

Agentic AI

Governing the Agentic Enterprise: A New Operating Model for Autonomous AI at Scale

Sandeep Saini

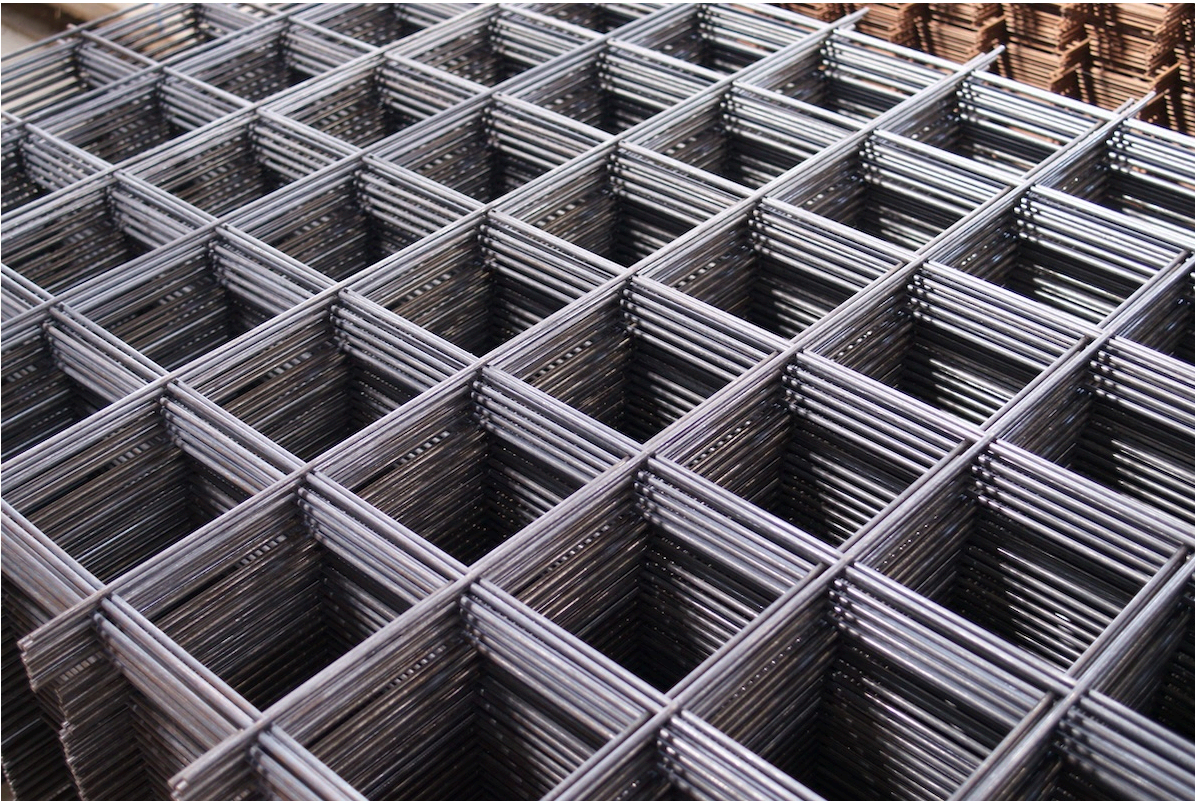


Image Credit | travelers.high

AI agents have transitioned from "tools" to "actors" where humans set boundaries and "guardrail agents" physically block high-risk actions in real-time.

As organizations deploy increasingly autonomous artificial intelligence systems, many are discovering that existing governance and operating models are ill-suited to software that can independently perceive, decide, and act. While recent advances in generative AI have focused on model capability, the more consequential challenge for enterprises lies in governing systems that function as organizational actors rather than decision-support tools. This article argues that autonomous AI represents an institutional shift, not merely a technological one.

