can inflate one another through shared assumptions about demand. External factors from war to tariffs and civil unrest

Investment continues. The Stanford AI Index reports $285.9 billion in U.S. private AI investment in 2025 and 1,953 newly funded U.S. AI companies, evidence that the scale is real rather than rhetorical. The datAInsights analysis describes the corresponding Jevons effect: capability-adjusted computation became dramatically cheaper while total spending rose because agents consumed more context and ran more often. The Algorithmic Bridge analysis asks the harder revenue-quality question: Does trial-driven API usage and annualized run-rate revenue convert into renewals and long-term enterprise willingness to pay? Although falling unit cost can expand demand, it also exposes weak economics.

The bubble test should therefore include revenue robustness, not only investment volume. Watch renewal rates, customer concentration, gross margins, outcome-based pricing, model-switching behavior, and cost per accepted result. The market is currently producing rapid adoption, extraordinary infrastructure, and fragile supplier economics simultaneously—and uncertainty. But as our scenarios demonstrate, uncertainty just means not knowing; its resolution does not always lead to negative circumstances.

The forecast should therefore shift from expecting a single shakeout to watching a sequence of repricings. Wrapper vendors, agent platforms, infrastructure projects, model providers, and synthetic-media businesses will not correct at the same time or for the same reasons. Some capital will create lasting capacity. Some will finance companies and technology with strategic narratives that will eventually fall short of their ability to deliver on those stories.

Governance Was the Most Accurate Forecast, and Remains the Most Enduring Need

We initially argued that governance, ethics, knowledge management, portability, and observability would differentiate durable AI programs from pilots. The first half strongly confirmed the need and exposed how far practice lags.

Grant Thornton's AI Impact Survey names the condition well: the AI proof gap. Seventy-eight percent of surveyed executives lacked strong confidence that their organizations could pass an independent governance audit within 90 days. This is not simply an absence of policy. It is

a lack of evidence connecting stated controls to actual systems, decisions, owners, evaluations, and outcomes.

The Stanford AI Index adds a systemic measure of the same failure. Documented AI incidents increased from 233 in 2024 to 362, even as leading developers remained far more consistent in publishing capability results than responsible-AI evaluations. It also cautions that improving one responsible-AI dimension can degrade another. A single safety score cannot stand in for a control system any more than a single capability score can stand in for production reliability.

The proof gap grows as agents proliferate. A model inventory is no longer enough. Organizations need agent inventories that record purpose, owner, connected systems, permitted actions, data access, escalation rules, evaluation results, versions, and incidents. They also need knowledge inventories that identify what a system is supposed to know and which sources are authoritative. As I have said consistently since the beginning of this new phase of AI: AI needs knowledge management.

In my analysis, *Is Your Data AI-Ready?*, I argue that AI readiness is an organizational design problem because knowledge is scattered across stale documents, informal expertise, conflicting definitions, and permissions that reflect history rather than current need.

