AI Agents: Security & Controls

Taught by an engineer who builds agents. What an agent really does when it acts on your systems, how it breaks, and what to demand before you sign it off.

WHO IT'S FOR	FORMAT	LENGTH
Compliance, legal, risk and security-governance teams asked to review AI agents and automation before they go live.	In-house workshop	Half-day

What your team walks away with

Review an AI agent before it is approved — how it gets attacked, how it fails silently, and which controls actually hold.

What you'll learn

- ✓ Describe in plain terms what an AI agent can reach, call and change inside your systems
- ✓ Recognise the attack paths that matter — prompt injection, data exfiltration and excessive permissions
- ✓ Explain why a passing test run is weak evidence that an agent will behave in production
- ✓ Specify the permission limits, approval gates and human checks an agent needs before approval
- ✓ Challenge a vendor's AI security claims and tell a real assurance from a marketing one
- ✓ Run every future agent proposal through a consistent, defensible approval checklist

Curriculum

01 What an AI agent actually does when it acts on your systems

- Chatbot versus agent: the difference is the ability to take action, not the quality of the writing
- Tools, connectors and permissions — the concrete list of what an agent can reach and change
- Autonomy levels: suggesting, drafting, acting with approval, acting alone
- Where agents are showing up in regulated firms: onboarding, screening, monitoring, reporting, service
- Mapping an agent's blast radius: the worst thing it could do if every judgement went wrong

02 Prompt injection, data exfiltration and the over-permissioned agent

- Prompt injection explained without jargon: instructions hidden inside the data an agent reads
- Why an agent cannot reliably tell your instructions apart from content it processes
- Data exfiltration paths — how confidential information leaves through an agent's own tools
- The over-permissioned agent: broad access granted for convenience during a pilot and never revoked
- Published references worth knowing: the OWASP Top 10 for LLM applications and MITRE ATLAS

03 Why you cannot test an agent the way you test software

- Non-determinism: the same input can produce different actions on different runs
- Why a green test suite is weaker evidence for an agent than for conventional software
- Evaluation sets, sampling and red-teaming — what each one does and does not prove
- Silent failure: agents that produce confident, well-formatted, wrong output
- What assurance to ask the business for, given that exhaustive testing is not available

04 Designing the leash: permissions, approval gates and human-in-the-loop that hold

- Least privilege for agents: scoping access to the specific task, not the whole system
- Choosing the actions that must never happen without a named human approval
- Rate limits, value thresholds and reversibility as controls rather than nice-to-haves
- Designing human review people will actually perform, instead of rubber-stamping
- Logging and traceability: reconstructing what an agent did, and why, months later

05 Reading a vendor’s AI security claims — and what to ask instead

- What common security certifications do and do not say about agent behaviour
- Separating model-provider assurances from the vendor's own application controls
- Questions on data handling, retention, training use and sub-processors that get real answers
- Where liability actually sits when a vendor's agent causes a loss
- Adding AI-specific items to your existing third-party risk questionnaire

06 An agent approval checklist your team applies to every deployment

- Building the checklist: purpose, data reached, actions permitted, human gates, logging, owner
- Risk-tiering agents so low-stakes automation is not reviewed like a payment system
- Setting conditions of approval and the triggers for re-review
- Monitoring after go-live and defining what counts as an AI incident
- Walking through the checklist against a real or realistic agent from your firm

You keep

> An agent approval checklist and control map for your review process.