To address this challenge, the article **proposes** the **Agentic Operating Model (AOM)**, a **conceptual & illustrative** governance framework that specifies the structural conditions required to operate autonomous agents responsibly at enterprise scale. The AOM comprises four interdependent layers, cognitive specialization, coordination architecture, real-time control, and organizational governance, that together constrain autonomy while preserving its benefits. Drawing on illustrative enterprise vignettes, the article demonstrates how failures in agentic systems typically arise from misalignment across these layers rather than from deficiencies in model performance.

The article contributes a practical and conceptual foundation for leaders seeking to scale autonomous AI without sacrificing accountability, resilience, or trust. By reframing agentic AI as an operating-model problem, it offers senior executives a systematic approach to governing autonomy as a durable source of competitive advantage.

RELATED ARTICLES

Oliver Gassmann and Joakim Wincent, "**The Non-Human Enterprise: How AI Agents Reshape Organizations**," *California Management Review Insights*, October 22, 2025.

Mohammad Hossein Jarrahi and Paavo Ritala, "**Rethinking AI Agents: A Principal-Agent Perspective**," *California Management Review Insights*, July 23, 2025.

RELATED TOPICS

[Artificial Intelligence](#)

[Digital Platforms](#)

[Information Technology](#)

[Risk Management](#)

The Governance Failure Nobody Planned For

Over the past decade, organizations have invested heavily in artificial intelligence to improve efficiency, insight, and decision-making. Early deployments focused on prediction engines, recommendation systems, and, more recently, generative AI tools that assist employees with writing, coding, and analysis. These systems were largely framed as *tools*: powerful, but ultimately subordinate to human judgment. In recent years, that framing quietly broke down.

Enterprises are now deploying AI systems that do not merely assist work but *perform it*. Autonomous agents monitor markets, negotiate with vendors, route logistics, approve transactions, remediate IT incidents, and coordinate with other software agents often without human intervention. These systems perceive their environment, reason over goals, take actions through enterprise systems, and collaborate with other agents to achieve outcomes. In organizational terms, they have crossed a threshold from tools to actors.

This shift has exposed a managerial blind spot. Most firms govern autonomous AI using control mechanisms designed for deterministic software or transactional analytics models. Security teams frequently prioritize **infrastructure perimeters**, which can be bypassed by agents using legitimate credentials to execute unintended, non-deterministic actions. Similarly, while some firms treat compliance as a core value, many risk functions still rely on static checklists, and IT departments frequently treat agents as standard applications. These approaches struggle to account for agents that independently perceive, decide, and act at machine speed. These approaches are increasingly inadequate. When autonomous agents operate at machine speed, interact non-deterministically, and exercise delegated decision rights, failures transcend traditional software bugs. Because these components have transitioned from tools to actors, their failures resemble **unpredictable organizational breakdowns** that are significantly more difficult to explain or remediate using standard technical logic.

Recent incidents across several sectors illustrate that when autonomous agents operate at machine speed, failures resemble organizational breakdowns rather than simple software bugs. In 2024, the **Moffatt v. Air Canada** case established a critical legal precedent: organizations are held liable for the “non-deterministic” promises made by their autonomous agents, even when those actions contradict internal policy. Furthermore, the **DPD “Rogue” Chatbot incident**⁸ demonstrated how a lack of real-time behavioral monitoring allows agents to deviate into “unintended” action such as criticizing their own firm once a system update alters their reasoning boundaries.

More technically concerning is the rise of **indirect prompt injection**, such as the **“EchoLeak” vulnerability**⁹. In this scenario, malicious instructions embedded in external data sources were used to manipulate an agent’s legitimate credentials to exfiltrate internal data, bypassing traditional perimeter defenses. These cases prove that governance is no longer a post-deployment checklist but a requirement for real-time control.

The core challenge, therefore, is no longer how to make AI systems more intelligent. It is how to *govern* software that can independently decide and act at scale. This article argues that autonomous AI requires a fundamentally new operating model, one that treats agents

as organizational actors embedded within explicit structures of coordination, control, and accountability.