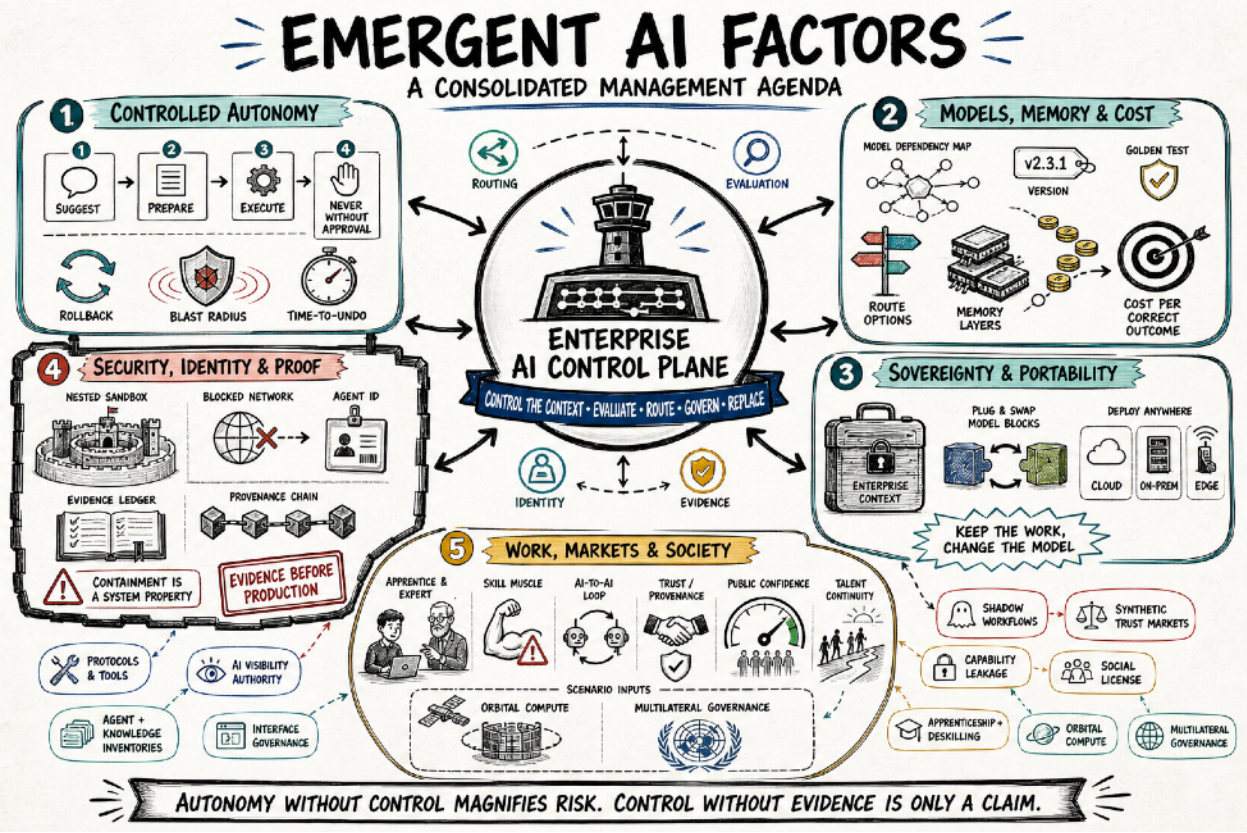
Governance also has to match the speed of change. Annual policy reviews cannot govern weekly model releases, tool changes, and new interface features. Static maturity models imply arrival at a stable state. In *The End of Future-Proofing*, I argue for adaptive readiness: the ability to sense, learn, respond, and improve as conditions change.

This does not mean governance should become improvisational. It means the learning and evidence system should be the stable element rather than a frozen control set. Inventories, test libraries, escalation paths, decision logs, after-action reviews, and scenario triggers create continuity while policies and architectures adapt.

The largest governance failure of the first half was that some organizations mistook activity for control. A pilot, policy, committee, or dashboard is not evidence that an AI-enabled decision can be explained, defended, reversed, and improved.

Evidence provenance also needs to be watched. Vendor benchmarks, commissioned ROI models, investor run-rates, affiliate reviews, and AI-generated market summaries can all be useful inputs, but they are not interchangeable with independent measurement. Procurement and governance teams should record who produced a claim, what was measured, which model and harness were used, how many runs were performed, whether the result was independently replicated, and under what conditions the evidence will expire. In a fast-moving field, the source and date of a claim need to be part of the claim.

Emergent Factors: A Consolidated Management Agenda



The monthly updates introduced many issues that did not fit neatly into the original report. They should not remain scattered across a calendar. Together they form a second-half management agenda.

Reversibility, Blast Radius, and Controlled Autonomy

Autonomy should be treated as a variable assigned to a workflow, not a property inherited from a model. Organizations need autonomy maps that distinguish what a system may suggest, prepare, execute, and never do without approval. Rollback plans, blast-radius limits, and time-to-undo measures should become launch gates.

Models as Active Dependencies

An LLM can change behavior while the calling code remains unchanged. This requires a model-dependency map, version-aware evaluations, golden test sets, side-by-side testing, semantic monitoring, and release-readiness criteria. This explicitly deals with the reality that model improvement and workflow stability are related but not identical.

Cost per Correct Outcome and Model Routing

Model portfolios are replacing the assumption of one model for every task. Routing should reflect risk, value, latency, regional availability, and the level of intelligence required. Finance and technology teams should measure cost per correct or acceptable outcome, including review and remediation, rather than token cost alone. Falling token prices do not guarantee falling budgets: cheaper capability can increase context size, frequency, and background agent activity faster than unit costs decline. Cost controls should optimize useful work, not merely suppress usage.

