JUPITER PRIME

Consulting • Training • Advisory

جوبيتر برايم للاستشارات والتدريب

COURSE BROCHURE

# Advanced AI Agent Security Training Course

**jupiterprime.uk**

info@jupiterprime.uk

Registered in the United Kingdom and Saudi Arabia

# Advanced AI Agent Security Training Course

AI Agent Security is a 2-week advanced training course for security architects, AI engineers and cyber risk managers who will produce an agent threat model and security control plan for an autonomous system their organisation runs or intends to run. Agents that browse, call tools, write code and move money create attack paths that conventional application security never anticipated. Jupiter Prime Training delivers this AI Agent Security course to technical security leads.

## Why AI Agent Security Matters Now

Three shifts drive this discipline: language-model agents gaining real permissions over email, code repositories, payments and customer records; attackers learning to hide instructions in web pages, documents and tool outputs; and boards asking who is accountable when an autonomous system acts wrongly. An organisation that secures agents well can grant them useful autonomy with contained blast radius; one that does not invites data exfiltration, fraud and unauthorised actions taken in its name. The course anchors controls in ISO/IEC 42001, the AI management system standard, and the NIST AI Risk Management Framework NIST AI RMF.

## Who Gains Most from AI Agent Security

- Architects who design how agents connect to tools, data stores and identity systems
- Engineers who build, orchestrate and deploy language-model agents
- Analysts who monitor detections, investigate incidents and respond to agent misuse
- Managers who assess cyber risk and approve agent permissions and autonomy levels
- Testers who run adversarial evaluations and red-team exercises

A security engineering function gains most when its architects, agent developers and detection staff attend together, because permission design, prompt-level defences and runtime monitoring only stop an attack when all three layers share the same threat assumptions.

## AI Agent Security Capabilities

Participants learn to treat every agent as a privileged, manipulable insider and to engineer layered defences around its reasoning, its tools and its identity.

- Build an agent threat model that maps trust boundaries, tool calls and untrusted inputs — security architecture boards
- Test agents against direct and indirect prompt injection, jailbreaks and goal hijacking — red-team and assurance leads
- Design least-privilege tool permissions, scoped credentials and human approval gates — identity and access owners
- Specify sandboxing, egress filtering and output validation to block data exfiltration — data protection officers
- Apply runtime monitoring, action logging and kill-switch controls to live agents — security operations centres
- Assess residual agent risk and autonomy levels against ISO/IEC 42001 and NIST AI RMF — cyber risk committees

## Key AI Agent Security Terms

- AI agent — a system that uses a language model to plan and take actions through tools, APIs or other software without step-by-step human direction.
- Prompt injection — an attack that inserts malicious instructions into model input so the agent ignores its intended task or rules.
- Indirect prompt injection — hidden instructions placed in web pages, files or tool outputs that an agent later reads and obeys.
- Tool permission scoping — limiting which functions, data and actions an agent may invoke, and under what conditions.
- Agent sandbox — an isolated execution environment that restricts an agent's file, network and process access to contain damage.
- Excessive agency — the risk created when an agent holds more autonomy, permissions or functionality than its task requires.

## AI Agent Security Day by Day

## Day 1: The Agentic Threat Landscape

- Agent Architectures: Planner, Memory and Tool Loops — Attack Surface Inventory
- Autonomy Levels Taxonomy — How Much Freedom Each Use Case Warrants
- Recent Agent Compromise Patterns — Lessons From Public Incidents
- Attacker Motivations Against Agents — Fraud, Espionage and Sabotage
- Agent Asset Register — Which Systems, Data and Credentials Agents Touch

## Day 2: Threat Modelling for Autonomous Systems

- Trust Boundary Diagrams for Agent Workflows — Where Untrusted Input Enters
- Data Flow Mapping Across Tools and Memory — Exposure of Sensitive Records
- Abuse Case Writing for Agents — Misuse Scenarios Before Build
- Attack Trees for Goal Hijacking — Prioritising Defensive Investment
- Threat Model Review Cadence — Updating as Tools and Prompts Change

## Day 3: Prompt Injection and Manipulation Defences

- Direct Injection and Jailbreak Techniques — Recognising Bypass Patterns
- Indirect Injection via Documents, Web and Email — Content Trust Rules
- Instruction Hierarchy and System Prompt Hardening — Precedence Decisions
- Input Classifiers and Guard Models — Filtering Hostile Content
- Spotlighting and Data-Instruction Separation — Reducing Confusion Attacks