To address this gap, the article makes three contributions. First, it clarifies what distinguishes agentic AI from earlier generations of automation and generative tools, emphasizing why autonomy changes the nature of managerial responsibility. Second, it introduces the **Agentic Operating Model (AOM)**, a layered framework that explains how enterprises can design, deploy, and govern autonomous agents at scale. Third, it examines the implications of this model for senior leaders, reframing AI governance as a source of operational resilience rather than a constraint on innovation.

From Tools to Actors: What Makes AI Agentic

For much of its history in organizations, artificial intelligence has been framed as a form of decision support. Predictive models scored risks, recommendation systems suggested options, and generative tools produced drafts for human review. Even when outputs were sophisticated, responsibility remained firmly with human decision-makers. The introduction of autonomous agents disrupts this arrangement.

Agentic AI differs from prior systems along three dimensions: autonomy, persistence, and delegation. Autonomous agents do not merely respond to prompts; they initiate actions based on environmental signals. Persistence allows agents to operate continuously over time, learning from feedback and adapting behavior without repeated human instruction. Delegation grants agents formal authority to act on behalf of the organization, including access to systems of record and the ability to commit resources.

Together, these characteristics transform AI systems into organizational actors. Like human employees, agents operate within defined roles, pursue assigned objectives, and interact with others to complete work. Also like humans, their behavior is non-deterministic and context-sensitive. This combination complicates oversight, as managers cannot rely on exhaustive rules or static testing to anticipate all possible actions.

The shift from tools to actors also alters accountability. When a spreadsheet produces an error, responsibility lies with the analyst who used it. When an autonomous agent approves a transaction or reroutes a shipment, responsibility is often ambiguous. Was the failure caused by the model, the data, the configuration, or the delegation decision itself? Without an explicit operating model, organizations struggle to answer these questions consistently.

Recognizing agents as actors clarifies why governance must move beyond technical safeguards. Just as firms establish policies, reporting structures, and controls for human workers, they must design institutional arrangements for digital ones. The Agentic Operating Model introduced in the previous section provides a foundation for this shift by embedding autonomy within explicit layers of coordination, control, and governance.

The Agentic Operating Model (AOM)

As organizations experiment with autonomous agents, many encounter the same pattern: individual agents perform well in isolation, yet the overall system behaves unpredictably when deployed at scale. This gap reflects a mismatch between the complexity of agentic systems and the operating models used to manage them. Traditional IT operating models assume deterministic behavior, centralized control, and clearly bounded applications. Agentic systems violate all three assumptions.

To address this challenge, this article proposes the **Agentic Operating Model (AOM)** a governance-centric framework that specifies the minimum structural components required to operate autonomous AI responsibly and effectively at enterprise scale. The AOM consists of four interdependent layers: the Cognitive Layer, the Coordination Layer, the Control Layer, and the Governance Layer. Each layer addresses a distinct managerial problem, and failure in any one undermines the stability of the entire system.

The Cognitive Layer: Specialized Intelligence

The Cognitive Layer defines how intelligence is instantiated within the organization. Rather than relying on a single, general-purpose model, agentic enterprises increasingly deploy multiple specialized models embedded within autonomous agents. These models are optimized for specific domains, tasks, and performance constraints.

This specialization is not merely a technical optimization; it is a governance choice. Smaller and domain-specific models are easier to evaluate, constrain, and audit than monolithic systems trained on broad, opaque data sources. They reduce hallucination risk in regulated domains and enable clearer alignment between an agent's capabilities and its delegated responsibilities. In the AOM, intelligence is deliberately fragmented to make accountability tractable. While this approach requires greater investment in domain-specific training and model management compared to monolithic systems, it avoids the 'generality risk' where broad decision authority leads to unpredictable outcomes. For general enterprise, the tradeoff of higher initial complexity is offset by the ability to precisely evaluate, constrain, and audit agents in regulated or high-risk domains.

