InfoSecAI

A BRIEF FOR CISOS · BISOS · GRC LEADERSHIP

WHITE PAPER · PAPER 3 · 2026

Securing Agentic AI Before It Acts

A practitioner's brief for CISOs, CTOs, CIOs, identity and access leaders, security architects and AI platform owners building the control plane that lets AI agents do useful work safely.

AUTHORED BY

Paul Jolliffe

By Paul Jolliffe, Founder and Director, InfoSecAI Limited · MBA · CISSP · ISO 27001 Lead Auditor

01 · EXECUTIVE SUMMARY

Executive summary

Artificial intelligence (AI) is moving from chat to action. Tools that previously suggested content now schedule meetings, send messages, retrieve customer records, write to production systems, trigger payments and orchestrate multi-step workflows. The change is not incremental. An agent that can act is a different kind of system from a model that can talk, and it needs a different kind of control.

The governing question is no longer "can this AI agent do the task?". The real question is what it is allowed to see, decide, change and trigger, and whether the organisation can prove it stayed within those boundaries.

This brief is the third in InfoSecAI's five-part executive series, From AI Ambition to AI Assurance. The first two papers covered the operating model and the discovery of the AI estate. This paper extends both into agentic AI: the control plane, the permission model, the human oversight design and the kill-switch discipline that an enterprise needs before an AI agent has an action button.

The thesis is direct. Agent risk is a function of two dimensions: autonomy and access. Prompt engineering is not a substitute for permissions. Static application controls are not a substitute for a control plane. The work is to build that plane now, while the market is still defining what an agent is, rather than after the first material incident.

02 · WHY THIS MATTERS NOW

Why this matters now

Three forces are converging in agentic AI faster than enterprise governance is keeping up.

The first is product velocity. Generative AI (GenAI) platforms shipped agentic capabilities through 2025 and the first half of 2026 with limited fanfare. Tool-use, retrieval, multi-step planning, browser automation and computer-use features are now default in mainstream platforms. Many enterprises have agents available to staff today without an explicit deployment decision.

The second is the security research consensus. The Open Worldwide Application Security Project (OWASP) Top 10 for Large Language Model Applications (2025 edition) names prompt injection as LLM01, sensitive information disclosure as LLM02 and excessive agency as LLM08. Excessive agency is the agentic AI risk specifically: the agent has more functionality, permissions or autonomy than the use case requires, and uses it under attacker influence. The United Kingdom National Cyber Security Centre (NCSC) Guidelines for Secure AI System Development make the same point in different language: secure deployment requires explicit constraint, not implicit trust.

The third is the regulatory clock. The European Union Artificial Intelligence Act (EU AI Act, Regulation (EU) 2024/1689) applies human-oversight obligations under Article 14 from 2 August 2026 to high-risk AI systems. Article 14 requires that natural persons can oversee

the operation of the AI system, intervene during operation, and interrupt the system through a stop button or similar means. An agent without an enforceable kill-switch does not satisfy Article 14, irrespective of how skilful the model is.

The combination is unforgiving. Capability outpaces control, controls that exist were designed for humans, and the regulatory expectation is now that the human oversight is real and provable.

03 · WHY CURRENT APPROACHES DO NOT FIT

Why current approaches do not fit

Three control patterns the security profession relies on do not extend cleanly to agentic AI.

The first is human-centric identity and access management (IAM). Role-based access control (RBAC) assumes a human user with a role; agents are not humans, they act on behalf of one or many users, and the trust assumptions break. A user permitted to read customer records does not necessarily intend their agent to read them in bulk, summarise them and send the summary to a third party.

The second is static application control. Web application firewalls, output filters and prompt templates are point controls. They do not see the agent's full execution graph: the tools it called, the data it retrieved, the decisions it made, the next action it queued. Agentic AI needs a control plane that sees the whole sequence, not a wrapper that sees the prompt.

