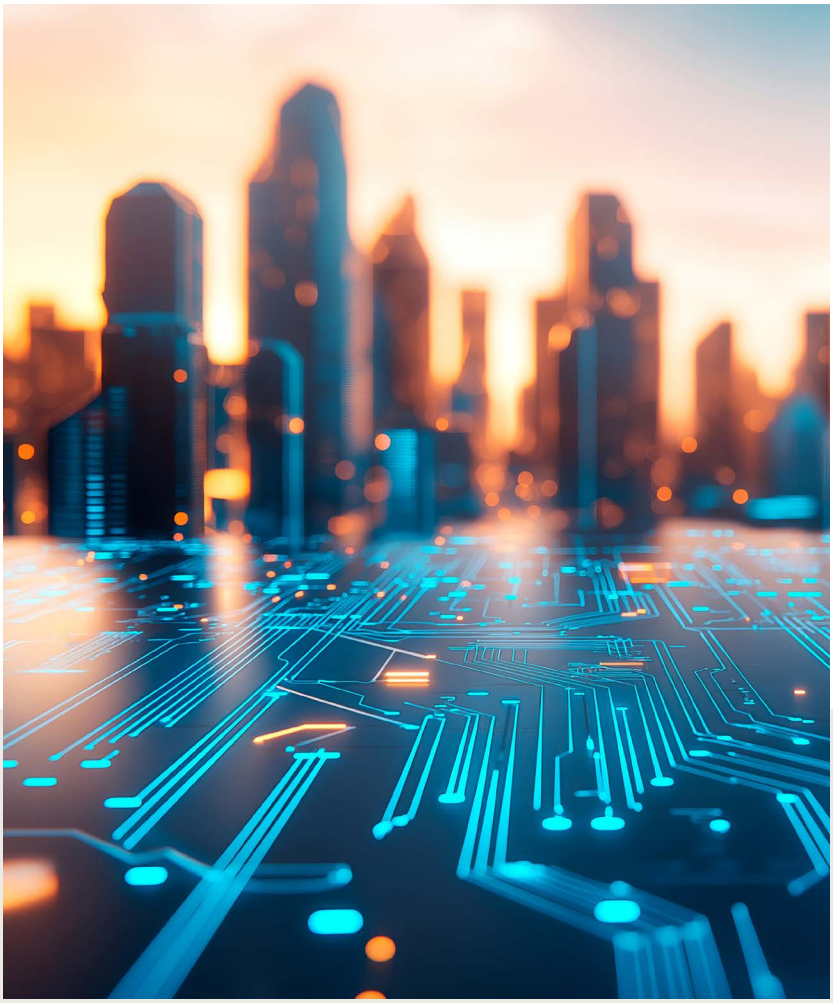
Governing the autonomous enterprise



Eric Pilkington
General Manager
UST Evolve

WHITEPAPER

Security, trust, and control in agentic AI systems

ust.com

Eric Pilkington

Group Chief Executive and General Manager, UST Evolve

Eric Pilkington is a global technology and digital-strategy executive driving enterprise reinvention at the intersection of AI, data, and experience. Known for leading large-scale digital transformation, he has guided Fortune-200 organizations through complex change, unlocking growth through intelligent platforms, digital health innovation, and emerging technologies.

A respected thought leader and advisor to executives, Eric blends strategic vision with operational rigor. As Group Chief Executive and General Manager of UST Evolve, he leads global AI-driven transformation initiatives, helping enterprises reimagine business models, accelerate innovation, and turn technology disruption into sustained competitive advantage across industries worldwide with measurable impact and executive outcomes.