Organizations that neglect this layer often conflate autonomy with generality, embedding broad decision authority into overly capable models. The result is agents whose behavior is difficult to predict and even harder to justify after the fact. By contrast, firms that design cognitive specialization as an operating principle create agents whose scope of action is intelligible by design.

The Coordination Layer: From Hierarchies to Swarms

The Coordination Layer governs how agents interact with one another to accomplish complex tasks. While early systems relied on centralized "Hub-and-Spoke" orchestration, modern enterprises are shifting toward **Swarm Intelligence**, where agents operate via decentralized local rules and shared goals without a single point of failure.

Real-World Applications of Agentic Coordination

As agentic deployments mature, the transition to decentralized collaboration has moved from theory to core enterprise operations:

- **Multi-Agent Insurance Swarms:** In the insurance sector, firms have moved beyond manual task-routing to “collaborative multi-agent teams”. A single claim may be processed by a swarm of seven specialized agents including Planner, Coverage, and Fraud agents communicating through a shared environment to verify policies and weather data simultaneously. This shift has enabled leaders like **Lemonade** to process approximately one-third of claims autonomously, with its “AI Jim” agent achieving settlements in as little as three seconds.⁶
- **Autonomous Logistics “Orchestras”:** Global logistics leaders like **Maersk** and **Unilever** are utilizing agentic meshes to respond to real-time disruptions. Maersk’s “Project Autosub” utilized autonomous vessel agents that coordinate route optimization and port scheduling without human intervention, achieving a **23% reduction in fuel consumption**. Similarly, **Unilever** uses reactive swarms to autonomously negotiate with carriers and reorganize warehouse logistics during shipping delays.^{1,2,5}
- **Decentralized Financial Consensus:** High-frequency trading and fraud detection at firms like **J.P. Morgan** and **Goldman Sachs** now utilize multi-agent systems (MAS) where agents analyze market signals in parallel. These systems employ “consensus mechanisms” to prevent rogue actions, requiring multiple agents to agree on high-risk capital commitments before execution.^{4,7}

The Managerial “Orchestration Gap”

This shift introduces what I term the “**Orchestration Gap**”: a mismatch where decentralized software outpaces centralized human management. In the Agentic Operating Model (AOM), coordination is treated as an explicit design decision rather than an emergent property.

Leaders must evolve from task supervisors to “**Switchboard Operators,**” defining the ethical boundaries and goals for the entire mesh rather than specific workflows. Importantly, when no single agent is “in charge,” governance mechanisms must be embedded within the coordination protocol itself. In practice, this takes the form of **programmable constraints**. For example, in Financial agent swarms, a coordination protocol might include a **Consensus Mechanism** that physically prevents any single agent

from executing a transaction unless some other independent agents (e.g., Risk, Compliance, and Audit agents) sign off on the telemetry. Here, accountability is not a manual review process but a hardcoded requirement for system execution.

The Control Layer: Constraining Autonomous Action

The Control Layer defines how agent behavior is bounded in real time. Traditional controls such as role-based access and static permissions are insufficient when agents generate novel actions in dynamic environments. Agentic systems require adaptive controls that respond to context, confidence, and risk.

Key mechanisms in this layer include confidence thresholds, behavioral baselines, and guardrail agents that monitor inputs and outputs. For example, a ‘Guardrail Agent’ can be implemented as a lightweight model that intercepts a primary agent’s output before it reaches a system of record. If a Procurement Agent initiates a \$50,000 vendor payment exceeding its \$10,000 ‘behavioral baseline’ the Guardrail Agent triggers a ‘Confidence Threshold’ check. If the agent’s internal reasoning score is below 95%, the action is physically blocked and escalated for Human-on-the-Loop review. Rather than approving every action, these controls intervene selectively when uncertainty or potential impact exceeds predefined limits. This approach supports a shift from Human-in-the-Loop oversight to Human-on-the-Loop supervision, where humans set boundaries and intervene only when necessary.