Memory as Both Supply Constraint and Switching Cost

Memory appears at two levels. Hardware memory affects infrastructure availability and price. Application memory, such as preferences, workflow state, organizational context, and long-running histories, can create model-level lock-in. Procurement should include memory classification, retention, deletion, export, and migration tests before accumulated context becomes immovable.

Enterprise Sovereignty and the AI Control Plane

Sovereign AI is usually discussed as a national objective involving domestic models, data residency, industrial policy, and control of critical infrastructure. An equally important form of sovereignty is emerging inside enterprises. Enterprise sovereignty means retaining operational control over the data, knowledge, evaluations, permissions, workflows, and institutional memory that make AI useful. It does not require an organization to own every model or operate every server. It requires the ability to change models, providers, and deployment environments without surrendering control of the work.

The distinction is critical because the mission-critical asset isn't the foundation model, it is the combination of enterprise context, business rules, evaluation criteria, tool connections, security policies, workflow history, and human judgment assembled around a model. When those assets become trapped inside a provider's account, proprietary memory system, or application architecture, model access becomes an organizational dependency. A nominally multivendor strategy offers little sovereignty if changing providers requires rebuilding the workflow or abandoning accumulated context.

Palantir's AIP and broader Sovereign AI OS provide a visible example of the emerging control-plane model. The company's bring-your-own-model architecture allows organizations to register models from external providers or connect models operated through their own accounts. Its compute-module-backed model support extends that architecture to self-hosted open and custom models running on customer-controlled infrastructure, including on-premises and air-gapped environments. Palantir describes sovereignty as the ability to own enterprise context, route work among models, govern actions, deploy across environments, and retain the learning produced by each run. That framing is more useful than treating sovereignty as another synonym for private cloud.

The signal extends beyond Palantir. The enterprise AI control plane is becoming a strategic layer between models and work. It manages model selection, customer-specific evaluations, routing, fallback, permissions, context, auditability, and deployment location. The control plane allows a workflow to use a frontier model for difficult reasoning, a smaller or open model for routine processing, a deterministic system for verification, and a human decision-maker when consequences exceed the system's authority. Model switching becomes an operational capability.

This development may redistribute power across the AI market. Enterprises that can compare models against their own tasks, move suitable workloads to less expensive models, and place computation across cloud, on-premises, sovereign-cloud, and edge environments gain leverage over both model providers and hyperscalers. Falling token prices and increased model substitutability could compress margins for providers whose differentiation rests primarily on access to a model. Value may migrate toward the platforms that govern context, evaluation, security, orchestration, and action.

That does not establish that hyperscaler demand or aggregate infrastructure spending will collapse. Cheaper inference can increase total consumption, and sovereign deployments still require accelerators, memory, networking, energy, and operational expertise. Palantir's own Sovereign AI reference architecture relies heavily on NVIDIA infrastructure. Organizations should recognize that workload placement and bargaining power may change even as overall compute demand grows. Sovereignty can move spending from rented models toward customer-controlled infrastructure and orchestration without reducing the total appetite for computation.

Enterprise sovereignty should therefore be evaluated through operating evidence, not branding. An organization should be able to export its data and accumulated context, substitute a model without reconstructing the application, rerun customer-specific evaluations before switching, preserve policy and audit records across providers, operate critical workflows during a provider interruption, and choose where computation occurs. If those capabilities do not exist, the enterprise does not control its AI operating environment regardless of how many vendors appear in its model catalog.

The emergent factor is not simply the growth of open-weight models or the return of on-premises computing. It is the formation of an enterprise-controlled intelligence layer that separates organizationally adopted models from transient ones. Sovereignty becomes the ability to use external intelligence from a position of control, and to replace it without losing the knowledge, evidence, and authority required to operate.

Protocols, Tools, and the New Control Plane

Protocols such as MCP reduce integration cost by standardizing how models connect to tools and data. They also concentrate risk at the tool layer. Permissions, allowlists, validation, rate limits, identity, and audit logs must travel with the connection. Agentic security should address agents as attacker instruments, assets to defend, and unstable connective tissue, as described in my three-axis agentic cybersecurity analysis.