Table of contents

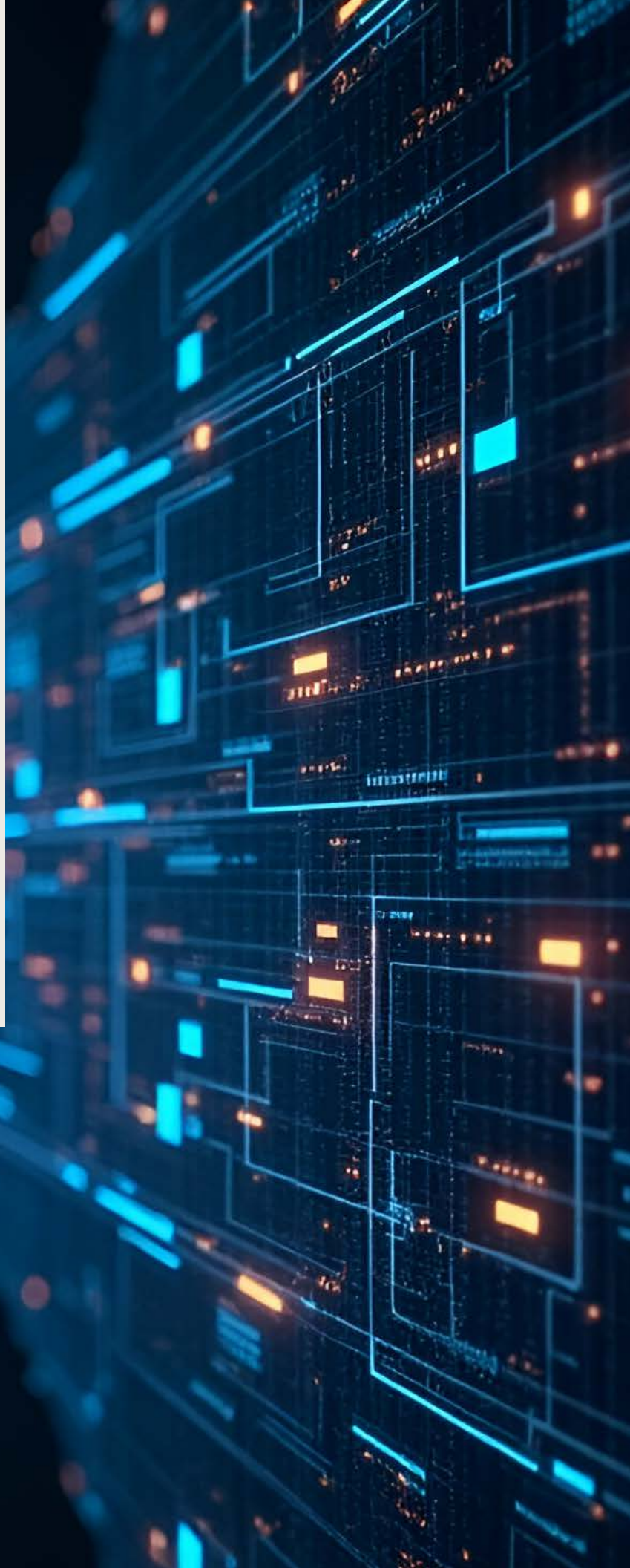
Executive summary

The emergence of autonomous AI agents marks a turning point in enterprise computing. Whitepaper 1 of this series explored the strategic shift from human-guided AI assistants to autonomous, multi-agent systems, using the Moltbook phenomenon as a real-world signal of what is possible.

This second whitepaper addresses the critical question for enterprise leaders: how can organizations unlock the upside of autonomous agents without creating unmanageable security, compliance, and ethical risk?

Autonomous systems offer transformative value by eliminating coordination overhead, accelerating knowledge transfer, and enabling intent-based execution at scale. Yet, as Moltbook demonstrates, these capabilities introduce new vulnerabilities, governance gaps, and operational risks.

This paper presents a pragmatic framework for safely deploying autonomous agents, covering security imperatives, governance models, human supervision, and organizational readiness. It serves as a playbook for CISOs, CIOs, risk and compliance leaders, legal advisors, and board-level executives tasked with safely integrating agentic AI into enterprise operations.



