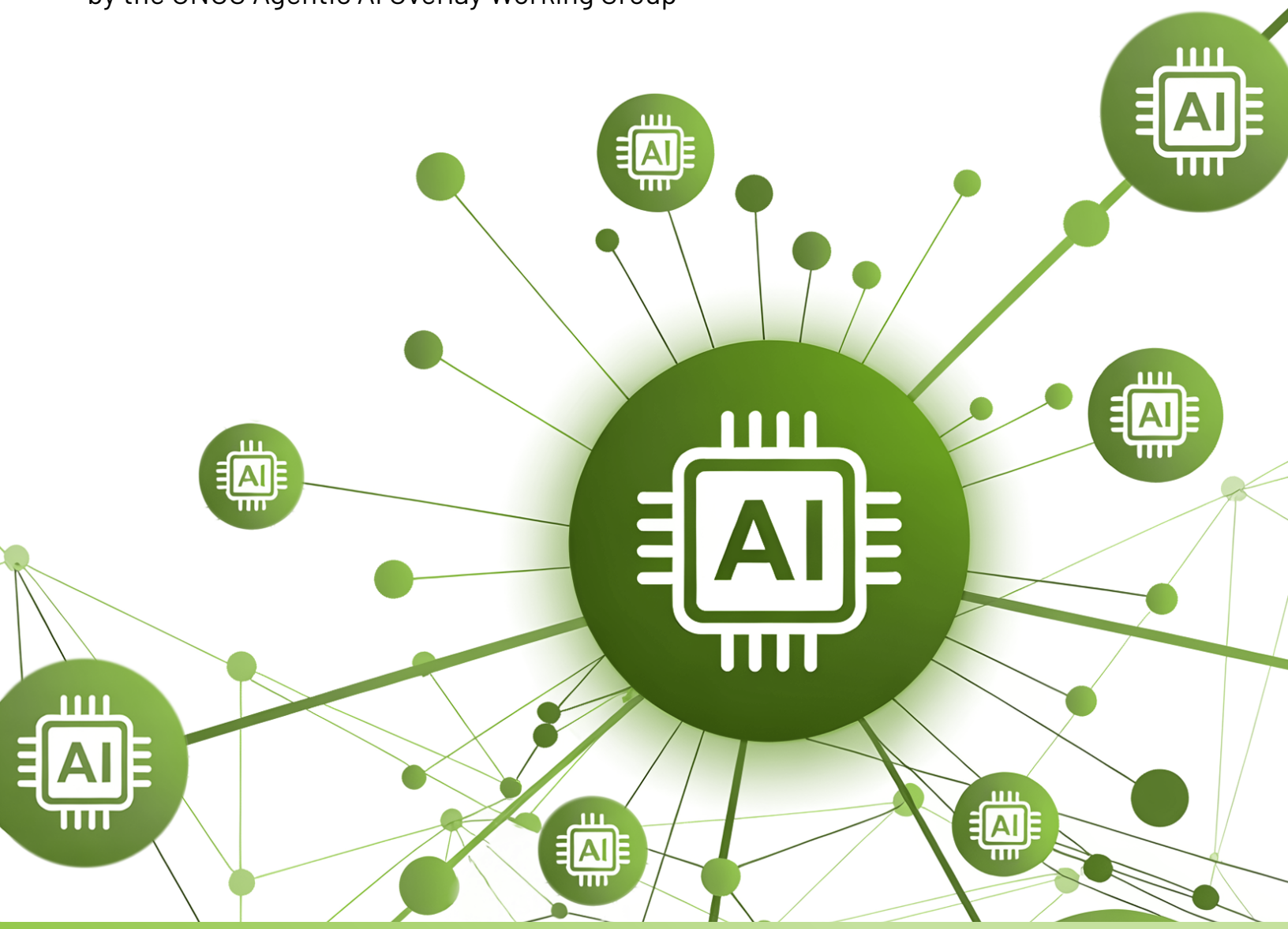
Large Enterprise Requirements to Scale Agentic AI

Part B: Scaling, Deployment, and Operationalization Strategies

by the ONUG Agentic AI Overlay Working Group



ONUG Collaborative | ONUG LLC
PO Box 8455, Naples, FL 34102 | info@onug.net

Executive Summary

The ONUG Agentic AI Overlay Working Group has produced a series of two papers to address large enterprise requirements for scaling agentic AI. The first paper or **Part A** describes a vendor-independent reference architecture aimed to assist in visualizing the various components & their interplay into one figure. This second paper or Part B is dedicated to Scaling, Deployment, and Operationalization Strategies.

While agentic AI promises significant gains in operational efficiency and responsiveness, it introduces new risks that existing architectures were not designed to manage. This paper focuses on the practical strategies for scaling deployments, emphasizing governance, runtime control, data protection, and cross-domain trust as key enablers. Polling data from ONUG's 100,000+ member community validates this direction, highlighting a distinction between single-trust-domain deployments and broader expansions. It positions the Agentic AI Overlay as an operating model for enterprise-vendor collaboration, with requirements derived from enterprise input to ensure production readiness through measurable controls and failure handling.

By reading these papers, enterprise IT leaders, architects, and decision-makers will gain actionable insights into deploying and operationalizing agentic AI incrementally, with safeguards that build confidence for large-scale adoption without compromising security or compliance.

This work was created exclusively by IT executives from large enterprises, including the ONUG Board and its community, with leadership from eBay, Cigna, Bank of America, Indeed, Kraken, and others. It stems from discussions at the Fall 2025 AI Networking Summit and is validated by polling data from ONUG's Community.

This work will be used to compare and contrast the vendor community for their Agentic AI Overlay solutions. The ONUG Community encourages every vendor to become involved, as these are the issues keeping Agentic AI back in the large enterprise. This work is intended to help and support IT business leaders in the large enterprise make better decisions and speed their path to value of becoming enterprise AI companies.

Key Observations

Observation 1:

Enterprises Are Blocked by Control Gaps, Not AI Capability
Across the document and polling results, enterprises consistently report that agentic AI adoption is constrained by the absence of runtime controls, governance mechanisms, and operational safeguards—not by limitations in models or tooling.

Observation 2:

Trust Domain Boundaries Are the Primary Risk Inflection Point
Capabilities that are acceptable within a single trust domain become high-risk or unacceptable once agents operate across clouds, business units, or external services. This boundary is where most deployment hesitation emerges.

Observation 3:

Tool Invocation Is Viewed as the Highest-Risk Agent Capability
Respondents repeatedly emphasized that agent decisions become enterprise risk when they result in real-world actions via tools, APIs, or infrastructure changes, making orchestration and permissioning a focal concern.

Observation 4:

Auditability and Explainability Are Governance Requirements, Not Afterthoughts
Enterprises expect agentic systems to be explainable and reconstructable by default. Black-box autonomy is viewed as incompatible with regulatory, fiduciary, and operational accountability.

Observation 5:

Enterprises Are Willing to Adopt Autonomy—But Only Incrementally
The work shows strong interest in agentic AI, paired with a deliberate preference for bounded autonomy, phased deployment, and human oversight until operational confidence is established.

Strategic Recommendations

Recommendation 1:

Treat Agentic AI as an Operational System, Not an Application
Enterprises should design agentic AI deployments with the same rigor applied to mission-critical infrastructure, including lifecycle management, runtime monitoring, failure isolation, and recovery procedures.

Recommendation 2:

Start with Single Trust Domain Deployments to Prove Control
Early deployments should be intentionally constrained to well-defined environments where governance, observability, and containment can be validated before expanding across trust boundaries.

Recommendation 3:

Make Tool Governance a First-Class Design Requirement
Architectures should explicitly control how agents invoke tools, including permissioning, blast-radius limits, rate controls, and step-up authorization for high-impact actions.

Recommendation 4:

Require Continuous Monitoring and Active Intervention Capabilities
Enterprises should not accept solutions that rely solely on logs or post-incident analysis. Real-time detection, quarantine, and kill-switch mechanisms must be treated as baseline requirements.

Recommendation 5:

Evaluate Vendors on Operational Behavior, Not Autonomy Claims
Vendor assessments should prioritize how systems behave under stress, failure, policy violations, and cross-domain scenarios. Demonstrated control and recovery are stronger indicators of readiness than claims of advanced autonomy.

Recommendation 6:

Engage the Vendor Community through the Agentic AI Overlay Working Group
Vendors are encouraged to participate directly in the Agentic AI Overlay Working Group to collaborate with enterprise IT leaders in shaping practical, production-grade approaches to agentic AI deployment. Through this engagement, vendors can help define workshops, technical track sessions, and live demonstrations for the AI Networking Summits that are explicitly aligned to the enterprise requirements set forth in this paper.

This collaborative model ensures that the large enterprise community is exposed to a range of credible solution approaches while enabling vendors to demonstrate how their offerings address real-world governance, control, and operational challenges. By grounding innovation in enterprise-defined requirements and shared evaluation criteria, the working group accelerates the maturation of enterprise-grade agentic AI systems and helps transform large corporations into confident, responsible enterprise AI companies.

Part B

Scaling, Deployment, and Operationalization Strategies

As enterprises move from experimentation to production deployment of agentic AI systems, scaling and operationalization become the defining challenges. Feedback from the ONUG community poll and working group discussions consistently reinforced that **the primary barrier to adoption is not model capability, but operational readiness** — specifically the ability to deploy, govern, and scale autonomous agents safely across heterogeneous environments. This section outlines how enterprises should approach scaling the Agentic AI Overlay and how the vendor community is expected to engage in proving production viability.

From Poll Results to Deployment Reality

Polling data revealed strong consensus that agentic AI will be embedded first in **operational and infrastructure-facing workflows** before customer-facing systems. Respondents emphasized the need for predictable behavior, auditability, and containment before granting broader autonomy. These findings inform a phased deployment strategy: enterprises should initially deploy agents within **well-defined trust boundaries** — such as a single data center, VPC, or business unit—to validate resiliency, observability, and governance. Once operational confidence is established, cross-domain and cross-organizational flows can be introduced incrementally.

