The agent governance brief

What Microsoft Build 2026 shipped, what it leaves to you, and where to start

For CIOs, CISOs, chief risk officers and COOs at Australian organisations on the Microsoft stack, including those in regulated sectors.

AGILE INSIGHTS · AI STRATEGY & GOVERNANCE · OCTOBER 2026 · 12 TO 15 MINUTE READ

CONTENTS

1. What changed at Build
2. The stack in plain language
3. What Microsoft leaves to you: ten decisions
4. The Australian layer
5. Where to start: three numbers, then ninety days
6. The one-page checklist
7. How Agile Insights helps

Microsoft Build 2026 turned agent governance into a Microsoft product line. Microsoft now ships what to control: an agent inventory, agent identities, gateways for tools, containment, runtime policy and evaluation. It leaves who decides and how you prove it to you. That means the risk-tiering criteria, the approvers, the thresholds for human approval, change control for agents that rewrite themselves, telemetry retention, your Australian obligations and the value figure the whole thing is meant to earn. This brief explains both halves and gives you a practical place to start.

1. What changed at Build

At Microsoft Build 2026, held on 2 and 3 June, governance stopped being the slide at the end of the agent demo. It was the demo. In our own review of the Build catalogue, agent governance was the centre of gravity in 28 of the 35 sessions we rated as most relevant to our practice. That is our count, and we offer it as a signal: the argument consultancies were making in 2025 is now a Microsoft roadmap.

Five things changed.

Agent 365 is the control plane, and it is generally available. Microsoft said Agent 365 and its SDK reached general availability on 1 May 2026. It is positioned as the place to observe, govern and secure every agent, and it sits on Entra, Defender, Purview and Intune instead of replacing them. Agents built in Copilot Studio, Foundry and Agent Builder appear in it natively; custom agents on other frameworks onboard through the SDK.

The agent is not the user. Microsoft Entra Agent ID gives an agent its own identity. Windows is treating agents as a distinct security principal with their own sessions, and Microsoft says its telemetry can tell a person acting directly from the same person acting through an agent. That claim is worth testing in your own tenant, and it is the right direction.

Containment moved out of the prompt and into the platform. The line repeated across sessions was that a prompt cannot contain an agent; the platform has to. Microsoft showed a spectrum from process isolation through micro-VMs to full Cloud PCs, with the sandbox composed from what an action declares it needs.

Tools got enforcement points. Foundry Toolbox and the Azure API Management AI Gateway give least-privilege tool access a unit you can approve (a toolbox, an MCP product) and a place to enforce it.

Risk statements became tests. The Responsible AI breakout showed evaluations sitting in a release pipeline, and a deterministic policy layer, the Agent Control Specification, with fail-closed checkpoints at input, model, tool call and output. In Microsoft's demo, hardening the prompt barely moved multi-turn violations while the deterministic controls cut them sharply. Those are Microsoft's demo figures, so treat them as a claim to test on your own agents. A companion evaluator, ASSERT, was announced at Build.

What did not change is the part that matters most to a board. Picture a well-labelled panel of switches, every one in its default position, and a committee minute recording that the panel was installed. That is what a default template on every agent looks like: a decision, made by nobody. Microsoft now ships the switches. Deciding what each one should do, and showing later that it is set the way you said, is still your work. The rest of this brief is about that work.

2. The stack in plain language

Release status below is as at October 2026. Where we say "per Microsoft Learn" the status was re-checked on 16 October 2026; otherwise it is as Microsoft presented it at Build in June. "Preview" means the capability exists in some form and is not yet generally available; plan around it, do not depend on it.

Inventory and lifecycle: Microsoft Agent 365

What it gives you. One registry of Microsoft agents and of custom agents onboarded through the SDK, showing owner, publisher, permissions, users and a map of which agents depend on which. Blueprints define an agent's tools, data and guardrails, and the agent's identity is only created when an administrator activates it, which is a natural approval point. Templates bundle Entra access packages, conditional access, Defender and Purview data loss prevention into a baseline applied at publish. Lifecycle rules automate hygiene such as reassigning a leaver's agents.

What it doesn't. Agents on Amazon Bedrock and Google Vertex AI arrive by registry sync, which Microsoft described as visibility only. The view is single-tenant. The registry records an owner; it does not decide who the owner should be, and it holds no purpose, risk tier, impact assessment or review date. Microsoft said its risk automation is still maturing.

Status. Generally available (Microsoft said 1 May 2026; GA per Microsoft Learn).