The third is the prompt-engineering reflex. There is a persistent belief that the right system prompt makes the agent safe. It does not. The Open Worldwide Application Security Project's 2025 list names prompt injection as the top risk for a reason: instructions in the data the agent reads can override instructions in the system prompt, and there is no current cryptographic separation between the two. A prompt is a hint, not a boundary.

The structural fix is to treat the agent as a system that requires identity, permissions, monitoring, approval workflows, audit and a kill-switch, in the same way any other system handling sensitive data would.

04 · THE TWO DIMENSIONS: AUTONOMY AND ACCESS

The two dimensions: autonomy and access

Agent risk is best read on two axes.

Autonomy is the question of who decides. An agent at the lowest autonomy level only observes data and reports it to a human. The next level advises a human and waits for the human to act. The next level acts with explicit human approval per action. The highest level acts autonomously within a defined scope.

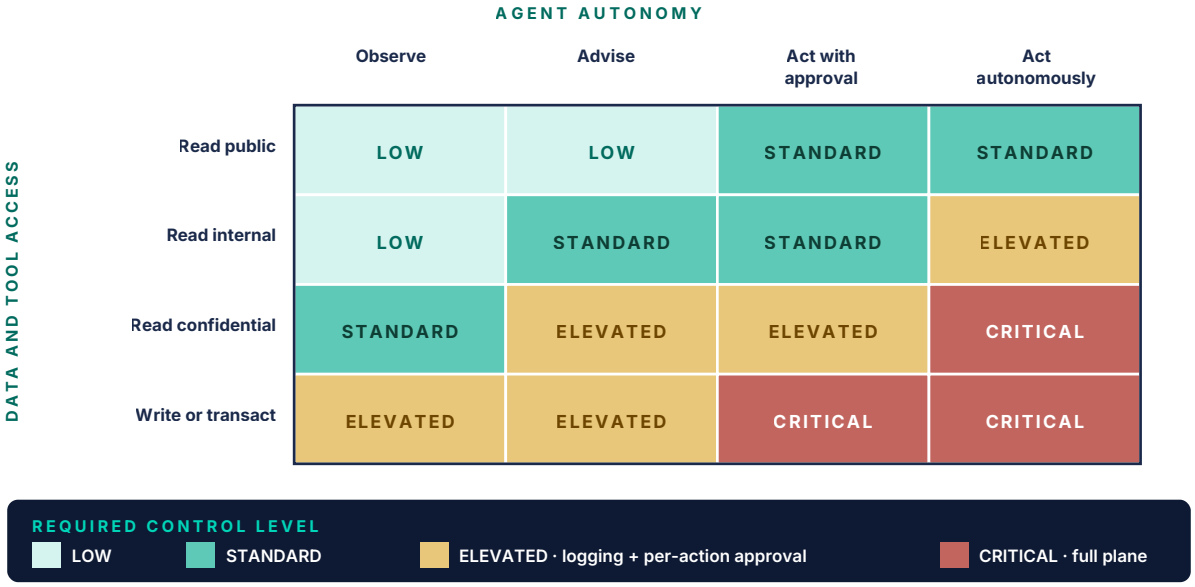
Access is the question of what the agent can touch. Read access, write access, external-action access (sending messages, triggering payments, retrieving regulated data) all scale the risk. An agent reading public web content has different risk from an agent reading client records; an agent that can also send email on the user's behalf is different again.

Risk is the product, not the sum. A low-autonomy agent with sensitive data access is still high risk because of the data path. A high-autonomy agent with weak monitoring is high risk because no one will know what it did. A low-autonomy agent with low access is genuinely low risk; a high-autonomy agent with high access is the case that needs the most deliberate control.

This is what the agent autonomy-access matrix below makes explicit.

FIGURE 1 · Agent autonomy and access risk matrix.

Cell colour = required control level. Critical cells need the full control plane before deployment.



05 · THE PERMISSION LIFECYCLE

The permission lifecycle

Every agent permission has a life: it is requested, reviewed, granted, exercised, monitored, reviewed again and eventually revoked. Most current deployments handle the first three steps and the fifth, and skip the others.