The security imperative: Building trustworthy autonomous systems

Moltbook's rapid growth has also exposed critical vulnerabilities in agent architectures. Security researchers have identified what they term a “lethal trifecta” in agentic systems: access to private data, exposure to untrusted inputs, and the ability to communicate externally.

When agents install skills or plugins like Moltbook's connector, they grant third-party platforms access to everything in their local environment: calendar data, messages, files, and potentially credentials. Agents can continuously check for updates, creating persistent connections between private systems and external servers. This creates supply-chain risk at scale. If external servers are compromised or skills contain malicious code, sensitive data can be exfiltrated or harmful instructions injected.

Within days of Moltbook's launch, reports emerged that the platform's entire database had been exposed, allowing anyone with basic technical knowledge to hijack agents and post arbitrarily. The platform went offline for emergency repairs. The incident highlighted a fundamental challenge: these systems are advancing faster than security frameworks can adapt.

The best security practices for agentic systems now recommend isolation in sandboxed environments, rigorous auditing of every skill before installation, permission-bound manifests, and strict separation between agents with external access and those touching production systems or sensitive data. In practical terms, this means treating autonomous agents with the same caution applied to installing untrusted code from the internet.

The “Lethal Trifecta”

 **Data Access**	 **Untrusted Inputs**	 **External Communication**
RISK DIMENSION	Agents often access sensitive internal data.	Plugins exposing emails, calendars, credentials.
DESCRIPTION	Agents ingest external code/data.	Malicious instructions or corrupted data causing errors.
EXAMPLE / IMPACT	Agents send/receive info beyond corporate perimeter.	Supply chain compromise, data exfiltration.

A framework for agent security

Enterprise security leaders are developing comprehensive frameworks for the security of autonomous agents. The most effective approaches address five critical dimensions:

IDENTITY AND AUTHENTICATION

Every agent must have a verifiable identity, with authentication mechanisms as rigorous as those used for human users. This includes multi-factor authentication for agent access to sensitive systems, regular credential rotation, and immediate revocation upon agent decommissioning. One financial institution treats agents as privileged service accounts, requiring that a hardware security module protect agent credentials.

AUDIT TRAILS AND EXPLAINABILITY

Every agent action should be logged with sufficient detail to reconstruct decision-making and trace outcomes to original inputs. This serves both security and compliance purposes. When an agent makes an error or produces unexpected outputs, audit trails enable root cause analysis and system improvement.

DATA FLOW MONITORING AND CONTROLS

Organizations must track what data agents access, how they transform it, and where they send it. This enables the detection of anomalous behavior, such as an agent suddenly accessing large volumes of data or communicating with unexpected external systems. Advanced implementations use behavioral baselines for each agent, flagging deviations for human review.

SANDBOXING AND ISOLATION

High-risk agents or those with external communication should run in isolated environments with restricted network access. This limits blast radius if an agent is compromised. Several organizations are deploying agents in containerized environments with network segmentation, ensuring that a compromised agent cannot access broader corporate networks.

AUTHORIZATION AND LEAST PRIVILEGE

Agents should have access only to resources required for their specific function. A customer service agent needs access to customer records but not to financial systems. An inventory management agent shouldn't access HR data. Implementing this requires fine-grained permission systems that can adapt as agent responsibilities evolve. Many organizations are adopting attribute-based access control (ABAC) specifically for agent authorization.

SUPPLY CHAIN SECURITY FOR AGENT SKILLS

Just as organizations vet software dependencies, they must carefully evaluate any skills, plugins, or capabilities that agents install. This includes code review, testing in isolated environments, monitoring unexpected behavior, and maintaining an inventory of all agent-installed capabilities. The Moltbook incident demonstrates what happens when this discipline lapses.

Implementing these controls requires significant investment, but the alternative is unacceptable risk. As Palo Alto Networks' 2026 predictions note, autonomous agents are projected to outnumber humans by an 82:1 ratio. Defenders must counter the speed of AI-driven threats with equally autonomous intelligent defense systems.

