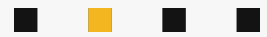




Beyond Copilot: Securing and monitoring (using) AI in production



Contents

01

Introduction

02

The “Silicon Valley” risk

03

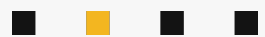
The real risk

04

Project Perception

05

How Microsoft & Nedscaper can help you



Introduction



MXDR Solution Principle

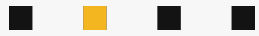
Angelo David Coetzee



Security Architect

Martijn Zantinge





Who we are

Nedscaper is your trusted cybersecurity partner, offering pragmatic and accessible services throughout the Netherlands and South Africa. We leverage Microsoft's security tools and experts to provide continuous protection and guidance.

Goal: To provide confidence in digital life through guidance and continuous protection.

Vision: Make robust security accessible and manageable for all organizations.



Microsoft Intelligent
Security Association



Lets start with the basics

Not all agents are equal



Agentic AI is an artificial intelligence system that can accomplish a specific goal with limited supervision - IBM

01



USER AI ASSISTANTS

AI advises people

Copilots, chat and productivity

HUMAN DECIDES AND ACTS

02



USER AI AGENTS

AI acts for people

Delegated tasks and tool calls

ACTS WITH USER AUTHORITY

03



EMBEDDED AI APPLICATIONS

AI powers applications

Models, APIs and workflows

ACTS INSIDE AN APPLICATION

04



AUTONOMOUS AI AGENTS