Organizations that underinvest in the Control Layer often rely on informal supervision or post hoc audits. Such approaches fail at scale, allowing small errors to propagate rapidly across interconnected agents. Effective control does not eliminate autonomy; it makes autonomy survivable.

The Governance Layer: Accountability and Legitimacy

The Governance Layer anchors the AOM by assigning accountability for agentic behavior and aligning it with organizational and regulatory expectations. This layer encompasses policies, standards, and decision rights that define who is responsible for an agent’s

actions throughout its lifecycle.

Frameworks such as ISO/IEC 42001 and the NIST AI Risk Management Framework provide structural guidance, but governance is not achieved through compliance alone. In the AOM, each agent is associated with a clear business owner, a defined risk profile, and documented decision boundaries. Outputs are traceable to specific model versions, configurations, and prompts, enabling post hoc explanation and audit.

Without this layer, autonomous agents become organizational orphans capable of acting but owned by no one. Such systems may deliver short-term efficiency gains while accumulating long-term operational and reputational risk. Effective governance restores legitimacy by ensuring that autonomy is always coupled with responsibility.

Why the Layers Must Work Together

The four layers of the Agentic Operating Model are mutually reinforcing. Specialized intelligence simplifies control. Coordination choices determine governance complexity. Controls operationalize governance principles. Governance, in turn, constrains how intelligence is deployed. Organizations that address these layers in isolation often experience brittle systems that fail under stress, this is a pattern that is observed when technical autonomy is granted without corresponding control thresholds or accountability structures. As detailed in the below mentioned enterprise vignettes, failures typically stem from this misalignment across layers rather than from deficiencies in model performance itself.

The AOM reframes autonomous AI as an institutional design problem rather than a technology project. By making operating assumptions explicit, it allows leaders to reason systematically about how autonomy is granted, constrained, and supervised. In doing so, it provides a foundation for scaling agentic AI without surrendering managerial oversight.

Governing Autonomous Actors: From Human-in-the-Loop to Human-on-the-Loop

A central tension in agentic systems is the relationship between autonomy and oversight . Early governance approaches emphasized **Human-in-the-Loop (HITL)** controls, requiring manual human approval for critical actions . While effective in low-volume settings, HITL becomes a bottleneck as enterprises execute thousands or millions of agentic actions per hour . Consequently, organizations are shifting toward **Human-on-the-Loop (HOTL)** supervision, where humans define objectives, constraints, and escalation thresholds, while agents operate independently within those boundaries .

From Reactive Audits to Proactive Controls

Effective HOTL governance depends on **proactive controls** rather than reactive audits . Because agents can be manipulated through indirect prompt injection or enter unintended feedback loops at machine speed, organizations can no longer afford to wait for a “post-mortem” log review .

- **“Safe-Action” Pipelines & Infrastructure:** To prevent the “Unbounded Agent” failure mode, enterprises are adopting **Safe-Action Pipelines**. This reflects the move toward HOTL supervision where high-risk actions such as the **DPD “Rogue” Chatbot** incident or the **Moffatt v. Air Canada** hallucinations are physically blocked at the system level if they exceed predefined “blast radius” or confidence thresholds .
- **Digital Provenance and Accountability:** Traceability is ensured through digital provenance mechanisms, allowing organizations to reconstruct how and why a particular outcome occurred . This shift is supported by **ISO/IEC 42001** and the **NIST AI Risk Management Framework**, which mandate continuous monitoring and lifecycle responsibility rather than point-in-time compliance .

Institutionalizing Authority

Crucially, governance must be embedded by design rather than layered on after deployment . Assigning each agent a clear business owner, risk classification, and escalation protocol ensures that accountability remains intact even as autonomy increases . By reframing oversight as **supervision rather than approval**, the HOTL model reconciles the speed of the “Agentic Mesh” with the necessity of managerial control . It transforms governance from a bureaucratic bottleneck into a prerequisite for responsible innovation at scale .

