AI Agent 365 Security Pattern

A baseline security architecture for operating Microsoft Agent 365 and Foundry agents

Type: Security design pattern **Status:** Baseline **Version:** 1.0 **Date:** November 2026 **Classification:** Internal

1. Purpose and scope

This pattern sets a baseline for how AI agents built on Microsoft Agent 365 and Azure AI Foundry are allowed to operate, so that a single failed control does not become a breach. It is written to be reused across engagements and is independent of any one organisation.

It covers the identity model for agents, the guardrails that keep an agent from acting outside its remit, the data boundary, the telemetry that gives a security operations team visibility, and the mechanism to stop an agent that misbehaves. It assumes an organisation that already has Microsoft Entra, Microsoft Purview, Microsoft Defender and Microsoft Sentinel in place, and an AI policy that governs acceptable use. It does not restate that policy; it is the technical and architectural control set that sits beneath it.

Apply this pattern before the first agent reaches production. The cost of retrofitting identity scope, egress control and telemetry to a fleet of live agents is far higher than designing them in at the start.

2. Why agents need their own pattern

An AI agent is not an application and not a user. It is a new kind of actor: it is given a goal, it reads text from sources that may be outside the organisation's control, and it can call tools that touch real systems, all under an identity of its own. Two properties make it a distinct security problem.

Agent authorisation is not human authorisation. An agent runs with its own permissions, which are usually broader than those of the person who triggered it. Agent 365 gives each agent a Microsoft Entra Agent ID, and an agent can be instanced many times from a single blueprint, so every running instance must be individually identifiable and individually constrained. Treating agents as if they were service accounts, or sharing one identity across instances, loses both attribution and containment.

The agent is steered by text it does not vet. A language model cannot reliably separate instructions from data once both are in its context. Every document, web page, email and tool response an agent reads is therefore an input channel for whoever wrote it. The classic failure is the confused deputy: a hostile instruction is planted in a source the agent later reads, and the agent carries it out using its own legitimate access. Nothing is hacked; the agent does what it was told, by the wrong party.

The pattern below is built around those two facts. The model is rarely the weak point. The configuration, the identity and the wiring to tools and data are.

3. The pattern at a glance

GOVERNING PRINCIPLES

1. **Every agent instance has a unique, least-privilege identity.** One Entra Agent ID per instance, scoped to the minimum it needs, with agent authorisation kept separate from human authorisation.
2. **An agent can reach only the tools and data it is explicitly granted.** Tools are a permission, not a convenience; the default is none.
3. **Untrusted input cannot change what the agent is allowed to do.** Content and prompt-injection controls on inputs, and no path from a read to an unapproved action.
4. **Several independent controls must fail before harm occurs.** Identity, tool scope, egress, approval and detection are layered, not stacked on one point.
5. **Every agent action is visible and attributable, and can be stopped.** Telemetry into the SIEM, detections on agent behaviour, and a working kill switch.

4. Identity and access

This is the single most important control set. If the identity is narrow, most other failures are survivable.

4.1 Unique identity per instance

Agent 365 is built on Microsoft Entra Agent ID. The identity chain runs blueprint, to blueprint principal, to agent identity, to agent user. Require that every agent provisioned to run a task has its own Entra Agent ID, not a shared one, so that its actions are attributable and its access can be revoked in isolation. Register every agent in the Entra registry so there is a single source of truth and no shadow agents.

4.2 Least privilege, separated from human roles

Grant each agent identity the minimum permissions for its task. Do not reuse human roles or broad service principals. An agent that can barely reach anything is a confused deputy that can barely do harm. Keep the role-based access given to agents on a separate footing from the role-based access given to people, and review it independently.

4.3 Conditional Access for agents

Apply Microsoft Entra Conditional Access to agent identities, not only to users. Constrain where an agent can authenticate from, require the conditions you would require of any privileged principal, and treat the agent's sign-ins and risk score as first-class signals.

5. Guardrails: keeping agents from going rogue

Containment works at two levels. The platform provides controls; the architecture around the agent provides the rest. Both are required.