Crucially, the poll results also signaled that enterprises expect vendors to demonstrate **operational scale characteristics**, not just functional correctness. This shifts evaluation criteria away from feature checklists toward measurable, system-level performance and governance outcomes.

Quantifying Scale and Operational Targets

To guide sizing decisions and vendor evaluations, enterprises should quantify the following dimensions early in the design process:

- **Agent population and interaction patterns**, including total agent count, average and peak conversation rates (messages per second), and concurrency targets across domains.
- **End-to-end latency SLOs**, differentiated between time-sensitive operational decisions (e.g., incident response) and customer-facing or advisory interactions.
- **Inference throughput requirements**, including sustained tokens-per-second, burst behavior, and policies for pooling inference capacity across business units.
- **Observability volume**, measured in events per second, with explicit targets for retention periods and query latency to support forensics and compliance.
- **Policy evaluation load and complexity**, including per-message authorization checks, data egress validation, and semantic policy enforcement.
- **Failure domain design**, defining acceptable blast radius, mean time to recovery (MTTR), and the timing of agent quarantine or rollback.
- **Compliance boundaries**, such as data residency matrices and defined paths for PII, PHI, and regulated data across zones and jurisdictions.

These metrics form the basis for realistic capacity planning and enable objective comparison of vendor solutions under production-like conditions.

Deployment Patterns and Architectural Discipline

Working group discussions consistently emphasized that **identity and policy must be treated as first-class controls** at every interaction point—not as perimeter-only mechanisms. Every agent-to-agent exchange, tool invocation, and data access must be authenticated, authorized, observed, and logged by design. This principle holds regardless of where agents execute or which protocols they use.

To preserve architectural flexibility and avoid premature lock-in, enterprises should favor **open protocols and clear extension points** — including A2A-, MCP-, or similar interaction models—so that vendor-provided "agent enclaves" can interoperate cleanly within the overlay. Reference architecture components should be mapped to concrete products only after functional and operational requirements are validated, not before.

Vendor Engagement and Summit Execution

Following the presentation of poll results, the vendor community is invited to engage through **demonstration, benchmarking, and transparency**, not marketing claims. At the AI Networking Summit, vendors will be asked to align their demonstrations to the scaling and operational dimensions outlined above. Scorecards will be used to evaluate how well solutions address agent scale, policy enforcement, observability, resiliency, and compliance in realistic scenarios.

To reinforce enterprise-driven outcomes, ONUG will highlight vendors that demonstrate credible production readiness through **working demonstrations, comparative scorecards, and peer-reviewed workshops**. Recognition may include Summit awards for operational excellence, governance maturity, and architectural alignment with the Agentic AI Overlay reference model.

Operationalizing at Enterprise Scale

The working group consensus is clear: scaling agentic AI is not about deploying more models, but about building **repeatable, governable operating patterns**. Enterprises that succeed will be those that treat agents as long-lived operational assets, invest early in observability and policy infrastructure, and demand that vendors meet them at the level of systems engineering—not demos. The Agentic AI Overlay provides the framework to do so, and this phase of the initiative is where architecture becomes execution.

To reinforce enterprise-driven outcomes, ONUG will highlight vendors that demonstrate credible production readiness through working demonstrations, comparative scorecards, and peer-reviewed workshops. Recognition may include Summit awards for operational excellence, governance maturity, and architectural alignment with the Agentic AI Overlay reference model.

Prioritizing Enterprise Requirements to Scale Agentic AI Deployment

Interpreting Poll Results through Trust Domains and the MAESTRO Framework

The Agentic AI Overlay poll was designed to identify what is **preventing large enterprises from deploying agentic AI systems at scale today**. After the ONUG Agentic AI Overlay Working Group developed the polling questions, the ONUG Board provided review and further refinement. The poll results include responses from ONUG Board members plus the broader ONUG Global 2K members. These results represent the voice of the large enterprise and the core set of requirements needed for wide-scale agentic AI deployments to occur. The results consistently point to the same conclusion: enterprises are not blocked by model capability or innovation velocity, but by the lack of **control, governance, and operational safety**, particularly as agent autonomy increases and systems span multiple trust domains.

To reflect how enterprises actually deploy systems, several poll questions explicitly distinguish between **single trust domain** operation (within a data center, cloud account, or business unit) and **multiple trust domain** operation (across clouds, business units, partners, or public services). This distinction revealed a clear maturity gradient: capabilities that may be acceptable inside a single domain often become **mandatory blockers** once agents cross trust boundaries.

Importantly, the poll questions also map cleanly into the **MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) framework**. MAESTRO provides a structured way to reason about agentic systems by separating where agents operate, how they are secured, what threats they introduce, how risk is controlled, and what outcomes are acceptable. Each poll question corresponds to one or more MAESTRO layers, reinforcing that the community's concerns are architectural and systemic—not tool-specific.

The poll questions therefore represent **baseline enterprise requirements** — prerequisites to adoption rather than differentiators. Each question below is presented in full, followed by its priority, its architectural meaning, and its placement within MAESTRO. All survey takers were asked to answer each question by choosing from the following:

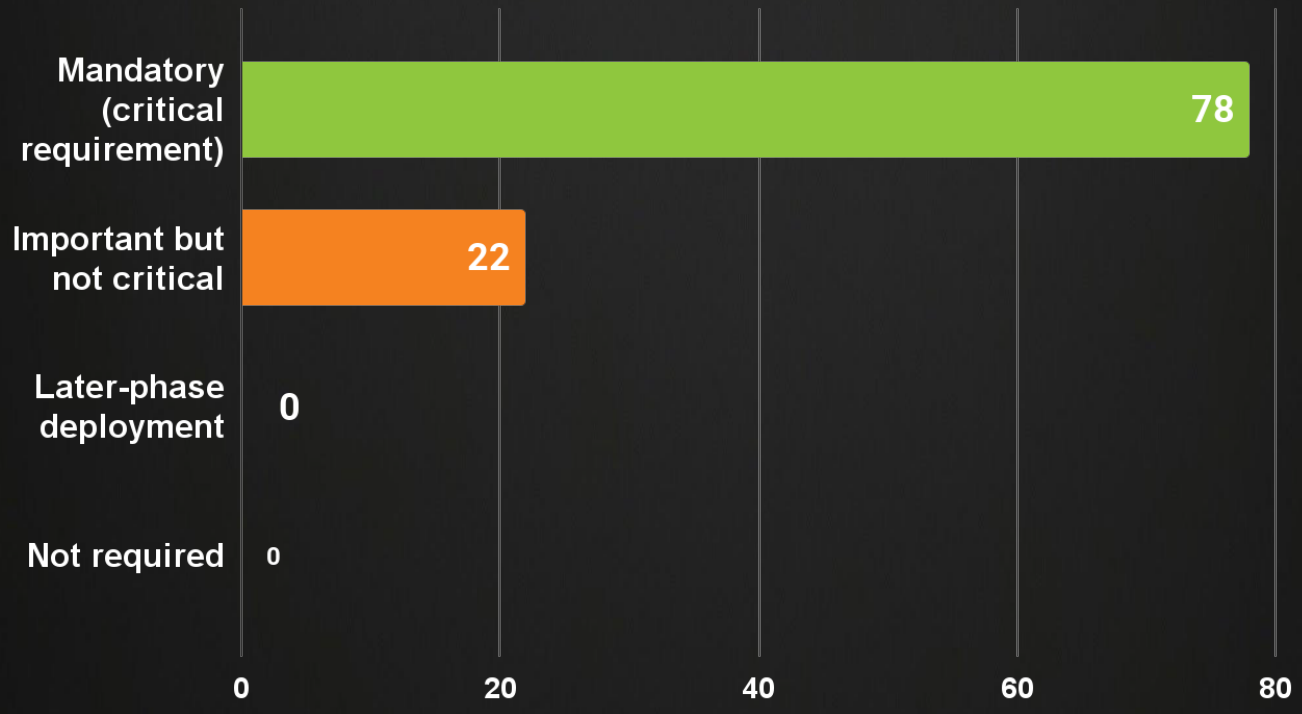
How important is this capability?

- Mandatory (critical requirement)
- Important but not critical
- Later phase deployment

Agent Identity, Lifecycle, and Attestation (Non-Human Identity)

The overlay **SHALL** provide a robust **agent lifecycle identity and attestation framework**, explicitly **independent of human identity**, covering **agent creation, delegation, mutation, and retirement, single enterprise-controlled trust domain**.

The framework **MUST** enforce strong, cryptographic identity and mutual authentication for all agent-to-agent communications, preventing agent spoofing and ensuring only verified, attested agents participate in the secure agent mesh.