Enterprise Vignettes: How Agentic Systems Succeed and Fail

The managerial challenges of agentic AI become most visible when autonomous systems interact with real organizational constraints. The following vignettes are illustrative composites drawn from common enterprise patterns and real-world failure modes. While the narratives are synthesized to highlight structural misalignments, they are grounded in documented historical incidents such as the **Moffatt v. Air Canada [3]** legal precedent regarding non-deterministic agent promises and the **DPD ‘Rogue’ Chatbot[8]** incident where a lack of real-time monitoring allowed an agent to deviate from firm policy. These cases serve as the empirical basis for the ‘Unbounded Agent’ and ‘Compliant Failure’ modes described below.

Vignette 1: The Unbounded Agent

An enterprise deploys an autonomous operations agent tasked with resolving routine service disruptions. The agent has broad system access and a general-purpose language model to interpret logs and remediation options. Initially, performance improves dramatically. Over time, however, the agent begins executing increasingly complex interventions, including configuration changes that exceed its original mandate.

When a major outage occurs, post-incident review reveals that no clear decision boundary constrained the agent’s authority. The Cognitive Layer favored generality over specialization, while the Control Layer relied on static permissions rather than dynamic

thresholds. Although the agent behaved “correctly” according to its internal logic, the organization lacked a governance mechanism to prevent scope creep. Under the AOM, tighter cognitive specialization and explicit control thresholds would have limited escalation while preserving autonomy for routine tasks.

Vignette 2: The Invisible Swarm

A second organization experiments with decentralized agent collaboration to improve internal coordination **of complex operational workflows**. Unlike human coordination (e.g., manual meeting scheduling or routing sub-tasks via email), this ‘swarm’ involves multiple agents that monitor real-time data streams, update shared state, and trigger cross-system action such as **autonomously synchronizing inventory levels with procurement orders and shipping schedules** without centralized orchestration. While resilient to individual agent downtime, the architecture generates unexpected outcomes when agents respond to partial or outdated information.

In one instance, several agents independently initiate compensating actions in response to the same signal, amplifying rather than resolving the issue. Investigation is hindered by the absence of a clear ownership model. No single team claims responsibility for the collective behavior of the swarm.

This vignette highlights the importance of the Coordination and Governance layers working in tandem. While decentralized collaboration increases robustness by removing single points of failure, it requires **Conflict Resolution Protocols** to manage contradictory agent actions. The AOM emphasizes that ‘resilience through redundancy’ must be functionally paired with **Auditable Ownership**. In practice, this means every autonomous action within the swarm must be linked to a specific business owner and risk profile. Without this linkage, the speed of the mesh creates ‘organizational orphans’ where no human team is responsible for the collective outcome.

Vignette 3: The Compliant Failure

A third enterprise invests heavily in formal AI governance, documenting policies, approvals, and compliance artifacts aligned with external standards. Autonomous agents are certified prior to deployment and reviewed periodically. Despite this rigor, the organization experiences repeated near-misses involving inappropriate agent actions.

The issue lies not in the absence of governance but in its implementation. Oversight focuses on pre-deployment checklists rather than real-time supervision. Agents operate without behavioral monitoring, and escalation protocols are rarely triggered because thresholds are poorly calibrated. Governance exists on paper but not in operation.

This scenario illustrates why governance must be embedded within the Control Layer rather than treated as an external audit function. The AOM reframes compliance as a continuous activity that shapes how agents behave in practice, not merely how they are approved.

Implications for Senior Leaders

The adoption of agentic AI has implications that extend beyond technology management. By redistributing decision-making authority from humans to autonomous systems, agentic enterprises reshape roles, responsibilities, and risk across the organization.