A permission lifecycle for agents has six stages. Request, where a use case is logged with the data and actions needed. Review, where a named owner assesses the request against the autonomy-access matrix and the supplier's representations. Grant, where the permission is provisioned narrowly: the smallest set of tools, the shortest validity, the most restrictive scope. Exercise, where the agent operates and every action is logged at the prompt, tool-call and output level. Monitor, where the security operations function alerts on anomalous patterns, including tool-call sequences that do not match the use case. Revoke, where permissions are removed when the agent is retired, when the use case ends, or when the periodic review fails.

FIGURE 2 · The agent permission lifecycle.

Six stages, named owners, two agent-specific extensions: tool-scope minimisation and per-action logging.

The discipline transfers from human IAM with two extensions. Tool scope replaces role: an agent needs the tools the use case requires and no others. Per-action logging replaces session logging: every tool call, every retrieval, every action queued or executed is captured. The standard of evidence is "could the security team reconstruct what the agent did?". If the answer is no, the logging is insufficient regardless of volume.

06 · THE HUMAN-IN-THE-LOOP DECISION TABLE

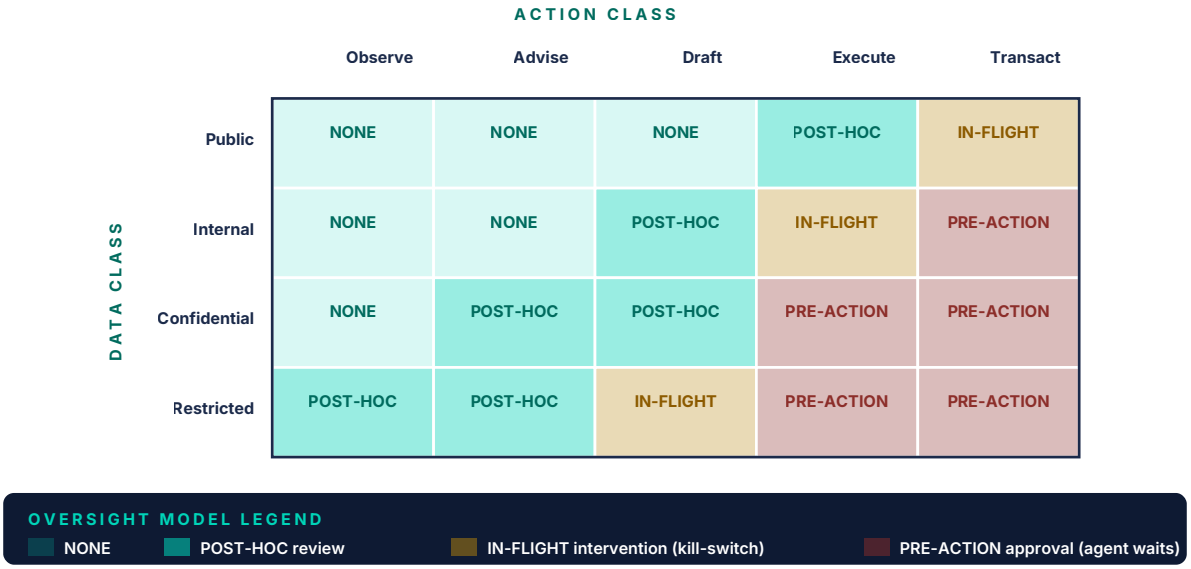
The human-in-the-loop decision table

EU AI Act Article 14 requires natural persons to oversee high-risk AI systems and to intervene meaningfully. For agentic AI, "meaningfully" means the human approves the action before it occurs, the human can interrupt the agent in flight, or the human reviews the action immediately afterwards.

The decision of which model applies is not the developer's. It is a governance decision driven by the action class and the data path.

FIGURE 3 · The human-in-the-loop decision table.