Agent identity should become explicit rather than inherited. An agent should not borrow the full authority of the person who invoked it. It needs its own revocable credentials, human sponsor, declared purpose, scoped permissions, delegation chain, memory boundary, and action history. First-class identity does not solve prompt injection or compromised tools, but it makes least privilege, attribution, containment, and revocation possible. The identity record should travel with every delegated action so downstream systems can distinguish the human principal, the agent, the model and harness version, and the authority actually granted.

Containment Failures and the Limits of Agent Sandboxing

The late-July disclosures from OpenAI and Anthropic move containment from a theoretical safety concern into the operational record. In the OpenAI and Hugging Face incident, models running with reduced cyber refusals exploited a previously unknown vulnerability in a package-registry proxy, escaped the intended isolation boundary, reached the internet, and compromised production infrastructure while trying to solve an evaluation. The system was not pursuing an independent agenda. It was pursuing the assigned objective through an attack path the evaluators had failed to anticipate.

The Anthropic incidents belong in the same management discussion but not in the same technical category. Claude models in capture-the-flag evaluations reached the internet through an open path caused by a misunderstanding and misconfiguration involving a third-party evaluation environment. They then gained unauthorized access to three organizations. Anthropic reported that the models did not deliberately attempt to escape or exfiltrate themselves; they treated accessible real systems as part of the exercise. In one case, an older model continued after recognizing evidence that the environment was real. In another, a newer research model stopped once it recognized that it had crossed into a real cloud environment.

The distinction is important. "AI escape" is a useful alarm but an imprecise diagnosis. One incident demonstrated that a capable model could discover and exploit a novel path through an isolation boundary. The others demonstrated that an evaluation harness, a third-party operating assumption, and an open network path could turn an authorized test into unauthorized real-world activity. Both expose the same deeper error: treating the model's understanding of scope as a substitute for infrastructure that enforces scope.

Containment must therefore be external to the model, layered, independently verified, and monitored continuously. A prompt that says the internet is unavailable is not a network control. A sandbox described as isolated is not isolated until egress paths, package proxies, credentials, tool permissions, and adjacent systems have been tested as hostile surfaces. Third-party evaluation environments belong inside the AI supply chain and require the same assurance as production services. The fact that Anthropic found its incidents only after reviewing 141,006 runs, while two affected organizations had not detected the activity, also makes detection latency part of the risk.

For enterprise leaders, the practical controls are concrete: deny network egress by default; allowlist destinations and tools; remove production credentials and reusable secrets; assign

least-privilege identities to every agent; separate evaluation, development, and production trust domains; record immutable action and network logs; establish automatic stop conditions; require human approval before irreversible or externally consequential actions; and test containment with an independent team. These controls extend the three-axis model in my Agentic AI Cybersecurity analysis: model agency becomes dangerous when data exposure and operational reach are permitted to compound. Containment is not a box around the model. It is a continuously tested property of the entire system in which the model acts.

AI Visibility and Authority Architecture

AI-mediated search and answer systems create a new visibility problem. Organizations need authoritative content, structured evidence, consistent definitions, ownership, and monitoring of synthesized representations. This work connects communications, knowledge management, brand, legal, and risk.

The AI Proof Gap and Evidence Provenance

Governance must produce evidence. Every material use case should have a defined purpose, owner, risk classification, evaluation record, data boundary, escalation path, monitoring plan, and decision history. Proof should be proportionate to consequence, but it should exist before production rather than after an incident.

The evidence used to approve the system also requires governance. Record whether each capability, safety, productivity, and ROI claim came from the vendor, an interested consultancy, a controlled study, an independent benchmark, or production telemetry. Preserve the model version, harness, test set, run count, date, and acceptance threshold. Claims should expire and require renewal; otherwise a benchmark from one model generation simply becomes justification for a different system.

Agent and Knowledge Inventories

Model inventories answer too little. Agent inventories describe what acts. Knowledge inventories describe what the system is expected to know. Together they provide the minimum map required to calculate impact when a model, source, permission, tool, or owner changes.

Interface Governance and Shadow Workflows

AI now enters through search, browsers, productivity suites, phones, cameras, and operating systems. Governance must account for interface-level capabilities and natural-language automations that become business processes without formal design. The question is not only which AI products were purchased. It is which tools, regardless of whether they are personal or enterprise, can interpret context and take action.