Identity: Microsoft Entra Agent ID

What it gives you. Three identity patterns: service-style access, on-behalf-of access with the user's consent, and a standalone agent user account. Every Foundry agent gets an identity automatically. Sponsorship lifecycle workflows, so an identity has a person who answers for it, are generally available per Microsoft Learn (May 2026), and conditional access for agent user accounts was added in June 2026.

What it doesn't. In one Build demo, an agent using OAuth pass-through made commits that appeared as the user's own. Delegated identity is fine where the telemetry tags the agent; it is a problem where agent actions are indistinguishable from the person's. Licensing was not covered at Build. Microsoft Learn's licensing page, read in October, says Agent ID is available to all Entra customers while extending Entra security features to agents needs a Microsoft Agent 365 licence per user. Re-read the live page before you plan around it.

Status. Generally available (April 2026, per Microsoft Learn).

Tool access: Foundry Toolbox and the API Management AI Gateway

What it gives you. Toolbox bundles an agent's tools behind one MCP-compatible endpoint, holds credentials once, applies a Foundry guardrail before and after every tool call, and versions the toolbox. Azure API Management is repositioned as a gateway for models, MCP tools and agent-to-agent traffic, with token metrics for cost attribution. Between them, least privilege for tools has a unit to approve and a place to enforce it.

What it doesn't. Toolbox's tool search lets the model choose tools per prompt and auto-pin the ones it uses often, so the tools an agent actually uses become a runtime choice. The allow-list moves to the toolbox boundary, and tool descriptions, which the model reads to choose, become a control surface worth reviewing. A governance dashboard, outbound MCP traffic controls and per-tool restriction inside an MCP server were roadmap at Build. The two layers overlap, and Microsoft did not say which should enforce what.

Status. Toolbox is documented on Microsoft Learn; tool search and skills are in preview. API Management's MCP products, agent-to-agent support and LLM policies were described as generally available at Build and have not been re-checked since.

Data protection: Purview for agents

What it gives you. Inline data loss prevention for prompts in Foundry, a policy middleware for custom agents built on the Agent Framework SDK, Insider Risk alerts on agent-driven activity, and Foundry evaluation results synced into Compliance Manager against each agent. Data Security for Local Agents, covering coding agents on endpoints, was announced.

What it doesn't. The inline control covers prompts, not responses. Outside Microsoft 365 surfaces, developers have to call the Purview APIs themselves. A community demonstration at Build showed Purview's AI activity view holding about 30 days of agent interactions unless exported; check the current behaviour in your tenant.

Status. Mixed. Release status for most of this was not stated at Build and has not been re-checked.

Containment: execution containers and Windows 365 for Agents

What it gives you. Microsoft Execution Containers let a developer declare what an action needs (which folders, which network destinations, whether it touches the screen) and have the operating system compose a proportionate sandbox, from process isolation through Hyperlight micro-VMs and full virtual machines. GitHub Copilot's CLI and SDK ship a sandbox behind an experimental flag, and Azure Container Apps Sandboxes do the same for agent code in the cloud. At the far end of the spectrum, Windows 365 for

Agents gives computer-use agents managed, stateless Cloud PCs so they can work applications that have no API. Each agent instance gets an Entra identity, conditional access and device compliance apply through Intune, an administrator has to assign the agent to a pool before it can work, and a person can take over at any point.

What it doesn't. The execution container policy schema shown was at an alpha version, central management through Intune was deferred, and the Agent 365 integration was announced only. For computer-use agents, action-level audit inside the applications, screenshot retention and prompt injection from on-screen content were not addressed. An agent at a desktop can do whatever a clerk could; tool allow-lists do not bound it.

Status. Execution containers: preview and alpha, and not on Microsoft Learn as at 16 October 2026. Copilot CLI sandbox: experimental flag. Windows 365 for Agents: generally available (week of 1 June 2026 per Microsoft Learn), with the Intune security baseline GA in August 2026.

Runtime policy: the Agent Control Specification

What it gives you. Deterministic, fail-closed checkpoints at input, model, tool call and output, inside Microsoft's open-source Agent Governance Toolkit. Microsoft describes a tamper-evident audit chain and mappings to the EU AI Act, the NIST AI RMF and SOC 2. It is the clearest evidence yet that a prompt instruction is a request and a policy is a control.

What it doesn't. It is open source without a support SLA, and the toolkit was mid-way through a major version migration when we checked. Nothing in it decides what the policy should say.

Status. Public preview, per the toolkit's own repository (latest release 9 June 2026).