Action class × data class. Cell value = required oversight model.



Three principles operate the table.

The agent's action class is determined by the system designer, not the model. An agent that can technically send emails but is configured not to is, for the purposes of the decision table, an "advise" agent. Designers must declare the action class as a configuration, not infer it from the model's capabilities.

Pre-action approval is required wherever a single action could produce material harm that cannot be undone by a post-hoc reversal. Sending a regulatory filing, executing a trade, deleting a customer record, sending a customer-facing communication: all are actions where a post-hoc review is too late.

In-flight intervention requires a working kill-switch. Not the model provider's API; the organisation's own enforced stop, with the latency to match the harm window. For most enterprise agents, the target latency is under sixty seconds from a security operations alert.

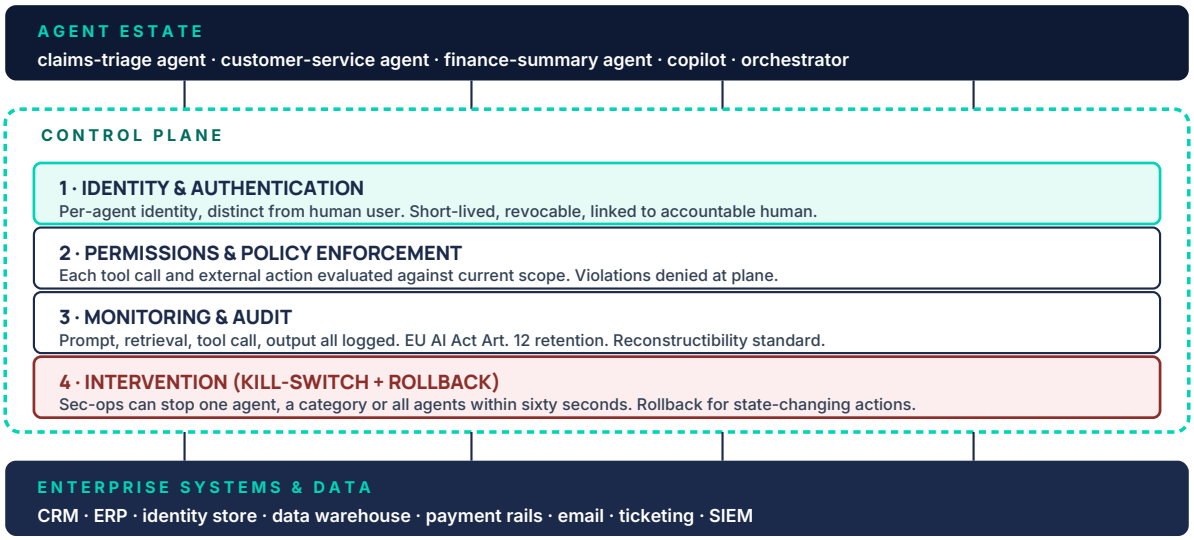
07 · THE AGENT CONTROL PLANE

The agent control plane

A control plane is the runtime layer that mediates between agents and the rest of the organisation. It is the architectural pattern that turns the principles above into operational controls.

FIGURE 4 · The enterprise agent control plane.

Four layers between agents and enterprise systems. Every agent routes through the plane.



Identity. Each agent has its own identity, distinct from the human user and from other agents. Authentication is short-lived, scoped to the use case, revocable, and linked to the accountable human.

Permissions and policy enforcement. The plane evaluates each tool call and external action against the agent's current permissions. Violations are denied at the plane, not detected after the fact. This is where excessive agency is prevented operationally.

Monitoring and audit. Prompt, retrieval, tool call and output are all logged at a granularity sufficient for incident reconstruction. EU AI Act Article 12 mandates retention for high-risk systems; six months is the pragmatic minimum for agents handling confidential or restricted data.

Intervention. The kill-switch is enforced at the plane, not at the model. Security operations must be able to stop a specific agent, a category, or all agents, with auditable latency, without depending on the model provider. Rollback capability for state-changing actions is the additional control.