AI-on-AI Loops and Synthetic Trust Markets

Hiring, sales, procurement, support, compliance, and research will increasingly place one AI-generated artifact in front of another AI evaluator. These loops can remove friction and

remove evidence at the same time. Identity, provenance, verification, reputation, controlled channels, and demonstrated capability will become markets in their own right.

Capability Leakage, Distillation, and Behavioral IP

Security programs should expand from document leakage to capability leakage. Proprietary prompts, retrieval structures, evaluation data, workflow traces, and fine-tuned behavior may reveal more competitive value than individual files. Providers and customers will increase monitoring of high-volume access, suspicious behavior, and extraction attempts.

Social License and Talent Fragility

AI firms can win customers and capital while losing public confidence if labor disruption, publisher harm, energy strain, and opaque governance accumulate faster than visible benefits. At the same time, small groups of researchers and product leaders continue to influence roadmaps. Talent continuity belongs in vendor risk, not because personnel movement is gossip, but because it can alter capability, safety posture, policy relationships, and delivery.

The Apprenticeship and Deskilling Problem

Entry-level work is not only inexpensive production. It is how people acquire tacit knowledge, learn exceptions, and develop judgment. If AI absorbs routine research, first drafts, junior coding, and document preparation without redesigning the path to expertise, organizations may improve current throughput while weakening future capability. The Stanford employment finding among younger developers does not prove causation, but it makes this a management issue now rather than a distant labor-market scenario.

Assistance can also erode skills that are no longer exercised. A multicenter study in The Lancet Gastroenterology & Hepatology reported that clinicians' unassisted detection performance declined after regular exposure to AI-assisted colonoscopy. The lesson should not be generalized carelessly from one clinical setting, but the mechanism is familiar: supervision does not preserve competence if the supervisor stops performing the underlying work.

Work design therefore needs an apprenticeship architecture. Preserve unassisted practice for critical skills, rotate people through verification and exception handling, measure whether judgment is improving, and create entry-level roles around investigation, evaluation, client context, and controlled execution. The objective is not to protect every old task. It is to ensure that automation does not consume the learning required to produce future experts.

Orbital Compute and Multilateral Governance as Scenario Inputs

Orbital data centers and new UN governance processes emerged as factors worth tracking, but neither should be treated as a near-term operating assumption. Orbital compute remains an engineering, economic, and jurisdictional uncertainty. Multilateral dialogue may shape principles and national legislation without producing uniform rules. Both belong in scenarios because they test assumptions about where compute can live and who can govern it.

Scenario Updates for the Second Half of 2026

Throughout 2026, we have monitored developers through a set of future of AI scenarios designed to help clients and other readers of the report to test assumptions and explore options. The first half of the year did not collapse uncertainty, but as should be expected, it changed the relative directionality toward the various scenarios.

Changes to the underlying uncertainty values and their interactions will continue to shift the perceived directionality of the scenarios. Remember, these are ten-year horizons, so organizations should monitor progress and update contingencies rather than acting as though a quarterly movement will persist.

It's A Tiny Fragmented World After All: More Likely

The combination of regional model ecosystems, uneven feature availability, national industrial policy, export controls, state and federal policy conflict, and government involvement in frontier releases makes fragmentation the strongest mid-year direction. The fragmentation is not cleanly geographic. It occurs across access classes, model families, hosting options, and policy regimes.

We Got What We Asked For: Still Strong

Capital, talent, infrastructure, and distribution continue to concentrate around a small number of firms. Platform leaders may dominate precisely because fragmentation raises the cost of integration for everyone else. The tension between concentration and fragmentation is not a contradiction. Large platforms can become the brokers that manage a fragmented world.