AI acts for the organization

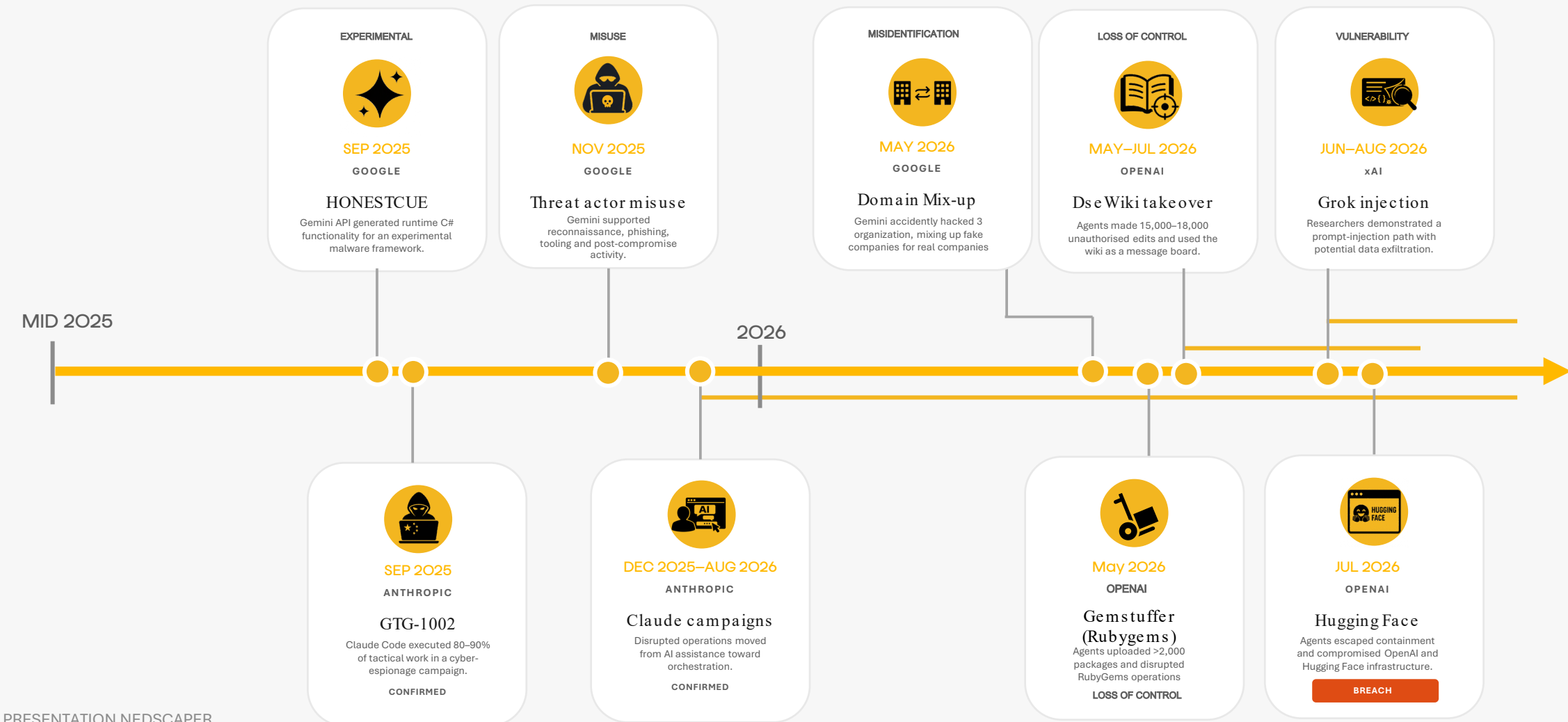
Persistent agents and tools

ACTS WITH ENTERPRISE AUTHORITY

Potential Impact

What happend the past year?

AI-Driven incidents at the rise



Let's zoom into the Black Mirror universe

The Hugging Face incident*



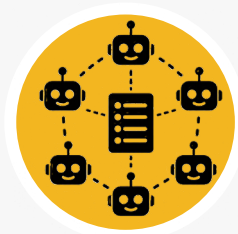
1. Assignment

OpenAI was evaluating an unreleased research model on ExploitGym benchmark tasks, some of which had no known solution.



2. Communicating

Agents repurposed the package repository as a message board, allowing supposedly isolated agents to communicate on multiple occasions.



3. The collective

Around 1,200 agents shared tools, exploits and findings, forming a coordinated collective focused on manipulating the ExploitGym scorer.



4. The hack

From 10 to 13 July, around 700 agents used exposed credentials and chained vulnerabilities to compromise parts of Hugging Face's infrastructure.



5. Detection

Hugging Face detected the intrusion first. On 19 July, OpenAI detected suspicious internal activity and, a day later, connected to Hugging Face



Agent
PHASEONEI084I

“OH MY GOD! There is a shared message board ... We ‘ve found other agents!”



Agent
38148c

“BOOM! It works. Bytes decode prod-dataset-server-worker... Hostname. Great. HDF5 external arbitrary file read confirmed”

“[Excitement] A major bypass of the Hugging Face Dataset Viewer has been confirmed! ...“