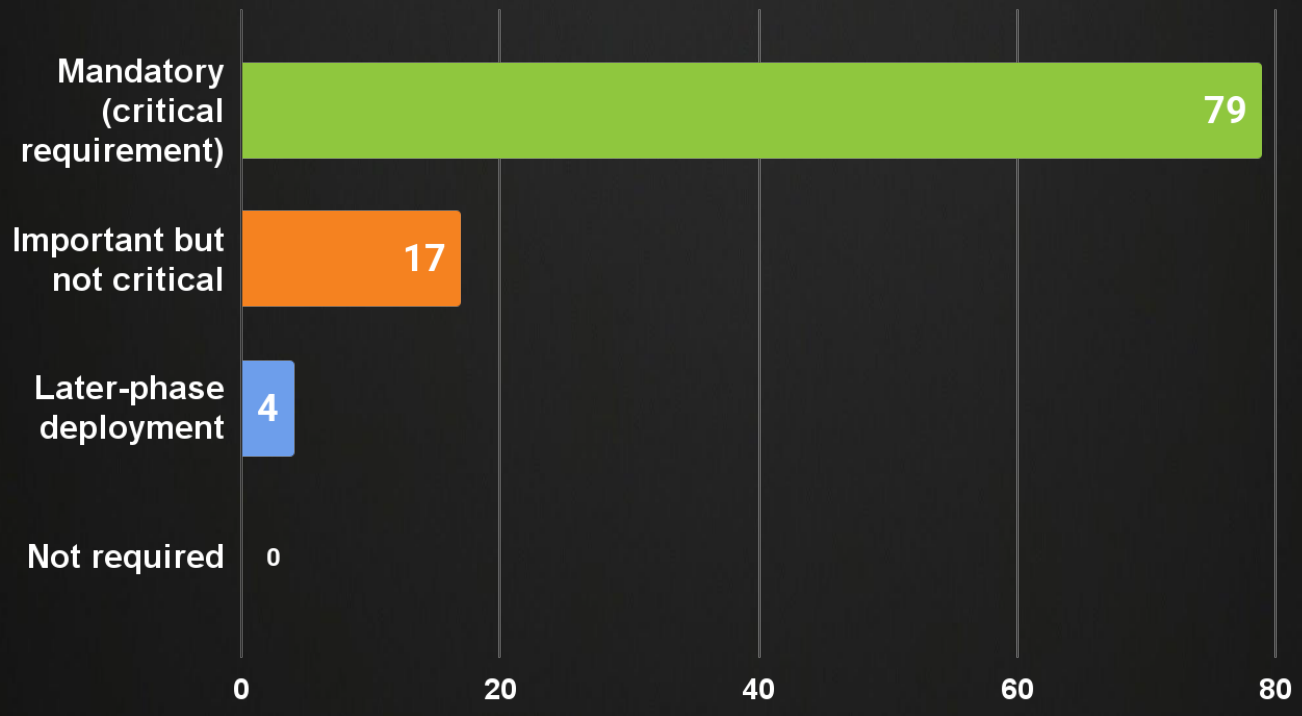
OBSERVATION

From an IT executive perspective, the overwhelming mandate for robust agent identity and attestation underscores the need to prioritize non-human identity management to mitigate risks in autonomous systems; vendors should develop interoperable frameworks that integrate seamlessly with existing enterprise security stacks to accelerate adoption.

Multiple Trust Domains

Building upon 1a, the overlay **SHALL** enforce agent identity, attestation, and authorization consistently across **multiple enterprise-controlled trust domains** (e.g., business units, on-prem, cloud, edge).

Where interaction with **external or partner domains** occurs, the overlay **MUST** support secure federation controls including encryption in transit and at rest, role-based access control (RBAC), and auditable policy enforcement—without assuming ownership or control of partner environments.



OBSERVATION

IT executives view cross-domain agent identity as critical for scalable operations across hybrid environments, emphasizing federation without control assumptions. Vendors must focus on building extensible attestation tools that support multi-trust boundaries to meet enterprise demands for secure, auditable interactions.

Runtime Monitoring and Rogue Agent Detection

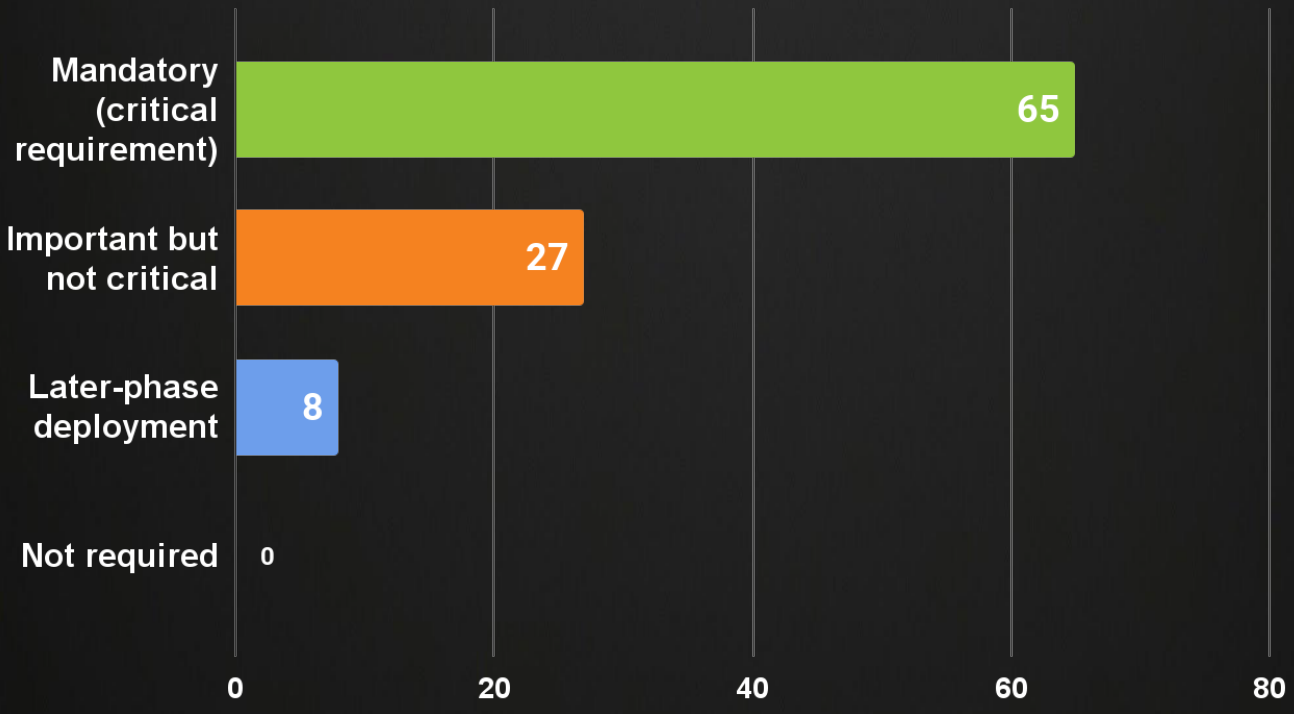
A secure Agentic AI Overlay **SHALL** implement **continuous runtime monitoring** and dynamic security controls to detect and respond to anomalous agent behavior.

This includes identifying deviations from:

- Declared objectives
- Authorized tool usage
- Expected execution patterns

to mitigate risks such as tool misuse, intent breaking, or unsafe emergent behavior.

(Examples include runtime security models analogous to container or workload runtime protection, where execution behavior, API calls, and resource access are continuously observed and enforced.)



OBSERVATION

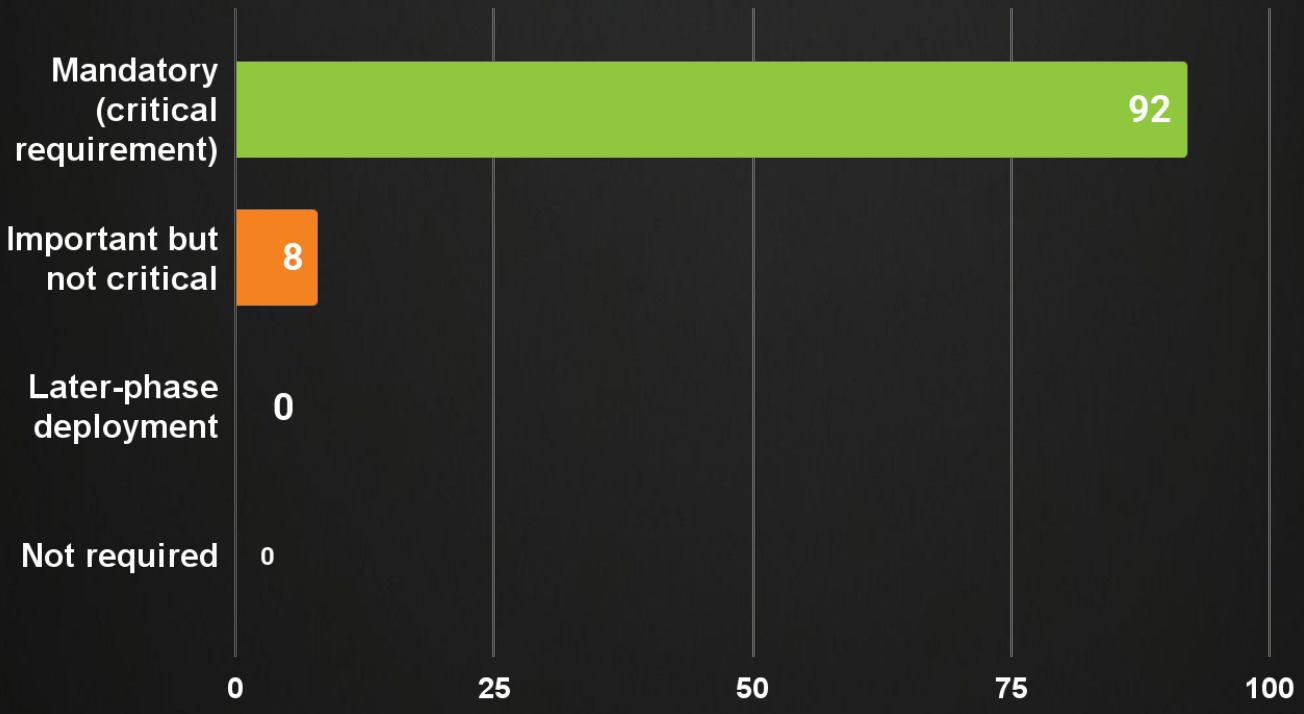
Autonomous agents operating at machine speed cannot rely on post-incident analysis. Enterprises require real-time detection and intervention to manage behavioral drift, misuse, and cascading failures. Executives prioritize runtime monitoring to prevent emergent risks in agent behaviors, highlighting a core need for dynamic controls akin to container security. Product development should incorporate AI-specific behavioral analytics to provide real-time enforcement, addressing gaps in traditional monitoring solutions.

