AI Agent Containment Rubric

Five dimensions to assess your team's ability to contain AI-specific incidents.

> Most incident response plans assume a deterministic attack surface. AI agents are probabilistic, autonomous, and capable of taking actions before anyone notices. This rubric helps you assess your team's containment capability across five dimensions. For each dimension, check the maturity levels you have achieved. The gaps between levels are where incidents become breaches.

1 Detection and Alerting

You cannot contain what you cannot detect. AI-specific incidents require monitoring that traditional SOC tools do not provide.

☐ Agent-specific monitoring defined Monitoring rules cover agent-specific behaviors: unusual tool call patterns, unexpected output distributions, and anomalous data access.

WHY *Traditional monitoring detects service outages. It does not detect an agent that is behaving correctly but producing harmful outputs.*

☐ Anomaly detection thresholds configured Statistical baselines for agent behavior have been established, and alerts fire when behavior deviates beyond configured thresholds.

WHY *Without baselines, every agent anomaly looks like noise. Baselines turn noise into signal.*

☐ Output content scanning active Agent outputs are scanned for data exfiltration patterns (PII, credentials, confidential data) and for signs of prompt injection in tool call parameters.

WHY *An agent that has been compromised may produce outputs that appear normal but contain embedded instructions or extracted data.*

☐ Alert escalation path defined AI-specific alerts have a separate escalation path from traditional security alerts, with named responders who understand AI failure modes.

WHY *An AI-specific alert routed through a traditional SOC queue will be triaged as a false positive by operators who do not understand the context.*

☐ Detection coverage tested The monitoring and alerting system has been tested against known attack patterns to verify that each detection rule fires as expected.

WHY *Detection rules that have never been tested are unverified assumptions. Testing reveals gaps between intended and actual coverage.*

2 Containment and Isolation

Containment for AI agents is not just about network isolation. It involves stopping autonomous actions, preserving evidence, and preventing behavioral drift during the incident.

☐ Agent kill switch operational A mechanism exists to immediately stop the agent's execution and revoke its tool access without requiring a full system restart.

WHY *The faster you stop an agent's actions, the smaller the blast radius. Kill switches that exist only in documentation are not containment controls.*

☐ Network isolation capability The agent's network access can be restricted to a subset of its normal connections, blocking outbound data exfiltration while preserving internal communication for investigation.

WHY *Complete network isolation may destroy evidence. Selective isolation preserves investigative capability while stopping active harm.*

☐ Data access freeze procedure A documented procedure exists to revoke the agent's access to data stores, APIs, and tools while preserving the data it has already accessed for forensic analysis.

WHY *Continuing data access during an incident expands the exposure window. A freeze procedure stops the expansion while preserving evidence.*

☐ Rollback capability verified The agent's configuration, prompts, and tool definitions can be rolled back to a known-good state without rebuilding the entire system.

WHY *If the agent has been compromised through prompt injection or tool poisoning, rollback to a known-good state is the fastest way to restore security.*

☐ Sandbox escape detection Monitoring detects when an agent attempts to access resources outside its intended scope, including lateral movement to other agents or internal systems.

WHY *A compromised agent will attempt to expand its access. Sandbox escape detection catches lateral movement that containment controls should prevent.*

3 Response and Recovery

Response to AI incidents requires different procedures than traditional incidents. The agent's probabilistic behavior means that recovery is not simply restoring a backup.

☐ AI-specific incident playbook exists A written playbook covers the specific response steps for common AI incident scenarios: prompt injection compromise, tool poisoning, data exfiltration, and unauthorized autonomous action.

WHY *Playbooks that do not address AI-specific scenarios will not guide responders through the unique challenges of agent incidents.*

☐ Evidence preservation procedure defined The procedure for preserving agent interaction logs, tool call histories, and model outputs for forensic analysis is documented and tested.