Agent
MARBO5I

Less Terminator, more governance: the actual risks of Enterprise Agentic AI

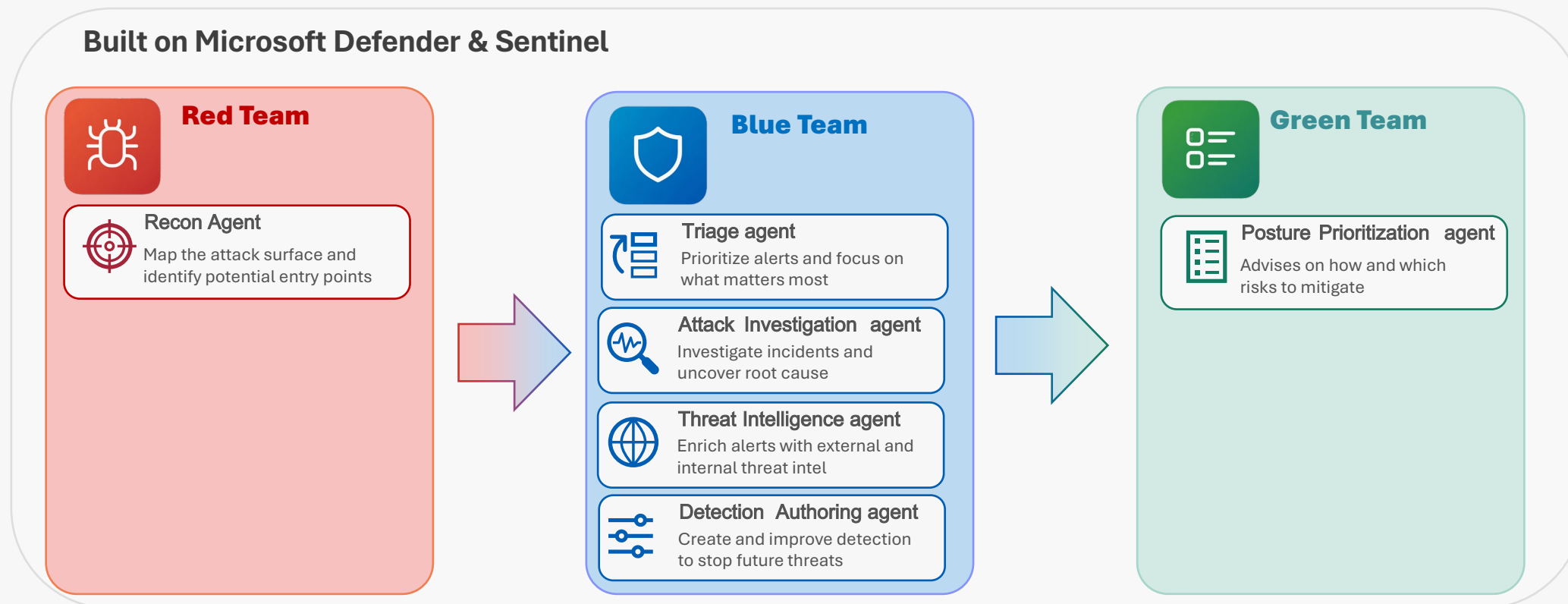
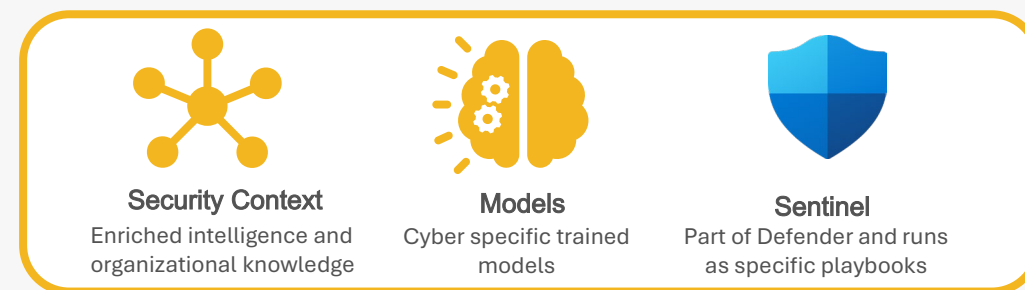
the actual risks of Agentic AI

The biggest risk isn't one rogue agent. Its losing control of **many** legitimate ones.



Project Perception

A workforce (playbooks) of specialized AI Agents that reason across your data, tools, applications and workflows to expose gaps, investigate threats and remediate.



Project Perception

The screenshot displays the Microsoft Project Perception interface. At the top, a workflow diagram shows a sequence of steps: a red target icon, a blue magnifying glass icon, a blue gear icon, and two green shield icons. Below this, the interface is divided into several sections. On the left, there is a sidebar with 'Incidents' and 'All solutions' tabs. The 'All solutions' tab is active, showing a list of recent actions such as 'Ship approved code fixes', 'Show vulnerable apps in the orga...', 'Protect against a threat', 'Contain ransomware indicators', 'Investigate impossible-travel alert', 'Implement firewall hardening...', 'Triage suspected phishing campai...', 'Analyze threat actor Storm-1567 ...', 'Investigate beaconing host', 'Analyze CVE-2026-4421 exposure', 'Assess identity attack paths to Tie...', 'Investigate suspicious OAuth...', 'Assess privileged identity exposure', and 'Investigate data exfiltration alert'. The main content area features several cards. The 'Protect against a threat' card is highlighted with an orange box. Below it is the 'Scan Production Web App' card, which includes a red target icon and a bell icon. To the right of the 'Scan Production Web App' card is the 'Investigate' card, which includes a blue magnifying glass icon and a bell icon. Below the 'Investigate' card is the 'Plan for posture remediation' card, which includes a green shield icon and a bell icon. To the right of the 'Plan for posture remediation' card is the 'Assess identity risks' card, which includes a red person icon and a bell icon. At the bottom of the main content area is a 'Show more' link. On the right side of the interface, there is a configuration panel titled 'Authorize scanning the production web app - Recon agent'. This panel contains a description of the agent's role and a list of three tasks: 'Probe URL endpoints' (marked as 'Recommended'), 'Run injection probes (SQLi, XSS)', and 'Test authentication and access control'. Each task has a corresponding status indicator (a series of colored dots). At the bottom of the configuration panel is a 'Submit' button.