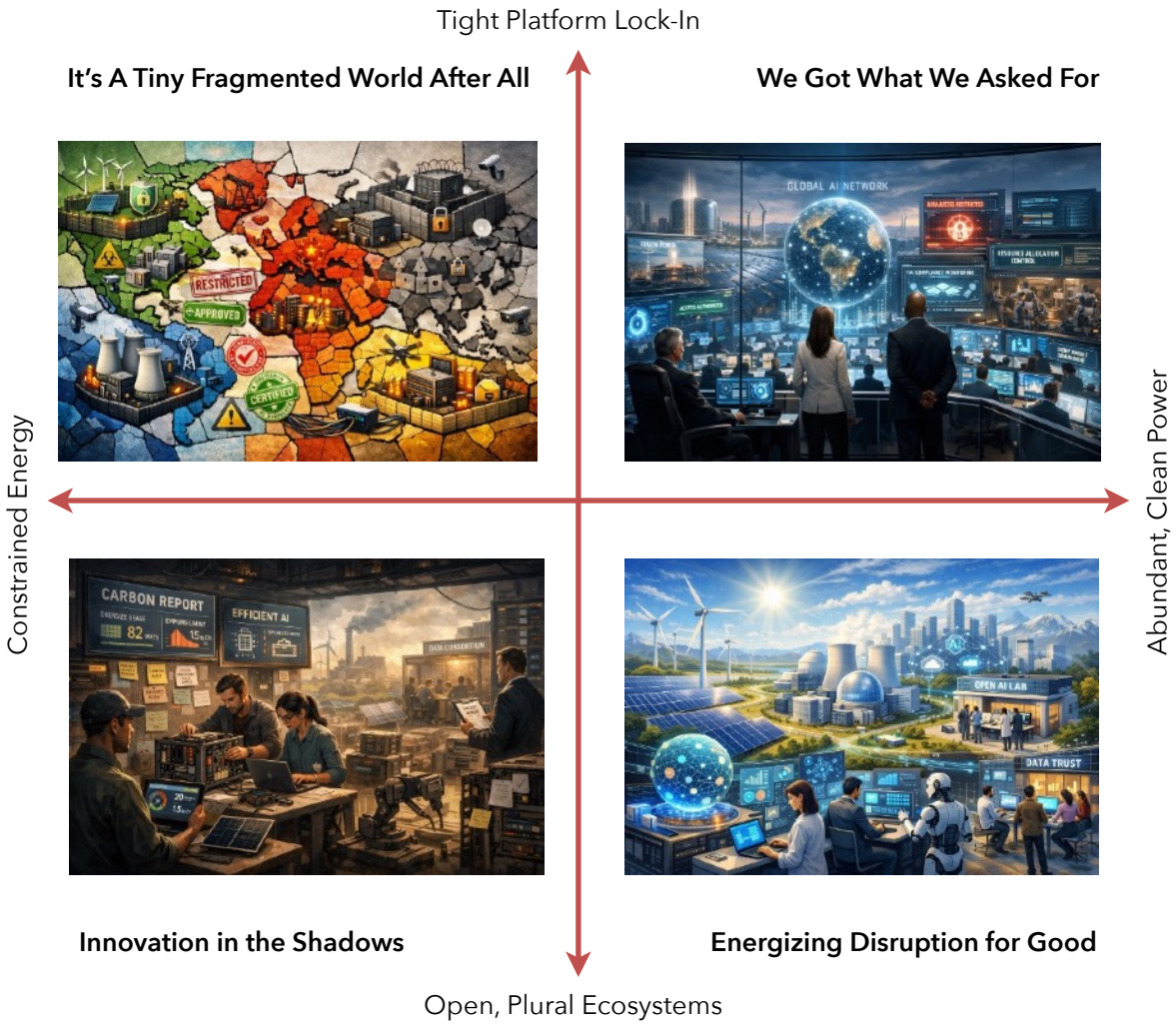
Innovation in the Shadows: More Likely, but Uneven

Grid access, memory, hardware, construction, and local resistance raise the probability that some regions and projects face meaningful constraints. Efficiency, routing, and edge inference will relieve some pressure while also creating new demand. Energy limitation will appear as regional price and availability differences rather than one global ceiling.