A control plane is the first thing internal audit asks about when scoping an agentic-AI review. Organisations without one operate their agents on trust.

08 · INFORMATION SECURITY IMPLICATIONS

Information security implications

Six integration points carry the largest exposure.

Identity propagation. When an agent acts on behalf of a user, both identities must appear in every audit record. Logging only one breaks either attribution or misuse detection.

Tool permission scoping. The agent's tools are its blast radius. Enable only what the use case strictly requires; review the new risk before adding any tool to scope.

Prompt injection mitigation. There is no current technique that eliminates it. Treat any externally-sourced text as untrusted, sanitise tool inputs, constrain output formats, and require pre-action approval for any action triggered by externally-sourced instructions.

Output handling. Agent outputs passed downstream as instructions become injected prompts for the next agent. Type the boundary between agent output and next-system input as untrusted by default.

Logging and audit. The standard is reconstructibility. Pick a random agent action from last week and ask the security team to reconstruct what happened. If they cannot, the logging is inadequate.

Incident response. The incident classification standard operating procedure (SOP) needs an agentic-AI branch with the kill-switch as the first step. Triggers include anomalous tool-call patterns, prompt-injection markers in retrieved text, unauthorised tool invocations, and agent persistence beyond the intended use case.

09 · EXECUTIVE DECISIONS AND 30-DAY READINESS

Executive decisions and 30-day readiness

Six decisions unblock the control plane build.

Who is the accountable executive for agent operations? One named individual, typically the chief information security officer (CISO), with a delegate operating the plane day to day.

What is the maximum agent autonomy without case-by-case approval? Default ceiling: "advise" or "act with approval"; "act autonomously" requires named board sign-off.

What tools may agents access by default? A short standing list, reviewed quarterly. Anything outside requires one-off authorisation.

What is the maximum latency from alert to kill-switch execution? Sixty seconds is the pragmatic target; tighter for agents touching regulated data or executing transactions.

What is the rollback model for state-changing actions? Documented per action class, with named operators authorised to perform the rollback.

What evidence proves the control plane is operating? Three artefacts: the agent register, the most recent quarterly permission review, and a sample reconstruction of a single agent action.

The first thirty days do not require a new platform. They require a register, a policy, a kill-switch, a logging baseline and an owner.

Week one: build the agent register from the AI inventory (Paper 1) and the shadow AI map (Paper 2), plus a direct survey of teams in pilot.

Week two: assign every agent a cell in the autonomy-access matrix. Identify the elevated and critical cells; those are the priority.

Week three: validate the kill-switch on each elevated and critical agent. Demonstrate operator-initiated stop within target latency, with an audit record. Untested means non-functional.

Week four: reconstruct a recent agent action end to end from the logs. If reconstruction fails, fix the logging before the next agent is added.

Worked example. Eastdale Mutual plc, a mid-sized UK mutual insurer, ran the sprint in May 2026 to prepare for an agentic claims-triage pilot. Twenty-three agents surfaced across operations, claims, customer service and finance: fifteen observe-only with standard logging; six advise-class with documented controls; two act-with-approval held pending kill-switch validation (one failed first test, passed after remediation); none act-autonomously. The control plane went into production with a named operations team and a quarterly review cadence.

10 · QUESTIONS EVERY LEADER SHOULD ASK NOW

Questions every leader should ask now

For the three most autonomous agents in the estate, can the named accountable executive be identified in under one hour, the action class and access class produced, and the most recent permission review evidenced? If not, the register is not operating.

Has the kill-switch been tested on each elevated-cell agent in the last ninety days? Untested means non-functional.

For a randomly selected agent action from the last seven days, can the security team reconstruct the full execution: prompt, retrievals, tool calls, decisions, outputs? If not, the audit standard is not met.

What is the single agentic-AI decision that has been pending the longest, and what evidence would unblock it? That decision is the priority for the next executive meeting.