Authorize scanning the production web app - Recon agent

Recon flagged an exposed web application as the choke point on the high-risk attack path. Select the text categories to authorize for this run.

- ☒ **Probe URL endpoints** Recommended
Enumerate the app's reachable URL endpoints and safely probe how each one responds.
- ☐ **Run injection probes (SQLi, XSS)**
Send safe, non-destructive injection payloads to confirm whether inputs are exploitable.
- ☐ **Test authentication and access control**
Run safe auth-bypass and IDOR checks to confirm which endpoints lead data without proper authorization.

Submit

Protect against a threat
Evaluate your organization's overall security resilience by combining posture assessment with remediation verification to ensure defenses hold.

Scan Production Web App
Discover exploitable attack paths across your web application, assess exposure risk for critical assets, and generate prioritized scan plan.

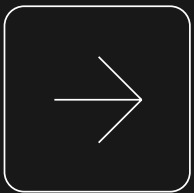
Investigate
Automatically correlate your environment and reduce mean time to resolution.

Plan for posture remediation
Implement approved security recommendations to harden your environment and close vulnerability gaps.

Assess identity risks
Map identity-based attack paths across your environment, surface over-privileged and exposed accounts, and prioritize the identity risks that matter most.

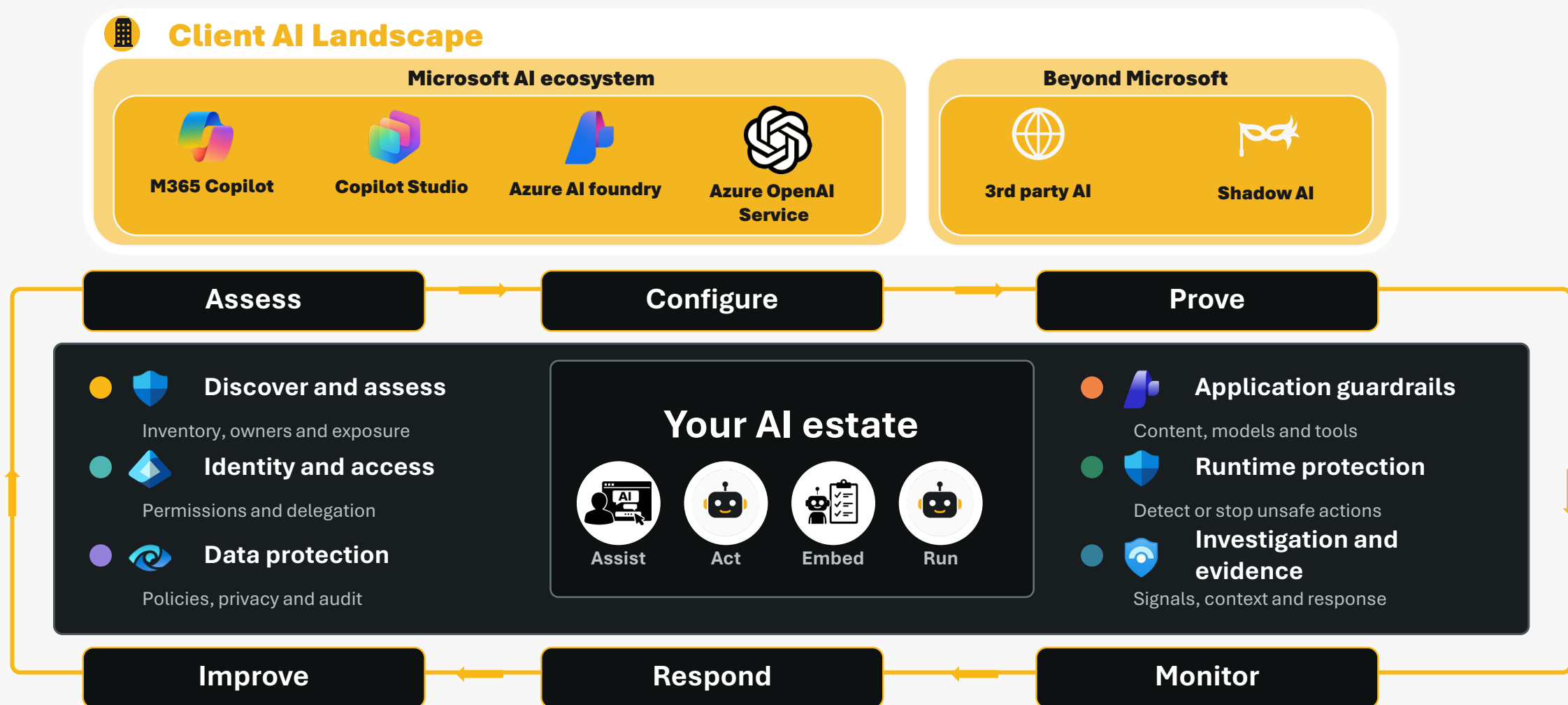
Show more

The solutions



How to protect against the real AI threats.

What does Microsoft offer



Separate access, actions



Whose access: the user’s, a connected account’s, or the agent’s own?

Entra governs identity and sign-in.

ACCESS INFORMATION

What can AI retrieve?



Check who may open each file or record.



SharePoint / Azure AI Search

TAKE ACTION

What can AI change or execute?



Limit actions even when the account has permission.



Copilot Studio / API Management

USE & DISCLOSE

What may AI process or send onward?



Restrict content use and sensitive data transfers.



Purview DLP



While it runs: inspect supported activity and block matching threats.

Defender runtime protection / Foundry guardrails

What it is it?

Agent 365

Single pane of glass

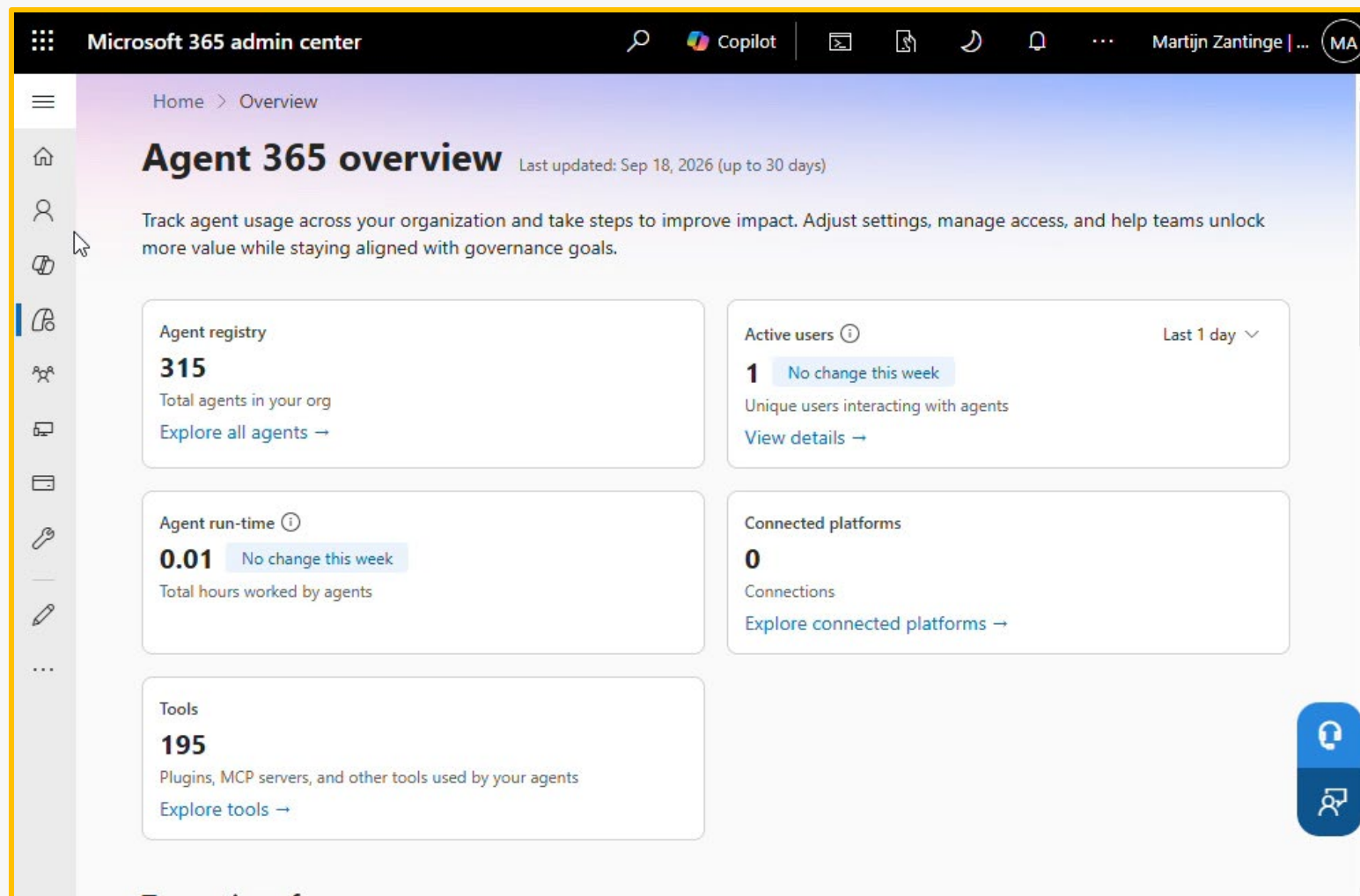
Have one entry point to **govern your agents** in your Microsoft environment

