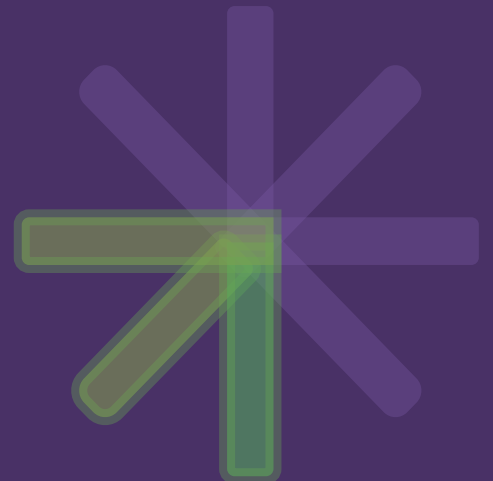

Governing Autonomous AI Agents: Guardrails, Permissions and Oversight Training Course





Governing Autonomous AI Agents: Guardrails, Permissions and Oversight Training Course

Programme Summary

Governing autonomous AI agents is a one-week training course for senior technology, risk and assurance leaders at advanced level, culminating in an AI agent control framework for their own organisation. AI agents now plan tasks, call tools and change records with little human input, so one faulty instruction can spread across systems unnoticed. The approach moves from autonomy levels and agent-specific threats to the design of permissions, guardrails, oversight and assurance. Mars Training Centre UK delivers this programme in governing autonomous AI agents.

The Business Case

Agents promise faster service and lower processing cost, yet most organisations deploy them with the access rights of a trusted employee and the monitoring of a simple chatbot. Without clear autonomy limits, accountable owners and audit trails, boards cannot tell which agent acted, on whose authority or with what data.

- Operational Loss Exposure: Reduced as agent permissions, spending limits and escalation points are fixed before deployment.
- Incident Detection and Containment Time: Shortened as agent actions are logged, monitored and stoppable.
- Assurance Readiness: Improved as agent controls are mapped to recognised AI risk frameworks.

Principles and Evidence Base

- NIST AI Risk Management Framework: Its Govern, Map, Measure and Manage functions organise the control framework built across Sessions 1 to 4.
- ISO/IEC 42001:2023: Its AI management system requirements shape roles, policy and continual improvement for agents in Session 4.
- OWASP Top 10 for LLM Applications: Its treatment of prompt injection and excessive agency informs the threat diagnosis in Session 2.
- OWASP Top 10 for Agentic Applications: Its agent risks, including tool misuse, privilege abuse and cascading failures, structure the testing in Session 2 and Session 3.

Intended Participants

- Leaders who approve the deployment of AI agents into business processes and customer services.
- Executives accountable for technology, data or digital strategy across the organisation.
- Heads of risk and information security who set limits and controls for automated decision-making.
- Internal audit leaders who must give assurance over AI-enabled processes.



The programme serves banking and insurance, government services, telecommunications and energy organisations.

Learning Outcomes

1. Participants will be able to classify AI agent use cases by autonomy level and business consequence.
2. Participants will be able to diagnose agent-specific threats, including prompt injection, tool misuse and excessive agency.
3. Participants will be able to design identity, least-privilege permissions and guardrails for AI agents.
4. Participants will be able to specify human-in-the-loop checkpoints, stop mechanisms and escalation rules.
5. Participants will be able to evaluate monitoring, logging, testing and third-party assurance for agent platforms.
6. Participants will be able to construct an AI agent control framework aligned with the NIST AI RMF and ISO/IEC 42001:2023.

Programme Structure

Day 1: Governing Autonomous AI Agents and Autonomy Levels

What makes an AI agent different to govern from other AI systems?

- Agent Architecture: Planning, Memory, Tools and Actions
- Autonomy Levels and Delegated Authority Mapping
- Use Case Inventory and Consequence Classification
- NIST AI RMF Govern Function Applied to Agents
- Accountable Owners and Agent Registers

Day 2: Diagnosing Agent Threats and Failure Modes

How can an AI agent be misled, misused or cause cascading harm?

- Prompt Injection and Indirect Instruction Attacks
- Excessive Agency and Tool Misuse Scenarios
- Memory Poisoning and Data Leakage Paths
- Cascading Errors in Multi-Agent Workflows
- Agent Threat Model Using the OWASP Agentic Top 10

Day 3: Designing Permissions, Guardrails and Oversight

Which controls keep an agent inside its intended authority?



- Agent Identity, Credentials and Least-Privilege Access
- Input, Output and Action Guardrails
- Human-in-the-Loop Checkpoints and Approval Thresholds
- Kill Switches, Rate Limits and Rollback Design
- Red Teaming and Pre-Deployment Testing Plans

Day 4: Monitoring, Assurance and Third-Party Platforms

How is agent behaviour evidenced, audited and managed over time?

- Action Logging, Audit Trails and Traceability
- Runtime Monitoring Metrics and Anomaly Alerts
- Agent Incident Response and Post-Incident Review
- Third-Party Agent Platform Due Diligence and Contract Terms
- ISO/IEC 42001:2023 Management System Integration

Day 5: Applying Agent Governance to the Organisation

Which agents should proceed, under what controls and with what evidence?

- Autonomy Decisions for Day 1 Use Cases
- Control Selection Against the Day 2 Threat Model
- Assurance Plan Using Day 4 Monitoring Measures
- Board Reporting and Risk Appetite for AI Agents
- AI Agent Control Framework Completion

Methodology

The programme relies on structured exercises built around case agents in customer service, finance operations and IT support. Each session moves from a short briefing to group analysis, then tests proposed controls against attack and failure scenarios.

- Suggested activity: Mapping the permissions of a procurement agent and removing unnecessary access.
- Suggested activity: Tracing an indirect prompt injection through a document-processing agent.
- Suggested activity: Designing approval thresholds for an agent that issues refunds.

Capstone Output

Each participant completes an AI agent control framework for their own organisation. It contains an agent register, autonomy classifications, a threat model, permission and guardrail standards, oversight checkpoints and an assurance plan. It can be adopted as policy or used to decide whether an agent goes live.



Frequently Asked Questions

Is governing autonomous AI agents training suitable for leaders without a technical background?

Yes. It is set at advanced level for executives who approve or assure AI agents, and assumes familiarity with risk or technology governance.

Which frameworks does AI agent governance training rest on?

It rests on the NIST AI Risk Management Framework, ISO/IEC 42001:2023, the OWASP Top 10 for LLM Applications and the OWASP Top 10 for Agentic Applications.

How is the AI agent control framework used at work?

You can use it as the basis for an agent policy, to approve or refuse a proposed deployment, and to plan assurance work.

How does governing AI agents differ from AI ethics programmes?

It concentrates on operational control of agents that act on systems and data, rather than on ethical principles for AI in general.