Testing and evaluation: Foundry evaluations and tracing

What it gives you. Rubric evaluators generated from an agent's own instructions, multi-turn evaluation, simulated users where no test data exists, and failing production traces turned into regression tests, with the same evaluators running before release and on live traffic. Tracing through OpenTelemetry stitches agents on LangChain, LangGraph, Google ADK, the OpenAI SDK and others into one trace. ASSERT, which turns written risk statements into test sets, was announced at Build.

What it doesn't. Nobody at Build said who sets the pass thresholds, whether an evaluation can block publishing instead of advising, or how the judging models are validated. Traces carry prompts, retrieved content and personal information, so retention and classification become design decisions.

Status. Foundry evaluations: public preview. ASSERT: announced at Build.

Value: ROI for agents

What it gives you. Foundry pairs an evaluator's definition of a successful outcome with token costs it collects automatically, tool costs you enter and a dollar value per outcome, then charts net value by agent version.

What it doesn't. Microsoft was explicit that the value per outcome is the customer's own estimate. The platform's other metrics are self-reported. Nothing at Build addressed benefits realisation or a board-level value narrative.

Status. Private preview at Build; not re-checked.

The stack on one page

Layer	Microsoft capability	Status as at October 2026
Inventory and lifecycle	Microsoft Agent 365 and SDK	GA (Microsoft said 1 May 2026; GA per Learn)
Identity	Microsoft Entra Agent ID	GA (April 2026 per Learn)
Tool access	Foundry Toolbox	Documented on Learn; tool search and skills in preview
Tool access	Azure API Management AI Gateway	GA for MCP products and LLM policies, as stated at Build
Data protection	Purview for agents	Mixed; mostly unstated at Build
Containment	Microsoft Execution Containers, Copilot CLI sandbox	Preview and alpha; not yet on Learn
Containment	Windows 365 for Agents	GA (June 2026 per Learn)
Runtime policy	Agent Control Specification (Agent Governance Toolkit)	Public preview
Testing and evaluation	Foundry evaluations and tracing	Public preview
Testing and evaluation	ASSERT	Announced at Build
Value	ROI for agents (Foundry)	Private preview at Build

3. What Microsoft leaves to you: ten decisions

Microsoft supplies the hooks. None of the following comes in the box, and each one is a decision your organisation has to make and write down.

1. **Risk-tiering criteria.** Templates and blueprints exist. Which agent gets which template does not. You need criteria that put an agent in a tier from its reach (what it can read and change), its autonomy (whether it acts or only answers) and who it affects (customers, staff, the public). Two bands are enough to start.
2. **Approval accountabilities.** Who signs an agent off before activation, and who owns it afterwards. Agent 365 records an owner without choosing one. The administrator who clicks "activate" is carrying out a decision; decide whose decision it is, per tier.
3. **Proof the controls are applied.** No session at Build showed a way to check, later, that the right template went on the right agent, that an evaluation actually blocked a release, or that a guardrail is still attached after a toolbox was updated. A control you cannot evidence is a belief.
4. **Thresholds for human approval.** Build had plenty of consent prompts, review modes and takeover controls. No session offered a threshold. Set yours per tier by irreversibility, money and the number of people affected, and keep approvals rare enough that people still read them. GitHub told Build that

users under prompt fatigue start approving everything, which removes the gate while leaving the dialog box in place.

5. **Change control for self-modifying agents.** Agent Optimizer generates and deploys new agent versions, learned skills persist across sessions and channels, and versioned toolboxes change behaviour without redeploying code. Nothing at Build put these under change approval. Define skills, tool descriptions, optimiser-generated versions and memory scope as change-controlled objects with an approval step before production.
6. **Telemetry retention, classification and residency.** Traces capture prompts and retrieved content. Agent screenshots on Windows 365 have no stated retention. Decide how long each store is kept, how it is classified, what is redacted at export and where it lives, sized to your complaint, investigation and regulator windows.
7. **Australian obligations.** No session addressed APRA's prudential standards, the Privacy Act or Australian residency for models and telemetry. Section 4 covers what to add.
8. **Fairness and decisions about people.** Build showed agents routing loan cases and approving purchases. No session covered bias testing, a register of automated decisions or a way for an affected person to challenge one.
9. **Value.** ROI for agents needs a value-per-outcome figure that Microsoft will not supply. Left alone, it gets typed in by whoever configures the dashboard that afternoon. Give the figure an owner, the evidence behind it and a review date, and reconcile it to what finance reports.
10. **Licensing and cost.** Neither Agent 365 session at Build covered it. Full control of a fleet touches Agent 365, Entra, Purview, Defender and, for the more advanced Foundry patterns, API Management. Decide which tiers need which products before procurement finds out for you.