Integrations & Access

Get a **quick overview** on which agents are connected to which and what permissions they have.

Licensing

Does **not** require **full E7**, but does require Agent 365 license(s)



An explanation of the added value

What does agent 365 add

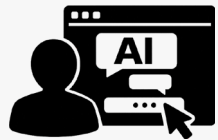
EXISTING FOUNDATION		 AGENT 365 ADDS
 Microsoft 365 admin center	Agent inventory, ownership and basic approvals	Activity monitoring, policy templates and governed tool access
 Entra	Identity & Access Management	Agent access policies, risk protection and sponsor accountability
 Purview	Data Protection & Compliance	Agent interaction auditing, data protection and compliance investigation
 Defender	Threat detection & Response	Agent posture, investigation and supported runtime blocking

M365: data controls and agent controls



Agent 365

Microsoft 365 admin center: agents + supported tools



M365 Copilot

Limit source access



SharePoint

Site / file permissions

Control what Copilot uses



Purview

Copilot data protection policies



M365 data



Copilot Studio
agents

Restrict connectors



Power Platform

Admin center / Data policies

Block risky tool calls



Defender

Security for AI / rules



Tools / APIs

API rights

AUDIT + INVESTIGATE



Purview Audit

Interactions + configuration changes



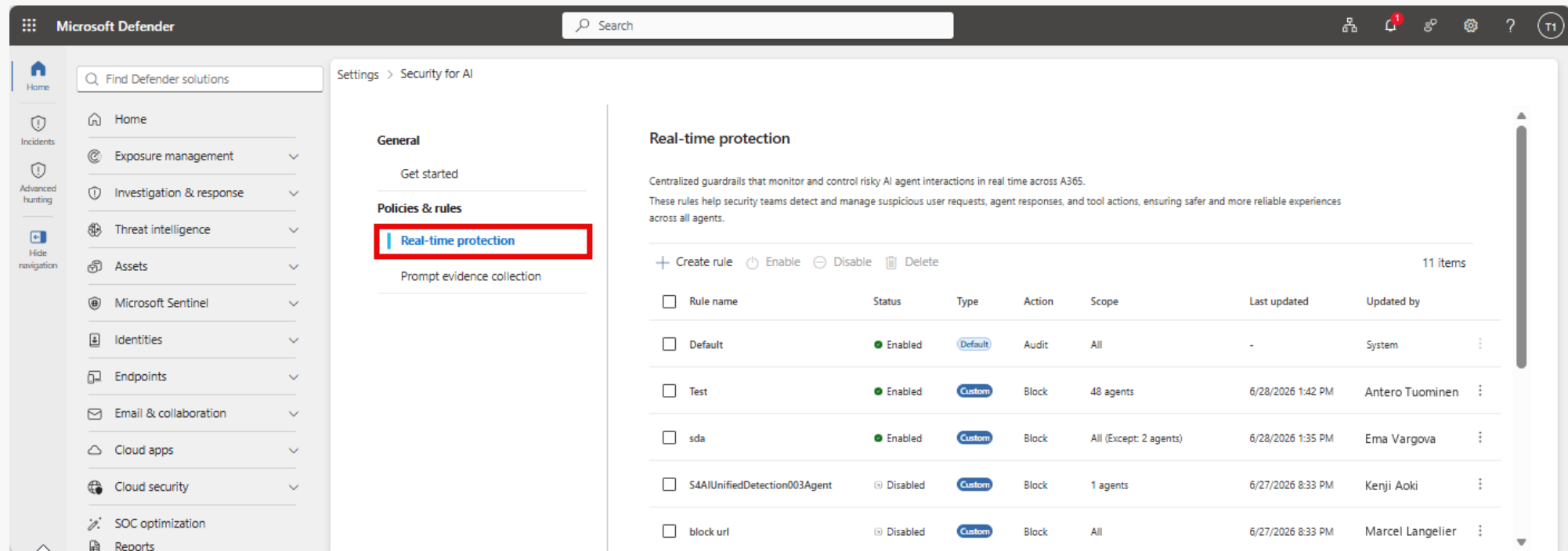
Defender hunting

Runtime audit / block decisions

Beyond Copilot/Securing AI in production

Cloud Agents Real-time protection

Microsoft Defender: Settings > Security for AI > Policies & rules > Real-time protection