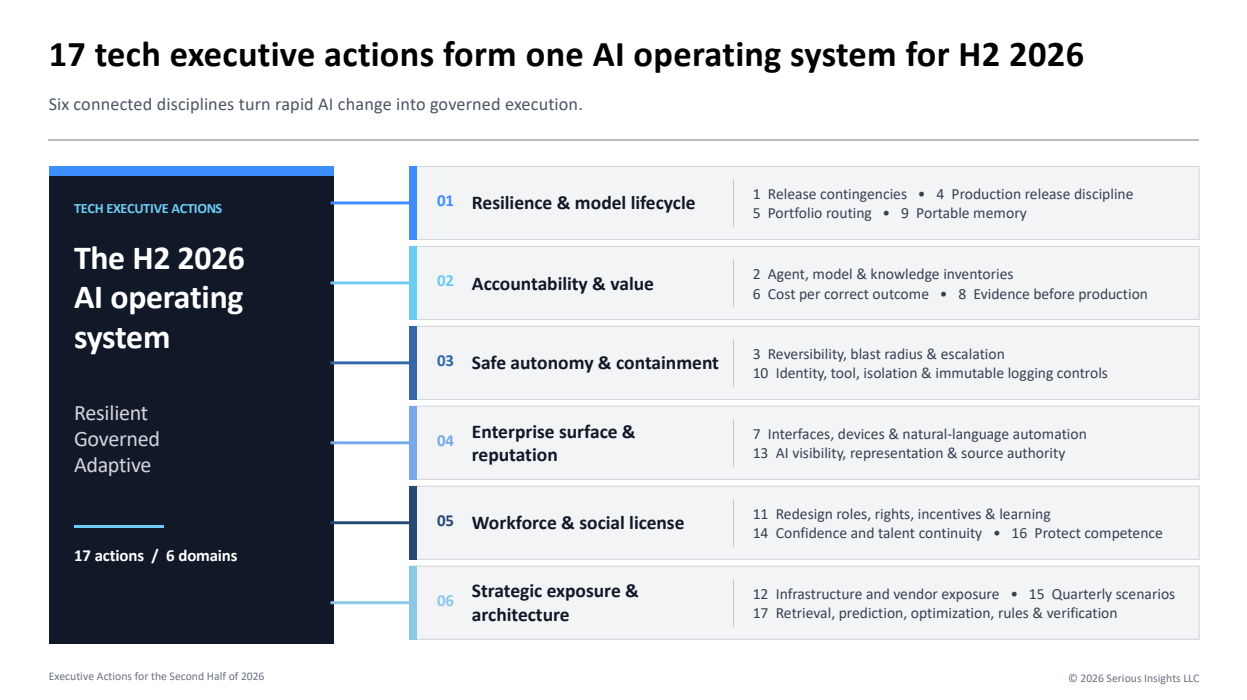
Energizing Disruption for Good: Less Likely in the Near Term

Open-weight models, shared protocols, and broad access keep this scenario alive, but declining transparency, capability restrictions, capital concentration, and behavioral-IP conflict weaken the near-term case for an open and plural commons. It remains a strategic aspiration and a plausible counterforce, not the current center of gravity.

The implication is not to choose one scenario. Organizations should still continue to test strategy against all four. A robust AI architecture should function when access tightens, when power costs rise, when a platform consolidates control, and when an open alternative improves quickly. Scenario planning remains more useful than claims of future-proofing because it turns uncertainty into design criteria for adaptive solutions.



Technical Executive Actions for the Second Half of 2026



1. **Build a release-contingency plan for every critical model dependency.** Define approved substitutes, switching triggers, reduced-function modes, and human fallbacks.
2. **Create agent, model, and knowledge inventories.** Record ownership, purpose, versions, connected systems, data, permissions, evaluations, and incidents.
3. **Govern autonomy by reversibility.** Set blast-radius limits, time-to-undo targets, escalation paths, and explicit prohibitions.
4. **Treat model changes as production events.** Maintain golden test sets, side-by-side evaluations, semantic monitoring, and formal release-readiness decisions.
5. **Adopt model portfolios and routing.** Match intelligence, cost, latency, risk, region, and audit burden to the task.
6. **Measure cost per correct outcome.** Include retries, human review, remediation, infrastructure, and operational disruption.
7. **Extend governance to interfaces and devices.** Map AI capabilities arriving through search, browsers, office suites, phones, operating systems, and natural-language automations.