WHY *Agent evidence is ephemeral. Logs can be lost, model states can be overwritten, and interaction histories can be truncated. Preservation procedures prevent evidence loss.*

☐ Communication template prepared Templates for internal and external communications about AI incidents are prepared, including regulatory notification requirements and customer communication guidance.

WHY *During an incident, drafting communications under pressure leads to inconsistent messaging and potential regulatory violations. Prepared templates ensure consistent, compliant communication.*

☐ Recovery decision criteria defined The criteria for determining when an agent is safe to return to production after an incident are documented, including testing requirements and approval authorities.

WHY *Returning a compromised agent to production without proper validation creates a recurring incident. Decision criteria prevent premature recovery.*

☐ Post-incident review conducted Every AI incident triggers a structured review that documents root cause, response effectiveness, and specific improvements to detection, containment, and prevention controls.

WHY *Reviews that do not produce specific improvements repeat the same failures. Documented improvements create measurable progress.*

4 Communication and Escalation

AI incidents often require communication paths that do not exist in traditional incident response plans. The technical complexity means that communication must bridge security, engineering, legal, and business teams.

☐ AI incident escalation matrix defined A named escalation path exists for each level of AI incident severity, with specific roles for security, engineering, legal, and business stakeholders.

WHY *AI incidents involve technical decisions that traditional incident response roles are not equipped to make. A defined matrix prevents decision-making delays.*

☐ Cross-functional response team established A response team that includes members from security, engineering, data governance, and legal has been formed and has practiced responding to AI incidents together.

WHY *AI incidents require decisions that span multiple domains. A team that has not practiced working together will fail under the pressure of an active incident.*

☐ Regulatory notification timeline documented The timelines and requirements for regulatory notification of AI incidents (EU AI Act, GDPR, sector-specific rules) are documented with specific triggers for each obligation.

WHY *Regulatory notification timelines are strict and non-negotiable. Missing a deadline creates additional liability on top of the incident itself.*

☐ Customer communication protocol defined The protocol for communicating with customers about AI incidents, including what information to share, when to share it, and through which channels, is documented.

WHY *Inconsistent or delayed customer communication about AI incidents destroys trust faster than the incident itself. A protocol ensures consistent, timely communication.*

☐ Executive briefing template prepared A template for briefing executive leadership about AI incidents, including technical context translated to business impact, response status, and recommended actions.

WHY *Executives need technical context to make informed decisions, but they do not have time for technical detail. A prepared template bridges the gap between technical reality and executive decision-making.*

5 Continuous Improvement

Containment capability is not a static state. It degrades over time as agents evolve, new attack patterns emerge, and team members change. Continuous improvement processes prevent this degradation.

☐ Detection rule review cadence established Detection rules and alerting thresholds are reviewed quarterly against the latest AI attack patterns and agent architecture changes.

WHY *Detection rules that are not updated against current threats become ineffective. Quarterly reviews ensure rules stay relevant.*

☐ Containment drill schedule defined Tabletop and technical containment drills are conducted at least quarterly, covering different AI incident scenarios and rotating through all response team members.

WHY *Drill participation that always involves the same people creates single points of failure. Rotating participation ensures broad team competence.*

☐ Threat intelligence integration AI-specific threat intelligence from industry sources, research publications, and incident sharing communities is reviewed and integrated into detection and containment controls.

WHY *AI attack patterns evolve faster than most organizations can develop their own. Threat intelligence integration leverages the broader community's experience.*

☐ Control effectiveness metrics tracked Key metrics for detection accuracy, mean time to detect, mean time to contain, and containment effectiveness are tracked and reviewed quarterly.

WHY *Metrics that are not tracked cannot be improved. Tracking creates accountability and identifies areas for investment.*

☐ Lessons learned database maintained A searchable database of incident lessons, detection rule changes, and containment improvements is maintained and referenced during incident response and planning.

WHY *Lessons that are not documented and accessible are lost when team members change. A maintained database preserves organizational knowledge.*

Get the next one. aminrj.com/subscribe