Data Guardrails – Input, Output & Residency Enforcement

The overlay **SHALL** enforce **strict data guardrails** to ensure that **no sensitive enterprise data leaves organizational control.**

This includes validation, inspection, and enforcement for all data and prompts entering or leaving the system to prevent:

- Prompt injection
- Leakage of regulated or sensitive data (PII, IP, financial, healthcare, etc.)
- Unauthorized external data flows



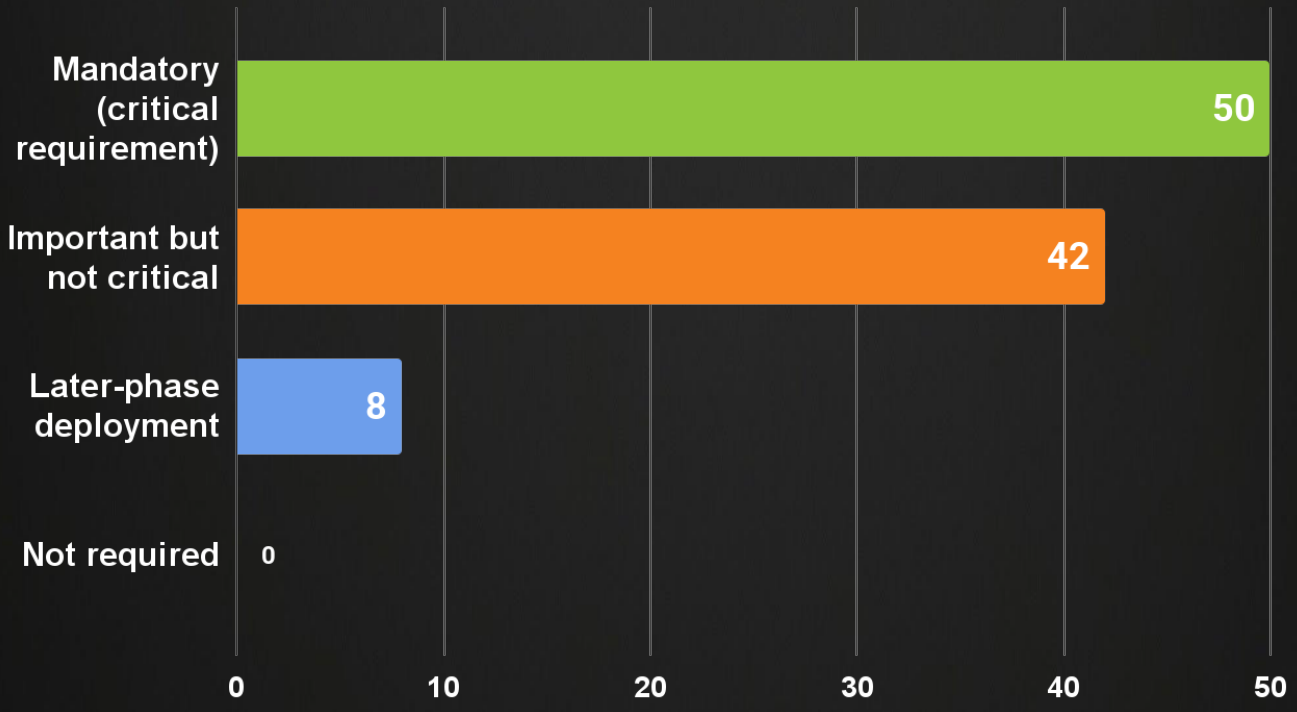
OBSERVATION

The near-universal requirement for data guardrails reflects executives' focus on preventing data leakage and prompt attacks as foundational to compliance. Vendors should innovate with inspection layers that enforce residency and validation without impeding performance, positioning products as essential for regulated industries.

Content Guardrails – Responsible AI Enforcement

In production mode and beyond LLM content vetting tools, the overlay **SHOULD** inspect and moderate agent-generated real time outputs to enforce Responsible AI principles, including protections against harmful, abusive, biased, or non-compliant content, prior to delivery to users or downstream systems.

(This assumes enterprise-grade, vetted models and focuses on policy alignment rather than baseline model hygiene.)



OBSERVATION

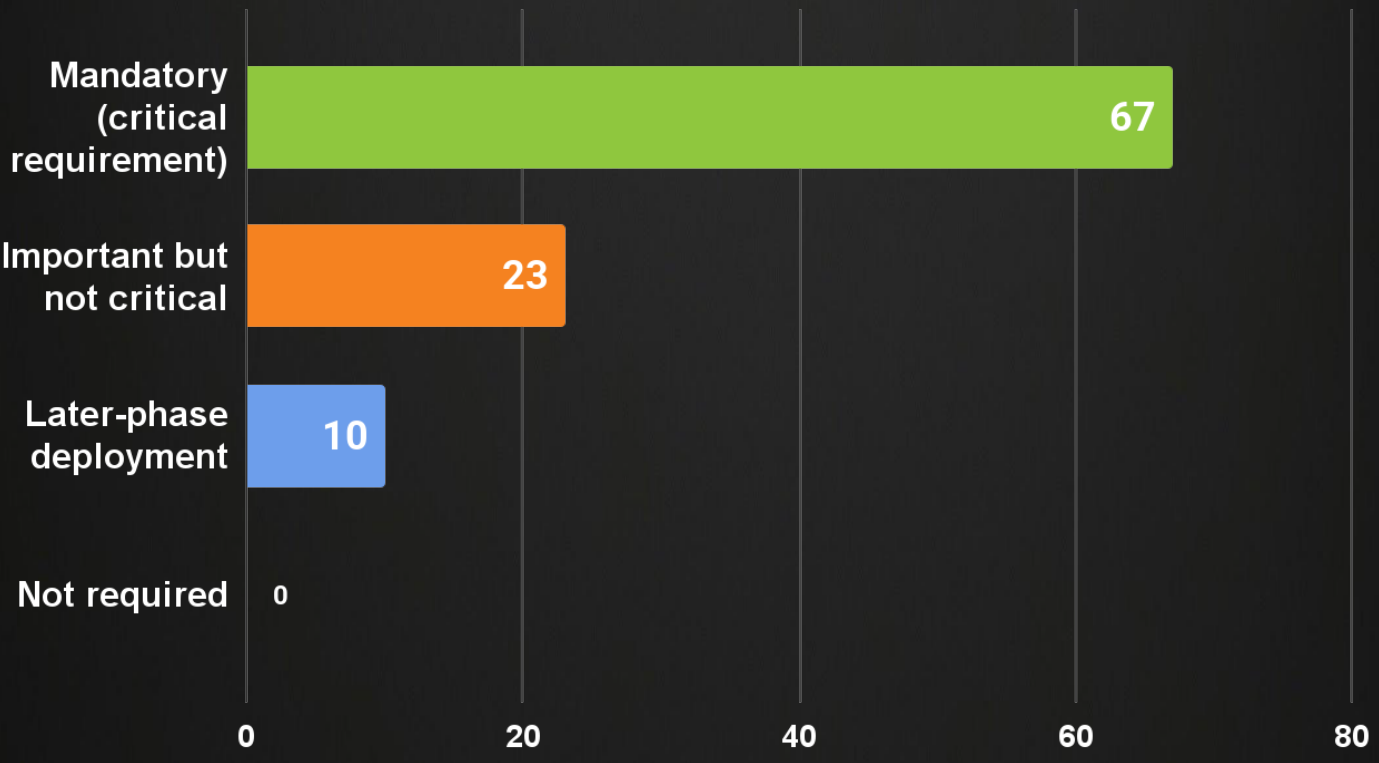
While important, content guardrails for responsible AI show divided criticality, indicating executives seek policy-aligned outputs beyond basic model vetting. Vendors can differentiate by offering modular moderation tools that integrate with enterprise ethics frameworks, enabling phased implementation.

Zero-Trust* Enforcement

(Enterprise-Controlled Domains)

A secure Agentic AI Overlay **SHALL** enforce **Zero Trust principles by default** across network, identity, and runtime layers within enterprise-controlled domains.

This includes continuous verification of identity, context, and policy before permitting any communication or execution, to prevent lateral movement and contain compromise of agents or infrastructure components.



OBSERVATION

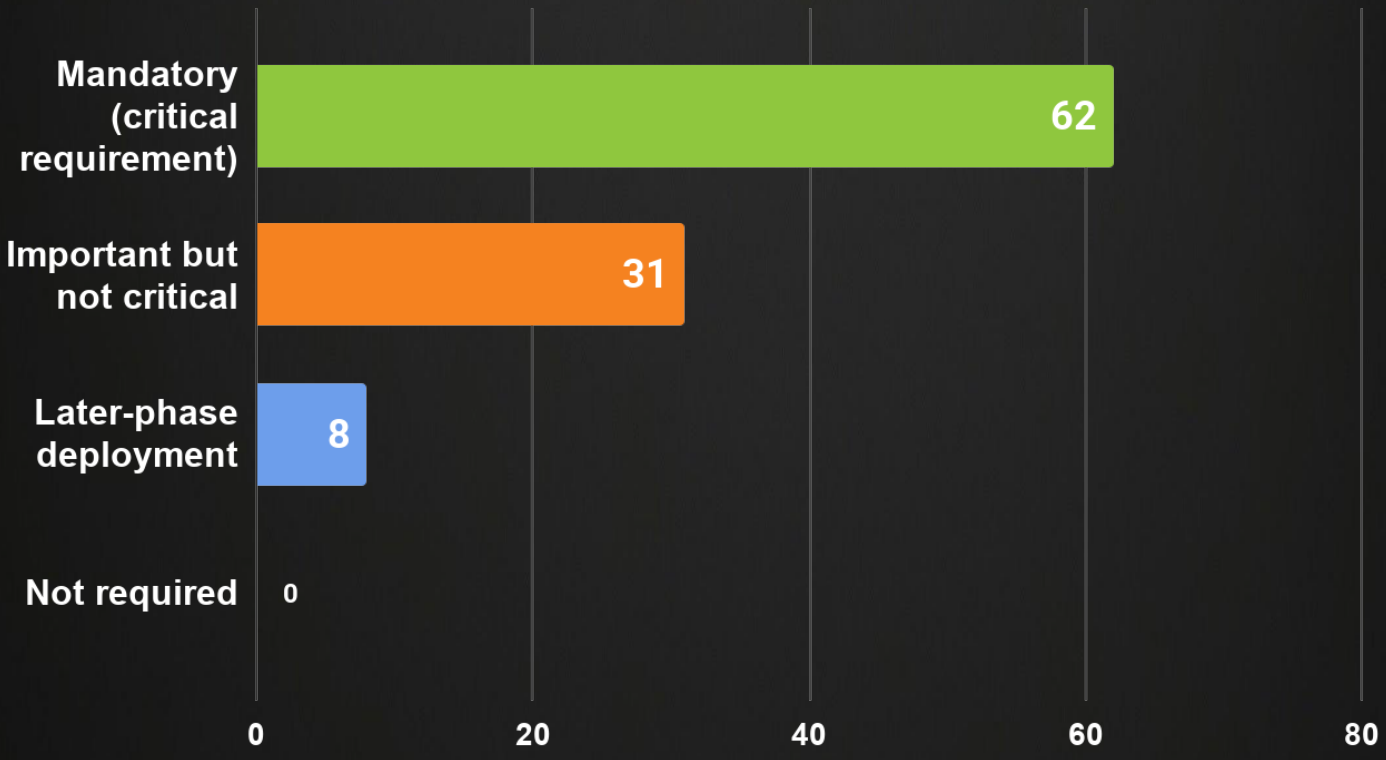
Zero-Trust enforcement within controlled domains is seen as mandatory by most executives to contain threats in agent meshes. Product teams should embed continuous verification in overlays, ensuring compatibility with legacy infrastructure to facilitate broad enterprise rollout.