## Day 4: Tool, Plugin and Permission Abuse

- Least-Privilege Tool Design — Narrow Functions Over General Capability
- Parameter Validation and Allow-Lists — Blocking Dangerous Arguments
- Confused Deputy Risks in Tool Chains — Cross-Tool Privilege Escalation
- Human Approval Gates for High-Impact Actions — Payments, Deletions and Messages
- Third-Party Plugin and Connector Vetting — Supply Chain Exposure

## Day 5: Agent Identity and Access Control

- Non-Human Identity Lifecycle for Agents — Issuance, Rotation and Revocation
- Scoped, Short-Lived Credentials and Token Exchange — Limiting Stolen Access
- Delegated Authority and On-Behalf-Of Flows — Acting for Users Safely
- Multi-Agent Trust and Inter-Agent Authentication — Preventing Impersonation
- Zero Trust Policy Enforcement for Agents — Continuous Authorisation Decisions

## Day 6: Data Exfiltration and Memory Protection

- Exfiltration Channels: Links, Images, Tool Calls and Logs — Leak Paths
- Egress Filtering and Output Validation — Stopping Covert Data Release
- Retrieval Store and Vector Database Access Controls — Document-Level Permissions
- Memory Poisoning and Persistence Attacks — Long-Term Agent Integrity
- Sensitive Data Redaction in Context Windows — Privacy by Design

## Day 7: Sandboxing and Execution Isolation

- Containerised and Virtualised Agent Runtimes — Isolation Strength Choices
- Code Execution Sandboxes for Agents — File, Network and Process Limits
- Browser Agent Isolation — Handling Hostile Websites
- Resource Quotas and Rate Limits — Preventing Runaway Agents
- Environment Separation for Testing and Production — Safe Promotion Paths

## Day 8: Red-Teaming and Adversarial Evaluation

- Agent Red-Team Planning and Rules of Engagement — Scope Decisions
- Automated Adversarial Prompt Generation — Scaling Attack Coverage
- Benchmark Suites for Injection and Tool Misuse — Measuring Robustness
- Finding Severity Rating for Agent Flaws — Remediation Priority
- Regression Testing After Model or Prompt Updates — Preserving Defences

## Day 9: Runtime Monitoring and Incident Response

- Agent Action Logging and Trace Capture — Forensic Evidence
- Behavioural Anomaly Detection on Tool Calls — Early Compromise Signals
- Kill Switches and Circuit Breakers — Halting Misbehaving Agents
- Agent Incident Playbooks — Containment, Eradication and Recovery
- ISO/IEC 42001 and NIST AI RMF Controls Mapping — Accountability Evidence

## Day 10: Practising the Agent Threat Model and Security Control Plan

- Threat Modelling Applied to a Participant's Agent Workflow — Risk Priorities
- Permission Scoping Applied to Its Tool Set — Least-Privilege Decisions
- Red-Team Findings Applied to Defence Selection — Control Gaps
- Monitoring Design Applied to Production Deployment — Detection Coverage
- Agent Threat Model and Security Control Plan Assembly — Owners, Controls and Review Points

## Applying AI Agent Security After the Course

- Produce a trust boundary diagram and abuse case list for one agent already in development and present the top five risks to its product owner.
- Run an indirect prompt injection test using planted instructions in documents the agent reads, and log which defences held and which failed.

## AI Agent Security Questions FAQ

### What is AI agent security, and why does it matter now?

It is the discipline of protecting autonomous language-model agents from manipulation, misuse and data leakage. It matters now because agents increasingly hold real permissions over code, payments and records, and attackers already hide instructions in content those agents read.

### Which roles gain most from AI agent security training?

Experienced security and AI practitioners gain most: security architects, agent developers, detection engineers and cyber risk managers. You benefit if you design, build, approve or monitor agents that act on behalf of your organisation.

### How does advanced AI agent security differ from a general cyber security course?

It focuses on threats specific to autonomous AI. A general cyber security course covers networks, endpoints and applications, while this course addresses prompt injection, tool abuse, agent identity, sandboxing and red-teaming of systems that reason and act.

## What do you bring back from AI agent security training?

You bring back an agent threat model and security control plan. It maps trust boundaries and attack paths, assigns permission, isolation and monitoring controls, records red-team findings and names the owners and review points for each defence.

## AI Agent Security: The Way Forward

The agent threat model and security control plan first informs whether an agent may move from testing to production and with what permissions. Security architects use it to approve designs, engineers to implement controls, and risk managers to accept residual exposure. After the first production incident or major model update, revisit the autonomy level assigned, as real behaviour often differs from design.