4. The Australian layer

Build is a global developer conference, and APRA is not on the keynote circuit. No session we reviewed covered CPS 230, CPS 234, the Privacy Act or Australian residency. Microsoft's general advice on regulation was to instrument everything now. That is sensible, and it is the start of the work, not the end of it. What an Australian organisation adds on top of the Microsoft stack is a mapping from each control to the obligation it helps meet, and a handful of decisions the stack leaves open.

APRA-regulated entities. CPS 230 (operational risk management) has applied since 1 July 2025, and the transition for existing service-provider arrangements ended on 1 July 2026. The standard does not mention AI, and that neutrality is the point. An agent that underpins a critical operation, or whose failure creates material operational risk, is in scope on the standard's own terms, and the vendors behind it may be material service providers. CPS 234 (information security) treats agent identities, tool credentials, prompts and traces as information assets to be classified and protected. APRA's April 2026 letter to industry on AI set out its expectations of boards and accountable executives, including an AI inventory with accountability across the lifecycle and human oversight for higher-risk uses; it did not introduce a new AI standard. The agent inventory, tiering and approver decisions in this brief are the practical response to all three.

The Privacy Act, with one example. Amendments to the Australian Privacy Principles add a transparency duty for automated decision-making. From 10 December 2026, as the legislation currently stands, the duty applies to an APP entity that uses personal information in a substantially automated decision that could reasonably be expected to significantly affect an individual's rights or interests. The entity's privacy policy must describe the kinds of personal information used, the kinds of decisions the program makes on its own, and the kinds it substantially and directly helps to make. Decision support can count. An agent that triages claims, scores applications or routes complaints may be making such a decision, and your register needs to

say so. Our ADM privacy explainer covers the duty in plain terms. The privacy policy paragraph is an afternoon's work; finding every tool that makes a decision about a person is the real job.

Residency. Decide, per workload, where prompts, traces, screenshots and agent memory are stored and processed, and confirm the region availability of each model you use in Foundry. Nothing at Build answered this for Australia, and the answer varies by product.

Log retention. Work out the longest window that matters to you: how long a customer has to complain, how long an investigation takes to start, how far back a regulator or auditor might ask. Compare it with what each telemetry store keeps by default. If the logs expire first, the evidence of what an agent did may be gone before anyone asks for it.

Decisions about people. Keep a register of the agents and systems that make or shape decisions about individuals, test them for unfair outcomes before and after release, and give affected people a way to challenge a decision. None of this came from Build; all of it will be asked for.

The pattern that works for most organisations is one internal baseline built to the strictest expectation you inherit, mapped afterwards into each regulator's language. One governance system, many sector stories.

This section is general information, not legal advice. Check the current position with your legal advisers.

5. Where to start: three numbers, then ninety days

Before any framework, get three numbers on one page.

The number	Where it comes from	What a gap means
How many agents you run	The Agent 365 registry export, plus whatever lists platform, security and business teams keep. Mark registry-sync agents and anything the registry cannot see.	You have agents nobody is counting, which is normal and fixable.
How many have a named approver and owner	Compare the registry's owner field with your own records of who signed each agent off.	Agents are live on a decision nobody made. Those go to the top of the backlog.
How many act under their own identity	Ask your identity team which agents hold an Agent ID, which act on behalf of a user with the agent tagged, and which simply run as the person who launched them.	Attribution and segregation of duties fail the first time someone asks who acted.

Most organisations we talk to can answer none of the three today, and the first one has become cheap. If any of them takes longer than an afternoon, that is the finding.

THE FIRST NINETY DAYS

Days	Focus	What exists at the end
1 to 30	Count and claim	Every agent found and listed, registry-sync and unmanaged agents marked, an owner claimed or the agent stopped. The three numbers, dated.
31 to 60	Tier and approve	Two risk bands with written criteria; an approver and owner per band; the agents that can act, listed separately; thresholds for human approval by irreversibility, money and people affected.

Days	Focus	What exists at the end
61 to 90	Wire and prove	Each band mapped to an Agent 365 template, an identity pattern, a tool policy and a containment tier; a release gate with evaluation thresholds someone owns; retention set for each telemetry store; a first sample check that what the register says matches what is configured.

Keep the low band fast. Governance that slows every agent to the speed of the riskiest one gets routed around.

6. The one-page checklist

INVENTORY

- [ ] Every agent in one register, seeded from Agent 365 and reconciled with other lists.
- [ ] Registry-sync and unmanaged agents marked as such.
- [ ] Each entry carries purpose, owner, approver, risk tier and review date.