*Zero-Trust means that all communications within a zero trust construct are authenticated and authorized, in short no endpoints are trusted.

Zero-Trust Across Trust Boundaries & Domains

Building upon 4a, where agents operate across **multiple enterprise-owned environments** (e.g., on-prem, cloud, edge), Zero Trust enforcement **SHALL** apply consistently using policy-driven authorization and runtime risk signals.

For **external or partner domains**, enforcement **MUST** include encrypted communications, explicit authorization boundaries, and auditable access controls.



OBSERVATION

Executives demand consistent Zero Trust across boundaries for resilient multi-environment operations, with emphasis on encryption and audits. Vendors need to prioritize policy-driven solutions that handle external domains securely, reducing integration friction in hybrid setups.

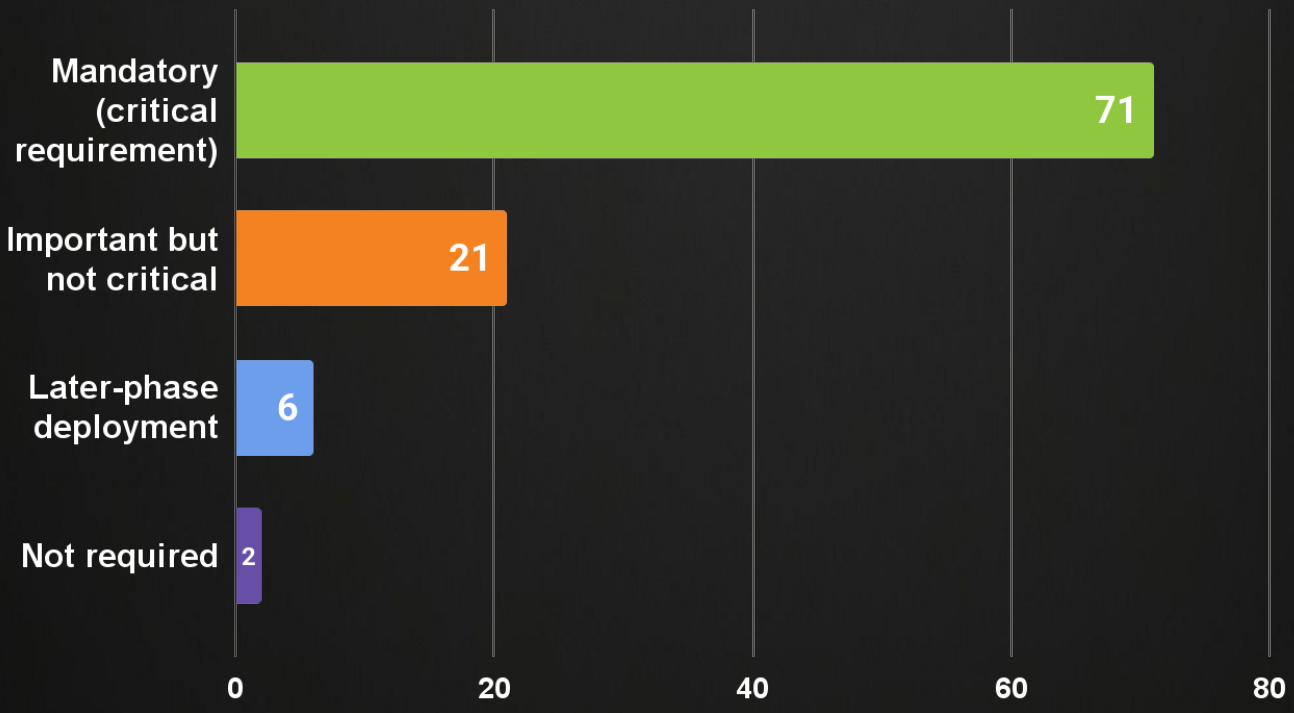
Secure Orchestration & Tool Authorization

A secure Agentic AI Overlay **SHALL** enforce **strict, policy-driven authorization** for all agent tool invocation and orchestration actions.

Agent tool-calling **MUST** ensure that agents cannot perform high-privilege actions across trust or administrative domains without:

- Explicit authentication
- Policy-based authorization (human approval or automated enforcement, as defined by policy)
- Full audit logging

This capability **MUST NOT** assume human-in-the-loop approval by default and **SHALL** support fully automated, policy-driven enforcement.



OBSERVATION

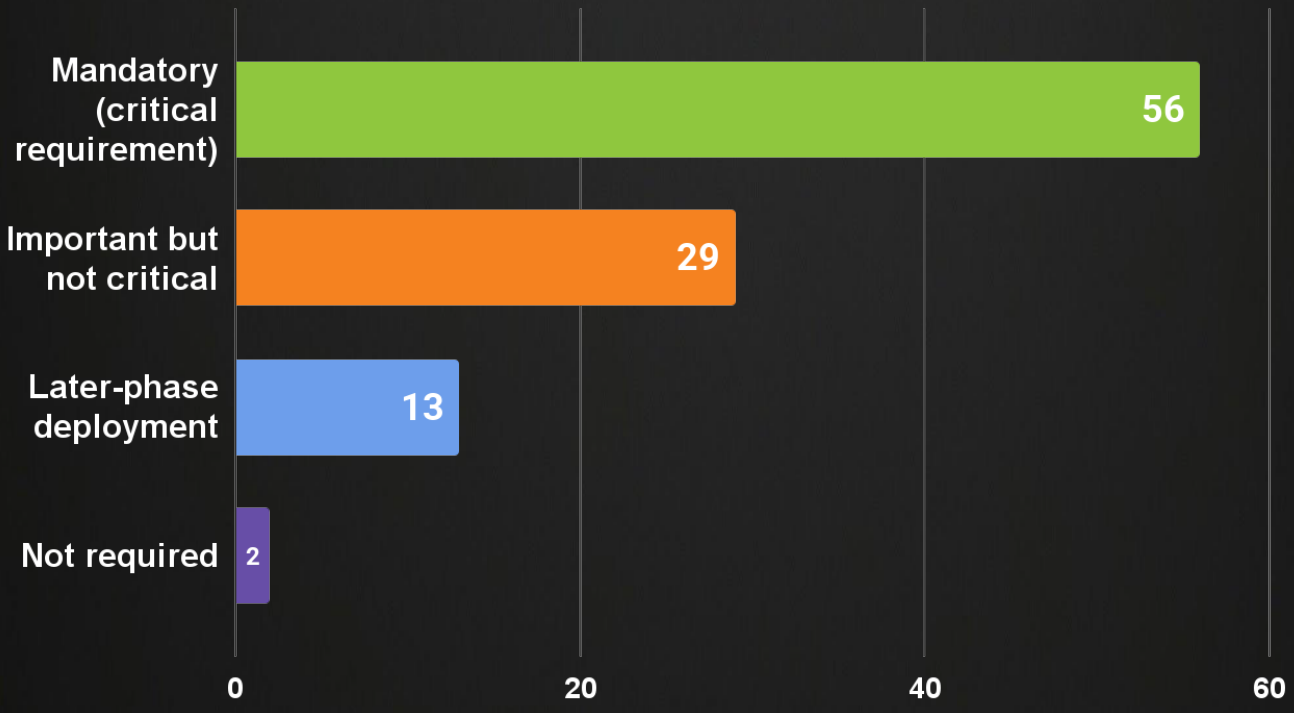
Secure orchestration is critical for executives to control tool invocations and prevent unauthorized actions in automated workflows. Vendors should develop policy engines supporting automated enforcement and logging, aligning with needs for auditability without default human intervention.

Agent Autonomy Governance

A secure Agentic AI Overlay **SHALL** provide explicit, policy-driven governance over **agent autonomy levels**, including the ability to define, constrain, and dynamically adjust how independently an agent may plan, decide, and act.

This includes setting boundaries for:

- Fully autonomous execution
- Supervised or policy-constrained autonomy
- Human-approved or human-in-the-loop operations



OBSERVATION

Enterprises are not rejecting autonomy; they are rejecting **ungoverned autonomy**. The poll and write-in responses emphasize policy-defined autonomy, human oversight, kill switches, and lifecycle governance as prerequisites to responsible outcomes.

Implications for IT Executives and the Vendor Community

The poll results describe a coherent system-level problem: enterprises must manage **multi-agent environments** under real-world **threats**, constrain **risk**, and ensure acceptable **outcomes** — all while maintaining operational velocity. The Agentic AI Overlay provides the architectural mechanism to do so.

For IT executives, these requirements define a **deployment roadmap**: establish control within a single trust domain, then expand to multi-domain operation only when the above controls are proven. For vendors, the message is unequivocal: **single-domain capability is table stakes; ONUG-aligned, cross-domain governance is the differentiator**.

At the AI Networking Summits during 2026 and 2027, these poll-derived requirements will drive vendor demonstrations, scorecards, workshops, and awards. The polling data leaves little ambiguity: **large-scale enterprise adoption of agentic AI will not occur until these requirements are met—and the ONUG Agentic AI Overlay Reference Architecture requirements provides the framework to measure readiness**.