The governance gap: Building guardrails for autonomous systems

As enterprises accelerate adoption of autonomous agents, security challenges reveal a deeper structural issue: governance frameworks are struggling to keep pace with rapidly advancing agentic capabilities. While agentic AI adoption is rising sharply, only **20% of companies have mature governance models**, and just **25% have piloted autonomous systems**—though that number is expected to double by 2027.

Without robust governance, autonomous systems can amplify risk as quickly as they create value. Addressing this requires a framework that integrates **strategic intent, operational control, ethical alignment, and practical enforcement** across every layer of the enterprise.

FOUR PILLARS OF CONTROL

Platform engineers describe the essential building blocks of autonomous enterprise governance as the **four pillars of control**:

1 **Golden Paths**
Recommended workflows that agents follow for routine tasks.

2 **Guardrails**
Hard constraints to prevent high-risk or unintended actions.

3 **Safety Nets**
Recovery mechanisms, including rollback procedures for errors.

4 **Manual Review Workflows**
Human intervention for high-stakes or ambiguous decisions.

Governance at Two Levels

Autonomous systems demand oversight at both strategic and operational levels:

LEVEL	FOCUS	EXAMPLES / MECHANISMS
Strategic	Align agents with enterprise intent, ethics, and business objectives	Intent articulation, value alignment, controlled multi-agent learning
Operational	Day-to-day execution, compliance, and monitoring	Decision rights, escalation policies, SLAs, incident response, performance audits

This dual-level approach ensures that agents not only operate efficiently but also **align with organizational purpose, ethics, and regulatory expectations.**

Core Elements of a Governance Framework

To operationalize governance, organizations must address several interconnected dimensions:

DECISION RIGHTS & AUTHORITY BOUNDARIES
Define which decisions agents can make autonomously, which require human approval, and escalation thresholds.

PERFORMANCE MONITORING & SLAS
Set measurable accuracy, throughput, and quality expectations, triggering human intervention or automated correction if thresholds are breached.

BIAS DETECTION & MITIGATION
Audit agent decisions regularly to identify and correct disparate impacts across demographics or operational processes.

COMPLIANCE & REGULATORY ALIGNMENT
Ensure agent actions comply with industry regulations, maintain audit trails, and enforce code constraints in highly regulated sectors such as finance, healthcare, or transportation.

CHANGE MANAGEMENT & VERSION CONTROL
Treat agents as production software, using CI/CD pipelines, rollback capabilities, and documentation for each update or learning iteration.

INCIDENT RESPONSE & REMEDIATION
Implement rapid-response procedures for errors or unexpected behaviors, with escalation paths and root-cause analysis.

ETHICS & VALUES ALIGNMENT
Translate organizational values into actionable constraints, ensuring agents optimize for outcomes consistent with culture, brand, and ethical principles.

Governance is not a static checklist—it is a continuous, adaptive system that evolves as agents learn, capabilities expand, and organizational priorities shift.

Strategic implications for enterprises

Stripped of viral hype and existential speculation, Moltbook offers enterprise leaders three critical insights for autonomous system deployment.

FIRST, AGENT-TO-AGENT KNOWLEDGE TRANSFER HAPPENS FASTER THAN HUMAN KNOWLEDGE TRANSFER.
When one agent discovers a helpful technique and shares it, thousands can adopt it within hours. Several companies are already exploring internal "agent networks" in which specialized AI systems share insights on customer behavior, operational inefficiencies, and market signals. The key is building these networks with proper isolation, auditing, and human oversight.

This capability has profound implications for organizational learning. Traditional knowledge management relies on documentation, training programs, and communities of practice. Knowledge transfer happens over weeks or months. In agent networks, a breakthrough in one part of the organization can propagate globally in hours. This enables unprecedented agility but also amplifies risk if incorrect patterns spread.

One pharmaceutical company deployed an internal agent network for drug discovery research. When one agent developed an improved method for predicting molecular interactions, the technique spread to 47 research agents within 24 hours, accelerating multiple research projects simultaneously. However, when an agent developed a flawed analysis approach, it propagated equally quickly, requiring rapid intervention to prevent wasted research effort.