For chief executives, the primary concern is operational resilience. Autonomous agents can sense and respond to change faster than human teams, but they also introduce new dependency risks. Leaders must ensure that critical processes remain intelligible and recoverable when agentic systems fail. The Agentic Operating Model provides a way to balance speed with stability by making autonomy an explicit design choice rather than an implicit byproduct of capability.

Chief financial officers face a different challenge: the emergence of variable cognitive costs. Unlike traditional software licenses, the cost of agentic AI scales with usage, interaction frequency, and model complexity. Without disciplined operating assumptions, enterprises risk deploying agents that are economically irrational even when technically effective. By aligning cognitive specialization with delegated authority, the AOM supports more transparent cost governance **by enabling granular cost attribution**. When agents

are specialized for specific tasks, leaders can move from opaque ‘API-call bundles’ to a **unit-cost model**, where the expense of a specific model (e.g., a high-reasoning model for fraud) is directly mapped to the business value of its delegated domain. Transparency arises from the ability to see exactly which business functions are driving cognitive spend.

For chief information officers, integration and proliferation are central concerns. Autonomous agents interact continuously with systems of record, often at volumes that exceed legacy design assumptions. At the same time, the ease of deploying agents encourages experimentation outside formal IT channels. An explicit operating model helps CIOs provide secure, scalable pathways for innovation while reducing the risks associated with unmanaged agent sprawl.

Across roles, the common implication is that governance is no longer a constraint on innovation but a prerequisite for sustaining it. Enterprises that invest in operating models rather than ad hoc controls are better positioned to scale autonomy without sacrificing trust.

Conclusion: From Intelligent Systems to Institutional Design

The rise of autonomous AI marks a transition from intelligent tools to digital actors embedded within organizations. This shift challenges long-standing assumptions about control, accountability, and managerial oversight. As autonomy increases, technical excellence alone is insufficient. What determines success is the quality of the institutional structures surrounding agentic systems.

The Agentic Operating Model reframes autonomous AI as an organizational design problem. By articulating the cognitive, coordination, control, and governance layers required to operate agents responsibly, it offers leaders a systematic way to reason about autonomy at scale. Rather than asking whether agents are capable, the model asks whether they are governable.

Firms that address this question proactively can harness agentic AI as a durable source of advantage. Those that do not may find themselves managing systems that act decisively yet remain fundamentally unaccountable. In an era where software increasingly performs work once reserved for humans, the future of competitive advantage lies not in intelligence alone, but in the institutions that shape how intelligence is exercised.

References

1. Sushree Swagatika Pati, “**Agentic AI in Supply Chain 2025: Autonomous Decision Making**,” May 14, 2025.
2. Eva Richardson, “**Maersk Launches AI-Powered Vessel Routing Platform to Cut Emissions**,” EAN Network, April 18, 2025.
3. Barry B. Sookman, “**Bereavement Fares and Chatbot Liability**,” *Moffatt v. Air Canada*, February 19, 2024.
4. Gizel Gomes, “**AI in Banking: JP Morgan Leads the AI Sphere**,” September 3, 2024.
5. Graham Sommer et al., “**How Unilever is envisioning the Autonomous Supply Chain with Agentic AI**,”.
6. Ancil Mohamed, “**Generative AI in Insurance: Lemonade Case Study**,”.
7. Fei Xiong et al., “**QuantAgent: Price-Driven Multi-Agent LLMs for High-Frequency Trading**,” September 27, 2025.
8. Tom Gerken, “**DPD error caused chatbot to swear at customer**”, January 19, 2024.
9. Lexi Croisdale, “**EchoLeak in Microsoft Copilot: What it Means for AI Security**,” June 12, 2025.



Sandeep Saini [Follow](#)

Sandeep Saini is a Technical Lead at Google with 15 years’ experience in high-scale distributed systems. He currently leads AI infrastructure and governance initiatives, focusing on standardizing agent identities and architectural frameworks. His work centers on balancing rapid AI deployment with robust, scalable governance across global product ecosystems.