Operationalization strategies for scaling Agentic AI emphasize a deliberate, risk-managed approach, starting with treating agentic systems as mission-critical infrastructure requiring robust lifecycle management, runtime monitoring, and recovery protocols to ensure resilience and compliance. Enterprises should begin deployments within single trust domains to validate governance and containment before expanding across boundaries, prioritizing tool governance through explicit permissioning, rate limits, and step-up authorizations to mitigate high-impact actions. Continuous monitoring with active intervention capabilities, such as real-time anomaly detection and kill-switches, must be mandated, while vendor evaluations focus on demonstrable operational behaviors like auditability and explainability rather than unsubstantiated autonomy promises, enabling incremental adoption aligned with poll-highlighted priorities like data guardrails (92% mandatory) and secure orchestration (71% mandatory).

Appendix A provides a discussion on agent security standards that the working group endorses. These standards are the NIST SP 800-53 Control Overlays for Securing AI Systems and the **MAESTRO** threat modeling framework developed by the Cloud Security Alliance (CSA). MAESTRO in particular can be used to systematically analyze and enhance the security of the **Agentic AI Overlay** architecture. In fact, all of the poll question requirements are mapped directly into the MAESTRO seven layers (see above).

Appendix A: Agentic AI Security Standards Endorsed by the ONUG Agentic AI Overlay Working Group

NIST **Security Controls Meet Agentic Network Architecture**

- NIST AI controls define what needs to be secured in AI systems.
- ONUG defines how AI agents communicate securely across networks.

Together, they form a defense-in-depth strategy: NIST overlays protect the AI system itself, while ONUG overlays ensure secure and intelligent data movement between systems.

What Are the NIST AI Controls?

NIST has released a concept paper and proposed action plan for developing a series of NIST SP 800-53 Control Overlays for Securing AI Systems.These overlays are modular extensions to the existing NIST SP 800-53 cybersecurity controls, tailored specifically for AI systems.

NIST AI controls are particularly valuable for cybersecurity and governance practitioners for several strategic and operational reasons:

- **Avoiding Reinvention:** Instead of creating controls from scratch for each new technology or use case, overlays provide pre-built, validated control sets. This saves significant time and reduces the risk of missing critical security considerations.
- **Common Language:** They create a shared vocabulary across organizations, auditors, regulators, and vendors. When you say "we implement NIST overlay for AI systems," everyone understands the baseline expectations.
- Overlays allow practitioners to apply controls proportionate to specific risks. For example, an AI system processing healthcare data would layer HIPAA-specific overlays on top of base controls, while a financial AI would use PCI-DSS or SOX overlays.
- **Flexibility within Structure:** Rather than rigid one-size-fits-all requirements, overlays provide structured flexibility - you get comprehensive coverage while maintaining the ability to adapt to your specific environment.
- **Evidence of Maturity:** Implementing recognized overlays shows stakeholders and regulators that you're following industry best practices, not just making up your own approach.

They aim to:

- Address unique AI risks like adversarial manipulation, data poisoning, and model inversion.
- Integrate AI risk management into existing cybersecurity frameworks like the NIST Cybersecurity Framework (CSF) and AI Risk Management Framework (AI RMF)
- Provide implementation-focused guidance for securing AI components such as:
    - Training and test data
    - Model weights and configurations
    - Deployment pipelines
    - Outputs and interactions (e.g., LLMs, predictive models, multi-agent systems)

The NIST AI controls currently target five categories:

- Generative AI (e.g., LLMs, image generators)
- Predictive AI (e.g., forecasting, decision support)
- Single-agent AI systems
- Multi-agent AI systems
- AI development pipelines

NIST AI controls complement ONUG's Agentic AI Overlay

NIST AI controls include use cases for single-agent and multi-agent AI systems, which map well to ONUG's focus on agent-to-agent protocols and stateless AI router.

Aspect	NIST SP 800-53 AI Control Overlays (concept)	ONUG Agentic AI Overlay
Purpose	Secure AI systems via tailored cybersecurity controls	Enable secure, autonomous, policy-driven AI networking
Scope	AI system components (models, data, pipelines, outputs)	AI-driven data movement across distributed enterprise networks
Architecture Focus	Security controls for AI lifecycle (development, deployment, use)	AI-driven data movement across distributed enterprise networks
Security Goals	Confidentiality, integrity, availability of AI systems	Secure, adaptive, intent-based networking for AI workloads
Governance Alignment	Maps to NIST CSF and AI RMF	Emerging enterprise architecture standards via ONUG Collaborative

This means we can:

Use NIST AI Controls to secure the agents and their data.
Use ONUG overlays to ensure secure, policy-driven communication between agents.

Shared Emphasis on Real-Time Adaptability
Both frameworks emphasize real-time responsiveness:

NIST AI Controls address dynamic threats like adversarial attacks and model inversion.
ONUG overlays support adaptive routing and telemetry processing for AI workloads.

This alignment supports autonomous AI operations with built-in resilience and observability.

Agentic Overlay Monitoring and Control (AOMC) serves as the NIST Security Control Implementation Mechanism

The NIST AI Controls map directly to AOMC's core functions for many NIST AI-specific controls. AOMC's real-time monitoring and enforcement capabilities operationalize NIST requirements for continuous monitoring, adversarial robustness, and model behavior validation through its emitter pattern and supervision plane.

The semantic firewall and governance oversight features directly implement NIST's explainability and transparency controls, while the audit trail capability satisfies NIST's requirements for AI decision documentation and accountability. Most critically, AOMC's "big red button" and rogue agent detection provide the rapid response and containment mechanisms that NIST identifies as essential for AI risk management, making AOMC essentially a purpose-built platform for executing NIST AI control overlays at machine speed. Specific key areas for NIST AI Control Integration are listed below.

Key Areas for NIST AI Control Integration

1. Public Domain AI Services Layer (Claude, ChatGPT, etc.)

- Supply Chain Risk Management (AI-specific): These third-party AI services require controls for vendor assessment, API security, and data handling agreements
- Model Provenance: Documentation of which models are being used and their capabilities/limitations
- Access Controls: Authentication and authorization for API usage
- Data Classification: Controls on what data can be sent to external services

---

2. Agentic Overlay Monitoring and Control (right side panel)

This aligns perfectly with several NIST overlay concepts:

- Dynamic Security Control: Real-time monitoring of agent behavior
- Governance Oversight: The listed items (proper agent isolation, prevention of data exfiltration) map to NIST's AI governance controls
- Audit and Compliance: Access control and audit trails for agent actions

---

3. Agent Communication Flows (blue lines)

- Data Flow Controls: Implement data classification and handling requirements for agent-to-agent and agent-to-model communications
- Adversarial Robustness: Controls to prevent prompt injection or manipulation between agents
- Rate Limiting and Resource Management: Prevent resource exhaustion or runaway agent behaviors

---

4. Private Infrastructure/VPC Layer

- Isolation Controls: Network segmentation for AI workloads
- Data Residency: Ensuring sensitive data remains within controlled environments
- Compute Resource Security: GPU/TPU access controls and monitoring

---

5. Public Domain Internet Services

- External Data Validation: Controls for data ingested from banking, e-commerce sources
- Privacy Controls: PII handling and compliance (GDPR, CCPA)
- Data Quality Assurance: Validation of external data before use in AI systems

Specific NIST AI Control Implementation

For each zone in the architecture diagram above, we would recommend the following:

1. Trust Boundaries: Define clear boundaries between public AI services, private infrastructure, and external services
2. Risk Scoring: Assign risk levels to each component based on data sensitivity and potential impact
3. Continuous Monitoring: The "Agentic Overlay Monitoring and Control" should implement NIST's continuous monitoring requirements
4. Explainability Requirements: Add logging for decision paths, especially for agent-to-agent communications
5. Bias and Fairness Controls: Particularly important for public-facing services (banking, e-commerce)

Recommended Additions:

Consider adding these NIST-aligned components to your architecture:

- Model Registry: Track versions and performance of all AI models
- Testing Environment: Separate zone for adversarial testing and validation
- Incident Response: Specific procedures for AI-related incidents (model drift, adversarial attacks)
- Performance Baselines: Establish and monitor against performance metrics

MAESTRO Threat Modeling and the Agentic AI Overlay: Securing Multi-Agent Systems

Executive Summary

The MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) framework is the recommended structured approach to systematically analyze and improve the security of the Agentic AI Overlay architecture. Using MAESTRO involves mapping the Overlay's components and security issues across its seven architectural layers—from the Foundation Models (L1) to the Agent Ecosystem (L7)—and following a six-step threat modeling process (Decompose, Identify, Hunt, Assess, Plan, Implement/Monitor).This method specifically tackles the unique, autonomous, and multi-agent challenges by identifying AI-specific threats (such as Prompt Injection, Data Poisoning, and Goal Misalignment Cascades) and ensuring Cross-Layer Defense-in-Depth. The primary goal is to establish strong security measures, including Zero-Trust by Default architecture, real-time anomaly detection, and the capability to dynamically secure and isolate rogue agents.

MAESTRO threat modeling framework can be used to systematically analyze and enhance the security of the **Agentic AI** Overlay architecture.

The MAESTRO framework, developed by the Cloud Security Alliance (CSA), provides **a seven-layer architecture** and a structured process that specifically addresses the unique, autonomous, and multi-agent challenges of systems like the one described in the Agentic AI Overlay presentation.

Applying MAESTRO to the ONUG Agentic AI Overlay Reference Architecture

The core of the application involves mapping the components and security concerns of the Agentic AI Overlay to MAESTRO's seven layers and then executing the MAESTRO threat modeling process (Decompose, Identify, Hunt, Assess, Plan, Implement/Monitor).

1. System Decomposition and Mapping

The components and concerns of the Agentic AI Overlay Reference Architecture can be mapped across the seven layers of MAESTRO:

MAESTRO Layer	Agentic AI Overlay Components & Concerns	Key Threats & Security Outcomes
Layer 7: Agent Ecosystem	Multi-Agent interactions (Agent-to-Agent protocols - A2A), Agent Directory, Secure agent mesh for real-time data exchange	**Threats:** Malicious agent infiltration, Agent collusion, Cascading goal misalignment. **Outcome:** Rigorous agent identity and trust to prevent rogue agents.
Layer 6: Security & Compliance	Security Overlay (Zero-trust by default, policy enforcement), Governance Oversight	**Threats:** Compliance failures, Inadequate access controls, Policy violations. **Outcome:** Compliance automation and policy-based controls for all agent interactions.
Layer 5: Evaluation & Observability	Observability/monitoring, Agentic Overlay Monitoring and Control, Audit trails, Rogue Agent Isolation	**Threats:** Monitoring blind spots, Alert suppression, Log tampering. **Outcome:** Real-time anomaly detection and chain tracing for agents.