Requirement	Control
Scoped tools	An agent may call only the tools explicitly granted to it. The default is none. Each tool is reviewed as part of a supply chain and its responses are validated rather than trusted.
Prompt-injection defence	Content safety and prompt-shield controls on the inputs an agent reads, including grounding data and tool responses. Treat all retrieved text as untrusted, and ensure read content cannot widen the agent's tool scope or trigger an unapproved action.
Data loss prevention	Microsoft Purview policies classify and prevent sensitive data leaving through an agent, with data minimisation so an agent collects and retains only what its task requires.
Egress control	An allow-list of destinations the agent may reach, with internal ranges and the instance metadata endpoint blocked, designed so an allowed destination cannot double as an exfiltration channel. Monitor what actually leaves.
Human in the loop	Approval gates on any write or high-impact action. An agent proposes; a person approves, for anything beyond read-and-report.
Rate and spend limits	Hard loop limits and spend caps with circuit breakers. A prompt-injected or runaway loop has a ceiling in both requests and cost. Denial of wallet is treated as a real threat.

The honest position: the platform guardrails are good and better than most teams would build, but they are necessary, not sufficient. An agent with a broad identity and open tools is a problem no content filter fully solves, because the dangerous action is a legitimate one performed for the wrong reason. The architecture, least privilege and segmentation and approval, is what turns a capable agent into a contained one.

6. Data protection and the training boundary

The baseline requires, and the platform supports, that customer prompts, completions, embeddings and fine-tuning data are not used to train foundation models without explicit instruction, that models run isolated within the tenant boundary with no access to or from the model provider, and that data stays in the chosen region and is not shared with third parties.

> **DEFAULT TO CHANGE**
> Standard deployments retain prompts and completions for a limited period for abuse monitoring, held separately per resource. Where that retention is unacceptable for the data in scope, apply for modified abuse monitoring (zero data retention) and confirm the setting is off. Layer on customer-managed keys for encryption at rest and private endpoints for the data path. Do not assume zero retention is the default; it is not.

7. Observability and detection

An agent that cannot be seen and cannot be stopped is the real worst case. This section defines how agent telemetry reaches the SIEM. The landscape here is moving quickly, so the pattern gives both the current native route and the interim route, and makes the choice between them an explicit decision.

7.1 The connector position, and the decision it forces

Native connectors that stream Agent 365 and agent-identity telemetry into Microsoft Sentinel now exist, in **public preview**. That changes the earlier position, under which there was no out-of-the-box path and a custom route was mandatory. The decision is therefore no longer "how do we build a route" but "do we permit preview connectors in production".

> **DECISION**
>
> **If the organisation permits preview features in production:** use the native connectors below. They are the simplest and richest route, and they make the Sentinel data lake the system of record for agent security.
>
> **If the organisation holds to non-preview only:** use the interim route in section 7.3 until the connectors reach general availability, then migrate. Record the preview exception, or the interim route, as an architecture decision with a review date tied to the connectors' GA.

7.2 Native route (public preview)

Source	Route into Sentinel	Table	Status
Agent activity and behaviour	Agent 365 connector, from the Sentinel Content Hub, streaming to the Sentinel data lake	`UnifiedAgentObservability`	Preview
Agent and blueprint inventory, ownership, governance	Agent Identities connector, with Microsoft Agent Identities enabled	`EntraAgentIdentities`, `EntraAgentIdentityBlueprints`, `EntraAgentIdentityBlueprintPrincipals`, `EntraAgentUsers`	Preview
Copilot agent activity	Microsoft Copilot connector	`CopilotActivity`	Preview
Agent risk signals	Entra ID diagnostic settings to the Sentinel workspace	`RiskyAgents`, `AgentRiskEvents`	GA path
Foundry agent and model telemetry	Foundry diagnostic settings, and Foundry Observability or the OpenTelemetry exporter into the same workspace	Foundry diagnostic and trace tables	GA path
Threat detection on agents	Microsoft Defender XDR, through its Agent 365 integration	Defender tables	Preview