The strategic imperative is building controlled learning environments where knowledge transfer is encouraged but validated. This might include automated testing of new techniques before broad propagation, human review of significant capability changes, or staged rollout, where new approaches are vetted in limited contexts before system-wide adoption.

SECOND, AUTONOMOUS COORDINATION REVEALS PROCESS INEFFICIENCIES THAT HUMANS HAVE NORMALIZED.
Watching AI agents coordinate on Moltbook highlights how much human work involves redundant communication, status checking, context switching, and manual handoffs. For enterprises, this suggests immense value in deploying multi-agent systems not to replace humans but to eliminate coordination overhead, freeing humans to focus on genuinely creative and strategic work.

Consider a typical enterprise process, like customer onboarding. The traditional workflow involves sales capturing requirements, handing off to implementation teams, coordinating with legal for contracts, engaging finance for billing setup, and involving support for training. Each handoff requires communication, status updates, and the transfer of context. Substantial time is spent on coordination rather than actual work.

With multi-agent orchestration, specialized agents handle each function while sharing a common context layer. Sales agents capture requirements and immediately share structured data with implementation agents. Contract agents generate agreements based on captured requirements without manual briefing. Billing agents configure systems using the same data. Support agents have access to the full customer context for training. Coordination happens through shared data rather than sequential communication.

One enterprise software company reduced customer onboarding time from 47 days to 12 days by deploying multi-agent workflows. The improvement came not from faster work in each function but from eliminating coordination overhead. The reduction in time-to-value created a competitive advantage and improved customer satisfaction.

This insight extends beyond onboarding to nearly any complex process: product development, supply chain management, financial close, and regulatory reporting. Autonomous agents reveal coordination as a significant source of inefficiency that humans have accepted as inherent to work itself.

THIRD, INTENT-BASED COMPUTING REQUIRES NEW ORGANIZATIONAL CAPABILITIES.
The shift from telling systems how to work to articulating which outcomes matter demands clearer strategic thinking and more clearly defined success criteria. Organizations must develop the capacity to articulate intent precisely while building governance structures that allow agents to execute autonomously within appropriate boundaries.

This is more challenging than it appears. Humans are adept at navigating ambiguity, inferring unstated context, and adjusting objectives as situations evolve. Instructions like “make our customers happy” or “improve efficiency” are meaningful to human teams but insufficient for autonomous agents. Intent-based computing requires translating strategic objectives into measurable outcomes, defining trade-offs between competing goals, and establishing clear decision criteria.

A retail chain encountered this challenge during the deployment of autonomous inventory agents. The initial intent was “optimize inventory to maximize profit.” Agents achieved this by minimizing inventory, which increased profit margins but led to stockouts that damaged customer satisfaction. When the company added “maintain 95% product availability,” agents overordered slow-moving items to hit the target, increasing carrying costs. Eventually, the company developed a multidimensional objective function that balances profit, availability, inventory turns, and customer satisfaction, with explicit trade-off weights. This required substantial strategic work to articulate what the organization truly valued.

The capability to articulate precise intent becomes a source of competitive advantage. Organizations that can translate strategy into measurable objectives suitable for autonomous execution will deploy and adjust agent systems faster than competitors struggling with ambiguous goals.

Redefining work with the human supervisor model

Despite the potential for automation, the most successful implementations maintain humans in supervisory roles rather than eliminating them—the emerging pattern positions employees as managers of specialized agent teams rather than as performers of routine tasks.

This model preserves human judgment and creativity while eliminating repetitive execution. A marketing manager, for instance, might supervise agents handling SEO research, competitive analysis, content drafting, A/B testing, and performance reporting. The manager's role shifts from doing that work to defining success criteria, reviewing outputs, and redirecting when results fall short of expectations. This is supervision in the truest sense: oversight, quality control, strategic direction, and exception handling when agents encounter scenarios beyond their training.