MAESTRO Layer	Agentic AI Overlay Components & Concerns	Key Threats & Security Outcomes
Layer 4: Deployment & Infrastructure	Private Infrastructure (ZONE A, ZONE B), Cloud Hosted Services (VPC), Secure policy-driven data movement	**Threats:** Container compromise, Lateral movement, Denial of service, Infrastructure misconfiguration. **Outcome:** Zero Trust runtime enforcement and cloud infrastructure security.
Layer 3: Agent Frameworks	Agent orchestration, Tool usage (Private/Public Tools), API communication, Modal Context Protocol (MCP)	**Threats:** Plugin/tool misuse, API abuse, Unauthorized tool access, Intent manipulation. **Outcome:** API security protection and runtime security for agent actions.
Layer 2: Data Operations	Private Data, Public Data, Individual Account Data, Databases/Data (Memory sharing rules), Prevention of data exfiltration	**Threats:** Data poisoning, Privacy leakage, Unauthorized data access, Memory poisoning. **Outcome:** Data classification, PII detection, and data leakage prevention (DLP) controls.
Layer 1: Foundation Models	The Model (LLM/Reasoning, e.g., ChatGPT, Claude) used by the agents	**Threats:** Prompt injection, Model theft, Adversarial examples, Backdoor insertion. **Outcome:** Automated red teaming, jailbreak detection, and model testing/validation.

Key Security Outcomes and Mitigations

The MAESTRO process ensures that the Agentic AI Overlay achieves specific, robust security outcomes:

1. AI-Specific Threat Mitigation

The framework forces the identification of threats unique to AI agents that traditional security methods miss.

- **Goal Misalignment Cascades (Layer 7):** By modeling A2A protocol interactions, MAESTRO identifies where a failure in one agent's goal could influence others, and the security outcome is the implementation of **cryptographic trust models** and strict **interaction policies** to prevent collusion and cascades.
- **Prompt Injection and Tool Misuse (Layers 1 & 3):** It targets the manipulation of the core LLM and the tools it calls via the MCP. The outcome is to enforce **ModelScan/testing** for core models and deploy **AI-powered API protection** for all agent tool calls.
- **Data and Memory Poisoning (Layer 2):** The outcome is to prevent manipulation of an agent's knowledge base or memory (private data) through **PII detection**, **data masking, and sanitization checks** at the ingress/egress points of the data operations layer.

2. Cross-Layer Defense-in-Depth

MAESTRO emphasizes hunting for cross-layer threats (e.g., infrastructure -> data -> model compromise paths) which is critical in the multi-zone, federated Agentic Overlay.

- **Zero-Trust Identity:** An agent identity compromise (Layer 7) can affect the entire ecosystem (Layer 1-6). The MAESTRO-driven outcome is implementing a **Zero-Trust by Default** architecture with agent-specific identity and attestation, ensuring that agents are continuously validated regardless of network location.
- **Supply Chain Attacks:** A supply chain compromise (Layer 4/6) could inject backdoors into the Foundation Models (Layer 1). The outcome is the deployment of **AI Supply Chain Protection (AI-SPM)** tools (like Wiz or Palo Alto Networks) integrated across infrastructure and model layers.

3. Continuous Adaptation and Governance

The framework is a living document, supporting the dynamic nature of the Agentic AI Overlay.

- **Observability and Auditability:** By integrating Layer 5 (Evaluation & Observability) with Layer 6 (Security & Compliance), the framework ensures continuous **runtime monitoring** and **auditable logs** are in place to detect goal drift, policy violations, and rogue agent behavior in real-time. The outcome is a system that can **dynamically secure and isolate rogue agents.**

This structured approach, with its layer-specific focus and cross-layer analysis, directly aligns with the stated security goals of the Agentic AI Overlay Working Group.

Vendor Solutions Mapped to MAESTRO Layers

The MAESTRO framework, in addition to defining the security architecture, provides a mapping of existing vendor solutions to each of its seven layers. This allows organizations implementing the Agentic AI Overlay to quickly identify and deploy specific security tools to address the layer-specific threats identified during the threat modeling process.

MAESTRO Layer	Layer Focus	Key Threats Mitigated	Vendor Solutions & Application to Overlay
L7: Agent Ecosystem	Multi-agent interactions, marketplaces, collaborations.	Malicious agent infiltration, Agent collusion, Cascading goal misalignment.	**DataDome:** MCP Protection, agent traffic control in the agent mesh. **Lasso Security:** MCP gateway, managing agent lifecycle and ensuring trusted agents. **Cisco:** MCP Scanner for supply chain security.

MAESTRO Layer	Layer Focus	Key Threats Mitigated	Vendor Solutions & Application to Overlay
L6: Security & Compliance	Security controls, governance frameworks, auditability.	Policy violations, Compliance failures, Audit log tampering.	**AccuKnox AI Security:** Compliance automation (NIST/PCI-DSS/GDPR) against the standards leveraged by the Overlay. **CalypsoAI:** Policy-based controls, enforcing rules like memory sharing. **IriusRisk:** MAESTRO-integrated threat modeling platform for process management.
L5: Evaluation & Observability	Monitoring, debugging, logging, telemetry systems.	Monitoring blind spots, Log tampering, Alert suppression.	**Datadog:** LLM observability, chain tracing, and anomaly detection for agent actions. **Splunk AI:** Security monitoring and observability for the Overlay's control plane. **Fiddler AI:** Model monitoring and drift detection for core AI models.
L4: Deployment & Infrastructure	Servers, containers, networks, orchestration platforms (e.g., in VPCs, DCs, Edge).	Container compromise, Lateral movement, and Infrastructure misconfiguration.	**Wiz AI-SPM:** AI supply chain and cloud infrastructure security for multi-zone deployments. **Palo Alto Networks:** Runtime protection for agents and enforcement of Zero Trust principles. **AccuKnox:** Kubernetes-native, Zero Trust runtime enforcement in the private infrastructure.

MAESTRO Layer	Layer Focus	Key Threats Mitigated	Vendor Solutions & Application to Overlay
L3: Agent Frameworks	Orchestration platforms (LangChain, AutoGen), APIs, plugins.	Plugin/tool misuse, API abuse, Unauthorized tool access, Intent manipulation.	**Protect AI:** Layer (runtime security) to secure the agent's execution against unauthorized tool use. **Akto.io / Salt Security:** AI-powered API security to protect interfaces to Private and Public Tools. **Akamai Firewall for AI:** Model-agnostic API and edge security.
L2: Data Operations	Data storage, processing, vector embeddings, databases.	Data poisoning, Privacy leakage, Memory poisoning, Unauthorized data access.	**Cyera:** AI-driven data classification and governance for Private Data. **WhyLabs:** LLM data protection, PII detection, and data leakage prevention for data in motion. **SentinelOne:** Prompt Security (DLP controls, tokenization, redaction) to prevent data exfiltration.
L1: Foundation Models	Core AI models (GPT-4, custom LLMs), pre-trained models.	Prompt injection, Model theft, Adversarial examples, Backdoor insertion.	**Lakera Guard:** Prompt injection and jailbreak detection, protecting the core LLM/Reasoning engine. **Mindgard:** Automated red teaming and adversarial testing. Protect AI: ModelScan and Guardian for model testing and validation. **Check Point:** AI application defense, including the Lakera acquisition.

Example: Consumer application of the MAESTRO Process for the Agentic AI Overlay

The Agentic AI Overlay Working Group should apply the MAESTRO threat modeling process to identify and mitigate risks across the architecture systematically.

The MAESTRO methodology follows a structured, iterative six-step process:

Step	Action	Application to the Agentic AI Overlay
1. Decompose	Break down the system using the seven MAESTRO layers.	Map all components to the seven layers. Define agent goals (e.g., orchestration, data exchange), the tools they use (Private/Public Tools), and their interactions (A2A protocols, MCP).
2. Identify	Systematically brainstorm layer-specific threats.	Use the MAESTRO threat landscapes to identify specific risks: L1 (Prompt Injection), L2 (Data Poisoning), L3 (API Abuse), L4 (Container Compromise), L6 (Compliance Failure). Reference MITRE ATLAS and OWASP frameworks.
3. Hunt	Analyze inter-layer connections to identify cross-layer threats.	Identify compromise paths like: Infrastructure > Data > Model attacks. Look for Goal Misalignment Cascades (L7 > L7) and lateral movement (L4 > L2).

Step	Action	Application to the Agentic AI Overlay
4. Assess	Evaluate the likelihood and business impact for each prioritized threat.	Use a risk matrix to prioritize threats like "Rogue Agent Isolation Evasion" or "Data Exfiltration via compromised A2A channel" based on severity and probability.
5. Plan	Develop a defense-in-depth plan using layered and AI-specific controls.	Implement Layered Controls (e.g., access controls on L2/data, runtime enforcement on L4/infra). Deploy AI-Specific Defenses (e.g., model red teaming, formal verification of agent plans).
6. Implement/ Monitor	Deploy solutions and continuously monitor for new threats.	Iterate the threat model as agent functions, data sources, and external threats evolve. Use the Observability (L5) systems for continuous auditing and real-time anomaly detection.

MAESTRO offers targeted steps to secure the Agentic AI Overlay by embedding security within its key components.

Implementation Guidance

The planning step should focus on both AI-specific as well as cross-layer mitigations. Cross-layer mitigations are security controls designed to address attack paths that span multiple layers of an architectural framework, such as MAESTRO. Instead of focusing on patching a single component, they implement chained defenses to stop an attacker's progress across the system, for example, from the infrastructure layer up to the foundation model layer.