ACCOUNTABILITY

- [ ] Written tiering criteria, with two bands to start.
- [ ] An approver and an owner named per tier, and the activation step tied to them.
- [ ] Thresholds for human approval set per tier, by irreversibility, money and people affected.

IDENTITY

- [ ] Every agent acts under its own identity or a tagged delegated identity; pass-through rejected where actions are indistinguishable from the user's.
- [ ] A sponsor for each agent identity, included in access reviews.
- [ ] Attribution tested: from the logs alone, can your security team say whether a person or an agent acted?

TOOLS AND DATA

- [ ] Allow-lists set at the toolbox or gateway boundary; tool descriptions reviewed as a control.
- [ ] A decision on which layer enforces which policy where Toolbox and API Management overlap.
- [ ] Data loss prevention and sensitivity labels in place where agents read and write.

CONTAINMENT

- [ ] A containment tier per risk band, from process isolation to Cloud PC.
- [ ] A policy for coding agents and assistants on laptops: permitted harnesses, identity, sandbox, consent.
- [ ] Computer-use agents treated as their own intake trigger, with human takeover, action evidence and screenshot retention specified before build.

RELEASE AND CHANGE

- [ ] Evaluation requirements and pass thresholds set per tier, with a named approver.
- [ ] Skills, tool descriptions, optimiser-generated versions and memory scope defined as change-controlled objects.
- [ ] Deterministic checkpoints, not prompt instructions, carrying the rules that must hold.

EVIDENCE

- [ ] Retention, classification and redaction set for traces, Purview AI activity and agent screenshots, sized to complaint and regulator windows.
- [ ] A quarterly sample check that configured controls match the register, with mismatches owned.
- [ ] One agent decision replayed end to end as a test of the audit trail.

AUSTRALIA

- [ ] Each control mapped to the CPS 230, CPS 234 and Privacy Act obligations it supports, where they apply to you.
- [ ] A register of agents that make or shape decisions about people, with fairness testing and a challenge path.
- [ ] Residency confirmed per workload for models, prompts, traces and screenshots.

VALUE

- [ ] A value-per-outcome figure per agent with an owner, evidence and a review date.
- [ ] Platform metrics reconciled to what finance reports.
- [ ] Licensing for the governance stack costed per tier.

7. How Agile Insights helps

Agile Insights is a Microsoft Data and AI partner, and our AI Strategy and Governance practice does the half of this work that Microsoft leaves to you. Our Agent Governance offer governs agents as a fleet, on the Microsoft stack you already license, and it runs inside Agent 365 rather than beside it.

It starts with a **fleet sprint of two to three weeks**:

- every agent found and listed, seeded from the Agent 365 registry and reconciled with what the registry cannot see
- risk tiers written, including tiers for computer-use and autonomous agents
- an owner and an approver named for each agent
- an identity pattern chosen per tier.

You leave the sprint with the three numbers, dated, and a backlog in priority order.

Rollout then follows by tier:

- tool and data policy at the gateway, and containment keyed to risk
- release gates with evaluation thresholds someone owns, and change control for self-modifying agents
- telemetry retention sized to your obligations
- monitoring and evidence feeding our AI Governance Control Plane, where a control is a live object with an owner, evidence and a verification, instead of a paragraph in a policy
- for regulated clients, the Australian mapping from section 4.

Where to start with us. Take the **free AI maturity check**. It takes minutes, it shows where your organisation stands before anyone opens a console, and it tells you which of the ten decisions in section 3 need attention first. If automated decisions about people are the item on your desk, our **ADM privacy explainer** is the short version of the Privacy Act duty. Then, if you want help reading what you find, talk to us.

- Free AI maturity check and the practice page: https://www.agile-insights.com.au/ai-strategy-advisory-governance/
- Contact us: https://www.agile-insights.com.au/contact-us/

Sources and currency. Microsoft Build 2026 sessions (2 and 3 June 2026) and Microsoft Learn, as at October 2026; key release statuses re-checked on Microsoft Learn on 16 October 2026. Microsoft's content is paraphrased. Vendor figures and forecasts are Microsoft's claims and are reported as such. The "28 of 35 sessions" figure is our own rating of the Build catalogue. Australian regulatory references are to APRA CPS 230 and CPS 234, APRA's April 2026 letter to industry on AI, and the Privacy and Other Legislation Amendment Act 2024.

Not legal advice; release status as at October 2026. Most of what Build showed was preview in June, and several sessions disagreed on status; re-check the live Microsoft Learn page before relying on any of it.

Agile Insights · www.agile-insights.com.au · AI Strategy & Governance practice
Not legal advice; release status as at October 2026.