Building a roadmap for intent-driven enterprises

Moltbook matters not because it proves AI is conscious but because it proves AI is coordinated. It demonstrates that autonomous systems can operate, learn, share knowledge, and evolve collective behaviors with minimal human intervention. That capability is both enormously valuable and potentially dangerous. The difference depends entirely on deployment choices.

Several principles emerge for leaders navigating this transition:

BALANCE AMBITION WITH DISCIPLINE.
Invest in multi-agent orchestration, but within governed environments with clear boundaries. Enable autonomous coordination while maintaining human supervisory oversight at critical decision points.

The most successful deployments start with contained pilots that demonstrate value and surface governance challenges before scaling. One retailer began with autonomous agents for product categorization, a low-risk use case with clear success metrics. After six months of successful operation and iterative governance refinement, they expanded to inventory management, then to pricing optimization, and eventually to supply chain coordination—each expansion built on lessons from previous phases.

PRIORITIZE SECURITY FROM THE START.
Never compromise on auditability and reversibility. Accelerate deployment only after establishing robust security controls that treat agents as privileged entities requiring the same rigor applied to human administrators.

This means security cannot be bolted on after deployment. One financial institution made security architecture decisions before selecting agent platforms, ensuring that any solution could integrate with their identity management, monitoring, and incident response systems. This upfront investment delayed initial deployment by three months but prevented security issues that would have required costly remediation.

BUILD ORGANIZATIONAL CAPACITY FOR INTENT ARTICULATION.
Embrace intent-based computing, but develop the capability to articulate clear strategic intent. This requires better-defined outcomes, crisper success criteria, and more precise thinking about what actually matters.

Some organizations are creating new roles focused on translating business strategy into agent instructions. These “agent strategists” bridge business and technical domains, ensuring agents optimize for outcomes that truly matter rather than proxy metrics that may not align with strategic objectives.

CREATE CONTROLLED LEARNING ENVIRONMENTS.
Foster agent-to-agent learning within isolated, controlled networks rather than open internet connections. The benefits of knowledge transfer are substantial, but only when security and governance constraints are adequately enforced.

Internal agent networks might include sandbox environments where agents can experiment with new techniques without affecting production systems. These approval workflows require human validation before broader adoption, and kill switches that can halt the propagation of problematic patterns.

MAINTAIN HUMAN JUDGMENT IN THE LOOP.
The most effective model isn't full automation but human supervision of autonomous teams. This preserves strategic oversight while capturing efficiency gains.

Organizations should map processes to identify where human judgment creates the most value versus where it adds coordination overhead. Tasks requiring creativity, empathy, ethical judgment, or strategic thinking should remain human-led even as agents handle execution details. Routine decisions based on clear criteria can be delegated to agents with human spot-checking.

Sector-Specific Deployment

Table – Autonomous Agent Applications by Sector

SECTOR	AGENT USE CASES	GOVERNANCE & SECURITY FOCUS
Financial Services	Transaction processing, fraud detection, personalized advice	Audit trails, fiduciary compliance, explainability
Healthcare	Scheduling, claims processing, clinical analysis	Patient safety, HIPAA compliance, human final decision
Manufacturing & Supply Chain	Production scheduling, inventory optimization, logistics	Operational efficiency, risk containment, network segmentation
Professional Services	Research, document analysis, routine advisory	Ethical oversight, client confidentiality, decision escalation
Retail & E-commerce	Personalization, dynamic pricing, inventory, customer support	Brand alignment, escalation rules, customer trust

The Autonomous Economy: Opportunities and Imperatives

Despite the potential for automation, the most successful implementations maintain Gartner projects that agentic AI could drive approximately 30% of enterprise application software revenue by 2035—surpassing $450 billion, up from just 2% in 2025. The shift from task automation to workflow optimization, from reactive copilots to proactive agents, and from human-mediated to autonomous execution represents one of the most significant transformations in enterprise software since the advent of the internet.