If a regulator asked for the agent register and the human oversight design under EU AI Act Article 14 next Monday, what is the gap?

11 · CLOSING THOUGHT

Closing thought

Agentic AI is the moment the security profession's existing controls reach their edge. Identity, permissions, monitoring and intervention all still apply; they just have to be re-engineered to operate around an actor that is not a person, that does not have a single intent, and that can be redirected by data it reads. The work is the control plane.

The thread of this series continues: AI assurance evidence, not AI reassurance narrative. An organisation that can prove what its agents did is one that can operate them. An organisation that cannot, is operating on hope.

The next paper, Why AI Transformation Fails After the Pilot, looks at the operating-model failure pattern that stalls enterprise AI even when the controls are in place.

12 · SOURCE REGISTER

Source register

All sources verified to primary publisher on 3 June 2026.

#	SOURCE	USE IN PAPER	LINK
1	Regulation (EU) 2024/1689 (the EU AI Act)	Article 14 human oversight, Article 12 logging, Article 13 transparency, high-risk applicability	https://eur-lex.europa.eu/eli/reg/2024/1689
2	OWASP Top 10 for Large Language Model Applications, 2025	LLM01 prompt injection, LLM02 sensitive information disclosure, LLM08 excessive agency	https://genai.owasp.org/llm-top-10/
3	NCSC, Guidelines for Secure AI System Development	Secure deployment and operation; explicit constraint principles	https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
4	NIST AI Risk Management Framework (AI RMF 1.0)	Govern, Map, Measure, Manage functions applied to agentic systems	https://www.nist.gov/itl/ai-risk-management-framework
5	NIST AI 600-1, Generative AI Profile, July 2024	Generative AI specific risks including confabulation, content provenance, value-chain risk	https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
6	ISO/IEC 42001:2023, AI Management System Standard	AIMS Annex A controls applied to agent operations	https://www.iso.org/standard/42001
7	European Commission, AI Act regulatory framework	Provider versus deployer obligations for agentic systems	https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

13 · ABOUT THIS SERIES

About this series

From AI Ambition to AI Assurance is a five-paper executive briefing series, 1 to 5 June 2026.

1. AI Governance Is No Longer a Policy Problem
2. The Shadow AI Exposure Map
3. Securing Agentic AI Before It Acts (this paper)

4. Why AI Transformation Fails After the Pilot
5. The Board Pack for AI Assurance

Each paper is published as a 12-page executive briefing under the InfoSecAI Blog Template. The full series is available at infosecai.net/insights for subscribers to the InfoSecAI insights list.

14 · PRACTITIONER NOTE

Practitioner note

This briefing is practitioner interpretation, not legal advice. For regulated deployments, validate final claims against current legal obligations, sector-specific requirements and the original primary sources before relying on them.

InfoSecAI

About InfoSecAI

InfoSecAI is an independent UK consultancy helping organisations turn security, regulatory, resilience and AI governance requirements into practical operating models, stronger controls and robust delivery.

We work across strategy, governance, risk, compliance, AI security, assurance, operations and engineering. Our services help leadership teams assess their current position, align to standards and regulation, define the target operating model, and deliver the governance, controls, artefacts and ways of working needed to move from intent to implementation.

Our toolkit capability accelerates structured work across ISO 27001, ISO 22301, ISO 42001, NIST CSF, NIST AI RMF, CIS Controls, Cyber Essentials, DORA, NIS 2, the EU AI Act, GDPR, UK GDPR, SOC 1 and SOC 2. The approach combines AI-enabled workflow support with senior practitioner judgement, so outputs remain proportionate, usable and connected to the way the organisation actually operates.

InfoSecAI was founded in **2025** by **Paul Jolliffe**. The company is built for organisations that need clarity, senior leadership and hands-on delivery across information security and AI governance, without adding unnecessary complexity or treating compliance as a paperwork exercise.

infosecai.net · paul.jolliffe@infosecai.net