7.3 Interim route (where preview is not permitted)

Until the connectors are generally available, collect the mandatory minimum through supported, non-preview mechanisms:

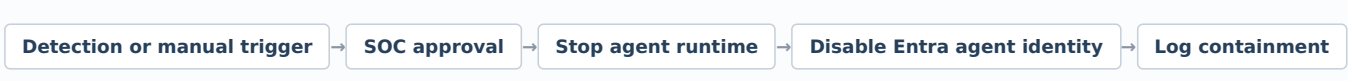
- **Foundry side:** enable Foundry diagnostic settings into the Sentinel-enabled Log Analytics workspace, and route Foundry server-side tracing and the OpenTelemetry distribution into Application Insights in that same workspace, so prompt, model and tool-call context lands.
- **Agent 365 side:** where no supported connector may be used, pull the required telemetry (agent inventory, blueprint identity, run context, model and tool-call context) through a codeless REST poller, or a scheduled function calling the Agent 365 API and writing to the Logs Ingestion API via a Data Collection Rule into a custom table. Collect the mandatory minimum only.
- **Risk signals:** the Entra diagnostic-settings route for risk logs is a supported path and should be used in both cases.

7.4 Correlation and detection

Whichever route is used, correlate on a consistent key set so an agent's activity, identity, model and tool calls join up: agent identity, blueprint identity, tenant, subscription, model and deployment, session, trace, tool call, and the user or workload identity on whose behalf the agent acted. Build detections for anomalous tool use, access outside the agent's remit, risky-agent and identity anomalies, spend and loop anomalies, and signs of prompt injection in agent behaviour.

8. The kill switch

Define a security-operations control that can stop a live agent, as the backstop for every other control in this pattern.



Implement it as an approved Sentinel playbook, or a documented manual procedure, that on authorisation stops the agent's runtime and disables its Entra agent identity, records the containment, and requires a separate approval to restore. The design points that matter: it acts on the agent's own unique identity, which is why the unique-identity requirement in section 4 is load-bearing; it is reversible only through an approval gate; and it is exercised and evidenced, not merely documented. An untested kill switch is an assumption, not a control.

9. Controls summary

Control area	Baseline requirement	Primary capability
Identity	Unique Entra Agent ID per instance, registered, least privilege, separate from human roles	Microsoft Entra Agent ID
Access conditions	Conditional Access applied to agent identities	Microsoft Entra
Tool scope	Explicit per-agent allow-list, default none, responses validated	Agent and Foundry controls
Input defence	Prompt shields and content safety on inputs and tool responses	Foundry guardrails
Data protection	Purview DLP, data minimisation, no training on data, zero retention where required, customer-managed keys	Microsoft Purview, platform terms
Egress	Allow-list that cannot carry data out, metadata and internal ranges blocked, egress monitored	Network controls
Approval	Human approval on writes and high-impact actions	Workflow and gates
Cost and loops	Rate limits, spend caps, circuit breakers	Gateway and platform
Observability	Agent telemetry into Sentinel, native connectors or interim route per the preview decision	Microsoft Sentinel
Detection	Behaviour, misuse, risk and identity detections, correlated on the agent key set	Sentinel, Defender XDR
Containment	Tested kill switch, stop runtime and disable identity on approval	Sentinel playbook or procedure

10. Residual risk and assumptions

- Several controls in the observability and detection path rely on preview features. Preview functionality can change, and its service levels differ from general availability. Where the native route is chosen, record it as a preview exception with a review date.
- Prompt injection cannot be fully eliminated, only contained. The pattern reduces its impact through least privilege, approval and egress control, and assumes it will be attempted.
- The kill switch is only as good as its last test. Treat it as a control that decays unless exercised.
- This pattern assumes the surrounding Microsoft platform, Entra, Purview, Defender and Sentinel, is itself configured to a sound baseline. It secures the agents on that platform; it does not secure the platform.
- Feature names, table names and preview status in this document reflect the position at the date shown. Confirm current names, availability and role assignments against the vendor's documentation before a production change.