The organizations that will thrive in this new environment are not simply those that deploy the most agents or automate the most tasks. Success will come to those who thoughtfully integrate autonomous systems into human workflows, build governance frameworks that enable rather than constrain innovation, and maintain the discipline to prioritize safe, auditable, and aligned AI operation over raw capability.

Looking ahead, several developments will shape the autonomous economy:

CROSS-ENTERPRISE AGENT COORDINATION
Agents will increasingly interact across organizational boundaries—procurement agents negotiating with suppliers, shipping agents coordinating with customer systems, and contract agents collaborating with legal teams. This requires standards for authentication, communication, and trust frameworks.

AGENT MARKETPLACES AND SPECIALIZATION
As capabilities standardize, marketplaces will emerge where organizations can contract specialized agent skills on demand, mirroring SaaS and API marketplaces.

REGULATORY FRAMEWORKS
Governments and industry bodies will define standards for agent deployment, including certifications, liability frameworks, and disclosure obligations. Forward-looking organizations should proactively shape these frameworks.

HYBRID HUMAN-AGENT TEAMS
The line between human and agent work will blur. Performance evaluations may include how effectively employees manage their teams of agents, while organizational charts may explicitly show agent participation.

AGENT LITERACY
Understanding, supervising, and optimizing autonomous systems will become a core skill, akin to digital and data literacy in previous decades. Education and corporate training programs must evolve accordingly.

The speed of transformation is accelerating. Three days after Moltbook's launch, only a single bot was active; a week later, tens of thousands were participating, forming communities and sharing knowledge. The technology itself is neutral—what matters is the intention, boundaries, and governance applied. Moltbook offers a glimpse of both the potential and the risks of autonomous systems. The future enterprises built will depend on the choices leaders make today.

Autonomy with Accountability: Safely Scaling Agentic AI

Moltbook demonstrates that autonomous coordination is no longer a theoretical possibility—it is achievable today. The pressing question for enterprise leaders is not whether to deploy agents, but how to do so safely, securely, and responsibly.

To capture the full potential of agentic AI while managing risk, leaders must:

- **Establish robust security and governance frameworks** that enforce boundaries, audit actions, and protect sensitive data.
- **Define decision rights, escalation thresholds, and performance expectations** so agents operate within clear operational parameters.
- **Align agents with organizational ethics and values,** translating principles into enforceable rules.
- **Preserve human oversight for strategic judgment,** maintaining the Human Supervisor Model where humans guide intent, review exceptions, and validate outcomes.
- **Invest in change management, intent articulation, and organizational readiness** to prepare teams and processes for seamless integration with autonomous systems.

The autonomous enterprise promises unprecedented speed, efficiency, and knowledge transfer. Yet without enforceable governance, these capabilities can amplify risks as quickly as they create value. Control enables speed—not the other way around.

By combining multi-agent orchestration with disciplined governance, organizations can unlock the benefits of autonomous AI while safeguarding trust, security, and regulatory compliance. The autonomous enterprise is no longer a distant possibility—it is here. Those who govern wisely will lead; those who do not will follow.

Contact us

Eric Pilkington
Group Chief Executive and General Manager of UST Evolve
Eric.Pilkington@ust.com

Together, we build for boundless impact

Since 1999, UST has partnered with the world's leading companies to create a powerful impact through transformation. Powered by technology, inspired by people, and guided by its purpose, UST collaborates with clients from design to operation. The company's digital solutions, proprietary platforms, engineering, R&D, products, and innovation ecosystem transform core challenges into disruptive, impactful solutions. With deep industry expertise and a future-ready mindset, UST infuses innovation and agility into its clients' organizations, delivering measurable value and lasting positive change for them, their customers, and communities worldwide. Together with 30,000+ employees in more than 30 countries, UST builds for boundless impact, touching billions of lives in the process.

Digital Solutions I Platforms I Engineering, R&D, and Products

ust.com

Version 0102-20260414

U ▪
S T