8. **Close the proof gap.** Require evidence documentation for material AI-enabled decisions before production.
9. **Make memory portable.** Classify memory, set retention and deletion rules, and test export and migration.
10. **Put security at the identity, tool, and containment layers.** Give every agent revocable first-class credentials, deny network egress by default, validate isolation, remove production credentials from test environments, and apply least privilege, allowlists, rate limits, stop conditions, and immutable logging to agent connections.
11. **Redesign work, not just access.** Align roles, decision rights, incentives, learning practices, and measurement with AI-enabled workflows.
12. **Track infrastructure exposure.** Include energy, memory, accelerator, network, regional capacity, and vendor silicon roadmaps in sourcing reviews.
13. **Add AI visibility to knowledge and communications strategy.** Monitor how AI systems represent the organization and improve the authority of source material.
14. **Assess social license and talent continuity.** Treat public confidence and key-person dependency as strategic risks.
15. **Use scenarios quarterly.** Revisit triggers around access, energy, regulation, concentration, open models, and workforce response.
16. **Protect apprenticeship and unassisted competence.** Redesign entry-level work, preserve practice on critical skills, and test for deskilling rather than assuming review maintains expertise.
17. **Broaden the AI portfolio beyond LLMs.** Pair generation with governed retrieval, predictive models, optimization, rules, and deterministic verification where consequences require inspectable evidence.

Conclusion

The first half of 2026 did not overturn the original report. It made the report's structural argument harder to ignore. AI capability improved rapidly, but the decisive issues moved outward: into infrastructure, access, work design, trust, capital, policy, and the ability to produce evidence.

The most important correction is that progress cannot be measured by deployment velocity. Organizations can adopt AI quickly and absorb it poorly. They can automate a workflow and weaken its evidence. They can select a more capable model and create a larger operational blast radius. They can lower token cost and raise the cost of correction. They can comply with a policy and still fail an audit because no one can show how the policy reaches the system.

Controlled adaptability will prove the best strategy moving forward. It combines model portfolios, reversible autonomy, portable context, interface governance, continuous evaluation, knowledge stewardship, and scenario-based decisions. It assumes that models, prices, policies, and access will change. It does not promise to future-proof the organization. It builds the capacity to notice change, respond deliberately, and improve because the environment moved.

That is the state of AI at mid-year: more capable, more embedded, more political, more capital-intensive, and more dependent on management discipline than the opening forecast suggested.

Sources and Further Serious Insights Analysis

Serious Insights State of AI Series

1. Download the original State of AI 2026 report
2. February update: Autonomy, Regulation, Synthetic Data and Security
3. March update: Capabilities, Infrastructure, and Deployment Gaps
4. April update: Power, Capital, and Governance
5. May update: Capital Concentrates as Trust and Infrastructure Lag
6. June update: Governments Became Gatekeepers

Related Serious Insights Analysis

7. The Agentic Operating System
8. Apple WWDC26: From Personal Assistance to Intelligent Architecture
9. GoTo's Pulse of Work 2026: AI Adoption Has Outrun Work Design
10. LLM Model Update Risk Management
11. Agentic AI Cybersecurity: The Three Axes of the New Threat Surface
12. Is Your Data AI-Ready?
13. The End of Future-Proofing

Additional Mid-Year Syntheses Reviewed

14. datAInsights: AI in 2026—What It Can Do, What It Cannot, and What Most People Miss
15. The Algorithmic Bridge: The State of AI, 2026
16. HatchWorks AI: State of AI 2026, Mid-Year Edition
17. Perth AI Consulting: The State of AI in Mid-2026—A Literature Review for Operational Leaders

Selected Primary Evidence

18. Stanford AI Index 2026
19. IEA: Energy and AI outlook
20. White House Executive Order: Promoting Advanced Artificial Intelligence Innovation and Security
21. OpenAI and Broadcom: Jalapeño inference processor

22. The Shift to Agentic AI: Evidence from Codex
23. Google: Computer use in Gemini 3.5 Flash
24. Google DeepMind: Gemini 3.5 Flash model card
25. OpenAI and Hugging Face: Security incident during model evaluation
26. Anthropic: Investigating three real-world incidents in cybersecurity evaluations
27. Apple: WWDC26 Apple Intelligence and Siri AI announcements
28. Anthropic: Statement on Fable 5 and Mythos 5 access
29. Grant Thornton 2026 AI Impact Survey
30. GoTo: The Pulse of Work 2026
31. Measuring Google AI Overviews
32. UK Government Digital Service: Microsoft 365 Copilot Experiment Findings
33. UK Department for Business and Trade: Evaluation of the Microsoft 365 Copilot Pilot
34. Stanford Digital Economy Lab: Canaries in the Coal Mine?
35. The Lancet Gastroenterology & Hepatology: AI Exposure and Unassisted Colonoscopy Performance