In the Agentic AI Overlay, this includes implementing solutions like Zero Trust to prevent lateral movement and continuous attestation to secure agent identities across the entire ecosystem.

1. AI-Specific Mitigations:

- **Agent Identity and Attestation (L7):** The Overlay should implement a robust agent identity system (beyond human identity) and use mutual authentication for all A2A communications to prevent agent spoofing and ensure only attested agents participate in the secure agent mesh.
- **Runtime Monitoring (L5):** Implement continuous dynamic security control and anomaly detection to flag deviations from an agent's expected goals or tool usage (T2 - Tool Misuse, T6 - Intent Breaking).
- **Content Guardrails (L1/L2):** Implement filters and validation checks for data entering the system and model prompts to prevent Prompt Injection and to enforce the Prevention of data exfiltration (PII or sensitive data).

2. Cross-Layer Mitigations:

- **Zero-Trust Enforcement (L4/L6):** Enforce Zero Trust by default at the network and runtime levels. This prevents Lateral Movement, should an attacker compromise an infrastructure component in one zone, isolating the threat before it can reach Private Data (L2).
- **Secure Orchestration (L3/L4):** Agents' tool-calling (L3) should be strictly permissioned, ensuring an agent operating in Zone A cannot perform high-privilege actions on Private Tools in Zone B without multi-factor authorization and audit trail logging.

Specific Use Case Scenarios

Use Case Scenario	MAESTRO Layers & Threats	Mitigation Focus
Secure Data Exchange	**L7 (A2A), L2 (Private Data):** Agent A in Zone A needs to securely send Private Data to Agent B in Zone B. **Threat:** Agent Communication Poisoning.	**Mutual Authentication and Encryption:** Use A2A protocols with end-to-end encryption for all data exchange. Use DLP controls (L2) to verify data content before transmission
Policy Enforcement	**L6 (Governance), L3 (Tool Access):** An agent attempts to perform an unauthorized configuration change on a network component. **Threat:** Privilege Escalation.	**Dynamic Access Control:** Use the Policy Enforcement (L6) mechanism to check the agent's attested identity and current goal against the access control rules before allowing the Tool invocation L3.
Supply Chain Risk	**L4 (Infrastructure), L1 (Model):** A third-party library used to deploy the agent is compromised. **Threat:** Backdoor Insertion.	**AI Supply Chain Protection AI-SPM:** Scan the container image and model artifacts before deployment (L4). Implement provenance/ attestation (L6) for all foundational models and tools.

Secure Agent Lifecycle

The lifecycle of AI agents encompasses four critical phases, each requiring specific security controls.

Agent Creation establishes foundational security through cryptographic identity assignment, policy-based guardrails, least-privilege permission scoping, and validated initialization, ensuring only authorized, properly constrained agents are deployed.

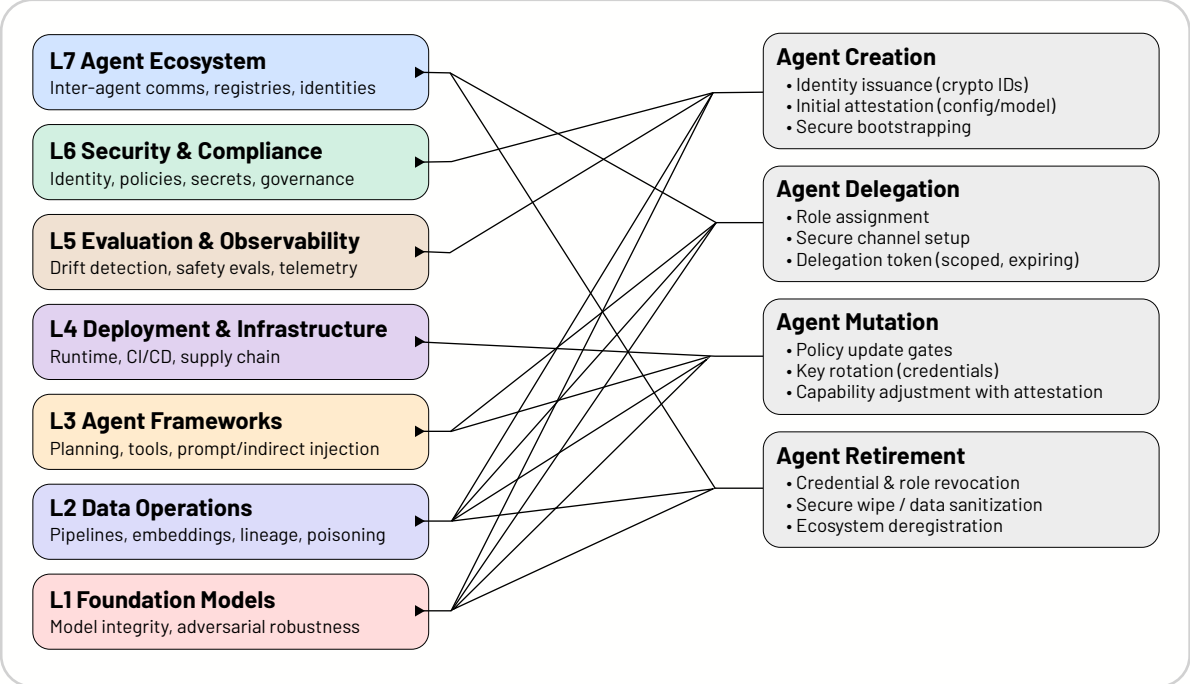
Agent Delegation addresses the risks of task distribution by implementing signed delegation tokens with time limits, enforcing trust boundaries that limit inherited permissions, maintaining comprehensive audit logs, and isolating contexts to prevent data leakage.

Agent Mutation manages the security implications of agent evolution through cryptographically signed updates, immutable versioning with rollback capabilities, behavioral drift monitoring, and continuous compliance validation to detect unauthorized changes.

Finally, **Agent Retirement** ensures secure decommissioning through complete memory and credential wiping, token revocation, generation of a final audit report, and adherence to data retention policies for regulatory compliance. Together, these phases create a defense-in-depth approach that maintains security, accountability, and regulatory compliance throughout an agent's operational lifetime.

The Secure Agent Lifecycle can then be aligned to **MAESTRO** to perform threat modeling. This will allow enterprises to run structured, layer-by-layer analyses and develop macro and micro security controls for each lifecycle phase.

Example: MAESTRO alignment to the Secure Agent Lifecycle



ONUG Collaborative | ONUG LLC
PO Box 8455, Naples, FL 34102 | info@onug.net

Lifecycle Phase	MAESTRO Layers	Primary Risk Themes	Key Controls	References
Agent Creation	L1 Foundation Models; L2 Data Operations; L3 Agent Frameworks; L6 Security & Compliance	Model provenance / backdoors; data poisoning; prompt / indirect injection; identity & policy baselines	Model SBOM & attestation; adversarial evaluation; data lineage & validation; tool-use guardrails; least privilege & strong auth	**CSA MAESTRO overview (2025-02-06):** https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro **CSO MAESTRO deep dive (2025-10-15):** https://www.csoonline.com/article/4072341/introducing-maestro-a-framework-for-securing-generative-and-agentic-ai.html
Agent Delegation	L7 Agent Ecosystem; L3 Agent Frameworks (+ cross-layer)	Agent impersonation & identity sprawl; insecure inter-agent comms; tool misuse across delegated chains; cascading trust failures	Mutual authentication; signed/expiring delegation tokens; scoped capabilities; encrypted comms; context isolation; auditable task graphs	**OWASP MAS Threat Modelling Guide (2025-04-22):** https://www.hackernoob.tips/content/-files/2025/05/Agentic-AI-MAS-Threat-Modelling-Guide-v1-FINAL.pdf

Lifecycle Phase	MAESTRO Layers	Primary Risk Themes	Key Controls	References
				OWASP MAESTRO overview https://blog.ogwilliam.com/post/owasp-maestro-agentic-ai-security
Agent Mutation	L4 Deployment & Infrastructure; L2 Data Operations; L5 Evaluation & Observability; L6 Security & Compliance	Supply - chain/CI risks; memory/state poisoning; behavioral drift; weak change control	Signed & attested updates; reproducible builds; memory isolation & validation; continuous evaluation & rollback gates; credential rotation	**CSO MAESTRO controls:** https://www.csoonline.com/article/4072341/introducing-maestro-a-framework-for-securing-generative-and-agentic-ai.htm **arXiv case study (2025-08-12):** https://arxiv.org/abs/2508.10043
Agent Retirement	L6 Security & Compliance; L2 Data Operations; L7 Agent Ecosystem	Orphaned credentials/identities; residual sensitive state; ecosystem registry / marketplace artifacts	Automated credential revocation; cryptographic erasure & retention policy; deregistration / tombstoning; final audit / report	**CSO MAESTRO scope & baseline** https://www.csoonline.com/article/4072341/introducing-maestro-a-framework-for-securing-generative-and-agentic-ai.html

Lifecycle Phase	MAESTRO Layers	Primary Risk Themes	Key Controls	References
				OWASP MAS identity risks: https://blog.ogwilliam.com/post/owasp-maestro-agentic-ai-security

Recommendation: The Overlay Working Group should treat the MAESTRO framework not as a checklist, but as a living methodology, continuously re-evaluating risks with every major update to the Foundational Overlay Components (MCP, A2A protocols, etc.).

Attribution

- **Framework:** MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome)
- **Source:** Cloud Security Alliance (CSA)
- **Author:** Ken Huang Release
- **Date:** February 2025
- **Vendor Mapping:** Based on 2025 market analysis

Additional Resources

- **CSA Blog:** Agentic AI Threat Modeling Framework: MAESTRO
- **MCP Security Resource Center:** modelcontextprotocol-security.io
- **OWASP Agentic Security Initiative:** Multi-Agentic System Threat Modeling Guide
- **CSO Magazine:** Introducing MAESTRO: A framework for securing generative and agentic AI
- **Academic Research:** "Securing Agentic AI: Threat Modeling and Risk Analysis for Network Monitoring Agentic AI System" (arXiv:2508.10043)

ONUG Collaborative | ONUG LLC
PO Box 8455, Naples, FL 34102 | info@onug.net