

AI Agent Incident Response and Recovery Architecture

A Comprehensive Enterprise Architecture for Detecting, Containing, Investigating, Recovering, and Learning from Agentic AI Incidents

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

Agentic AI systems can plan, invoke tools, access enterprise data, delegate tasks, and execute transactions at machine speed. These capabilities create incident patterns that differ from conventional application or cybersecurity events because a compromised, misconfigured, or unsafe agent may continue to reason and act while propagating effects across systems. This paper proposes an AI Agent Incident Response and Recovery Architecture that integrates detection, triage, containment, identity revocation, tool isolation, evidence preservation, rollback, human escalation, recovery validation, and post-incident learning. The architecture defines incident severity, telemetry requirements, forensic evidence, kill-switch patterns, bounded recovery, safe fallback, and business continuity controls. It also addresses multi-agent coordination, prompt injection, unauthorized actions, data leakage, policy bypass, and model or tool failures. The proposed framework enables enterprises to respond to agentic AI incidents with clear accountability, rapid containment, auditable evidence, and resilient restoration of business services.

Keywords — Agentic AI, Incident Response, AI Recovery, AI Security, Autonomous Agents, AI Governance, Forensics, Kill Switch, Business Continuity, Enterprise Architecture

I. INTRODUCTION

AI agents introduce incident scenarios in which the system can continue to take actions after an initial fault or compromise. Response therefore requires controls that can stop delegated execution, revoke authority, preserve evidence, and restore safe service quickly.

Incident response must treat agents as active digital actors with identities, tools, data access, and delegated permissions rather than as passive model endpoints.

II. INCIDENT TAXONOMY

Agent incidents can include prompt injection, unauthorized tool use, sensitive-data disclosure, harmful action, policy bypass, compromised credentials, runaway task loops, incorrect delegation, external model failure, and unsafe autonomous decisions.

A common taxonomy allows consistent severity classification, routing, reporting, and lessons learned.

III. DETECTION AND TRIAGE

Detection should combine agent telemetry, policy violations, identity anomalies, tool-call patterns, abnormal data access, user reports, model evaluations, and business-impact indicators.

Triage should identify affected agents, sessions, credentials, systems, data, business processes, and possible propagation paths.

IV. CONTAINMENT ARCHITECTURE

Containment mechanisms include agent pause, session termination, credential revocation, tool isolation, transaction blocking, network restrictions, and temporary reduction of autonomy.

High-risk environments should support emergency kill switches that operate independently of the model or orchestration layer.

V. EVIDENCE PRESERVATION

Forensic evidence should capture agent identity, prompts, retrieved context, model version, tool calls, authorization decisions, outputs, approvals, timestamps, and downstream effects.

Evidence stores should be integrity protected and access controlled so that investigations and regulatory reviews can rely on them.

VI. RECOVERY AND ROLLBACK

Recovery should restore business service using validated model versions, trusted prompts, clean credentials, approved tools, and verified data sources.

Where reversible transactions are supported, automated rollback can reduce impact; otherwise compensating business actions and human review are required.

VII. MULTI-AGENT INCIDENT HANDLING

Multi-agent systems require containment of delegation chains so that compromised authority does not continue through sub-agents.

Response workflows should identify parent-child relationships, delegated scopes, active tasks, and shared resources before service is restored.

VIII. HUMAN ESCALATION AND DECISION RIGHTS

Incident severity should determine whether business owners, security, legal, compliance, model-risk, and executive stakeholders must participate.

Human decision rights are essential for high-impact recovery choices such as resuming autonomous execution or accepting residual risk.

IX. BUSINESS CONTINUITY AND SAFE FALLBACK

Critical processes should have fallback modes that reduce or remove AI autonomy while preserving essential operations.

Fallback may include deterministic workflows, read-only operation, manual approval, restricted toolsets, or temporary human execution.

X. POST-INCIDENT LEARNING

Root-cause analysis should identify failures in prompts, models, identity, authorization, data, tools, governance, monitoring, or operating procedures.

Findings should update evaluation suites, detection rules, architecture patterns, training, risk scores, and recovery playbooks.

XI. METRICS AND ASSURANCE

Useful metrics include detection time, containment time, recovery time, evidence completeness, repeated incidents, failed rollbacks, unauthorized actions, and percentage of incidents with corrective controls completed.

Periodic exercises and simulations should validate that kill switches, recovery paths, and communication mechanisms work under realistic conditions.

XII. CONCLUSION

AI Agent Incident Response and Recovery Architecture provides a structured approach for controlling agentic AI failures while preserving business resilience.

By integrating identity, telemetry, containment, recovery, and continuous learning, enterprises can reduce the impact of autonomous-system incidents and improve trust in AI operations.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance specialists, security professionals, standards bodies, and technology researchers whose published work informed the concepts synthesized in this paper.

REFERENCES

- [1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
- [2] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
- [3] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
- [4] ISO 31000:2018, Risk management — Guidelines.
- [5] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
- [6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.

- [7] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [9] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [10] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [11] OECD, Recommendation of the Council on Artificial Intelligence.
- [12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.