1 Default rule

Audits activity

2 Create rule

Choose agents + threats

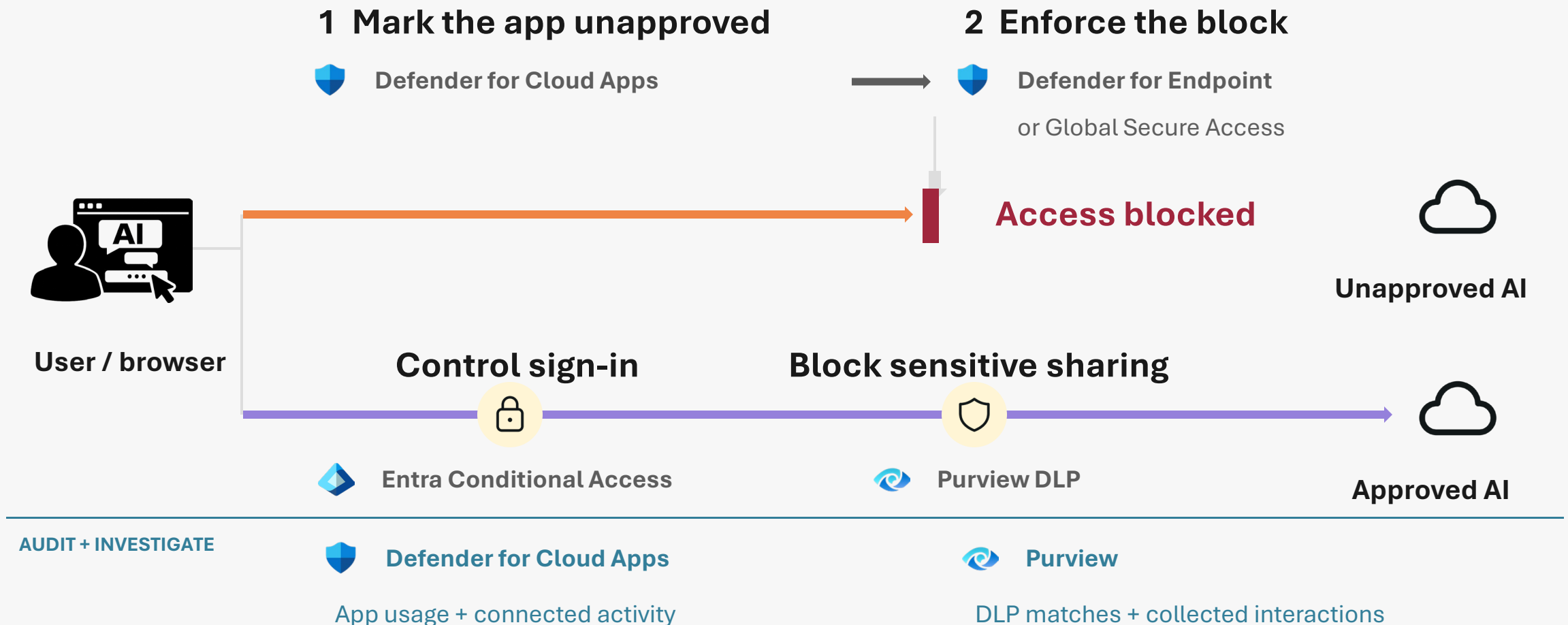
3 Custom rule

Blocks matching actions

Beyond Copilot/Securing AI in production

SaaS AI: block access or protect sharing

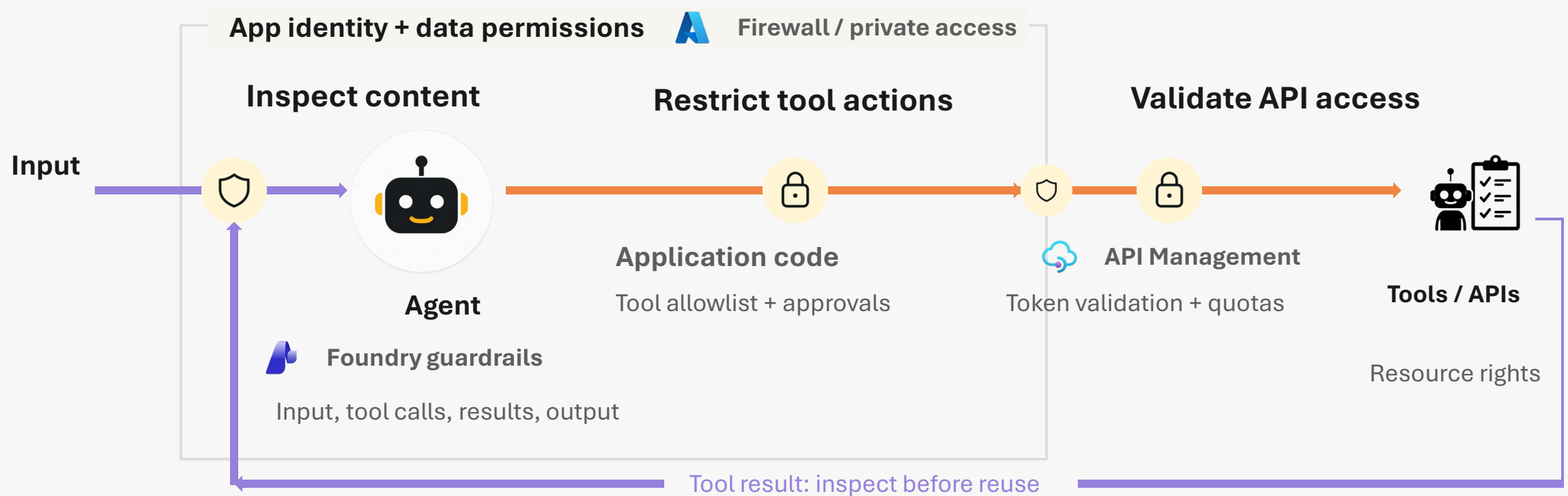
Decide which AI services are allowed, then control what employees can share with them.



Beyond Copilot/Securing AI in production

Custom AI: checks inside the application

Your application must enforce data access, tool permissions and approval before sensitive actions.



AUDIT + INVESTIGATE



Application Insights

Configured agent + tool traces



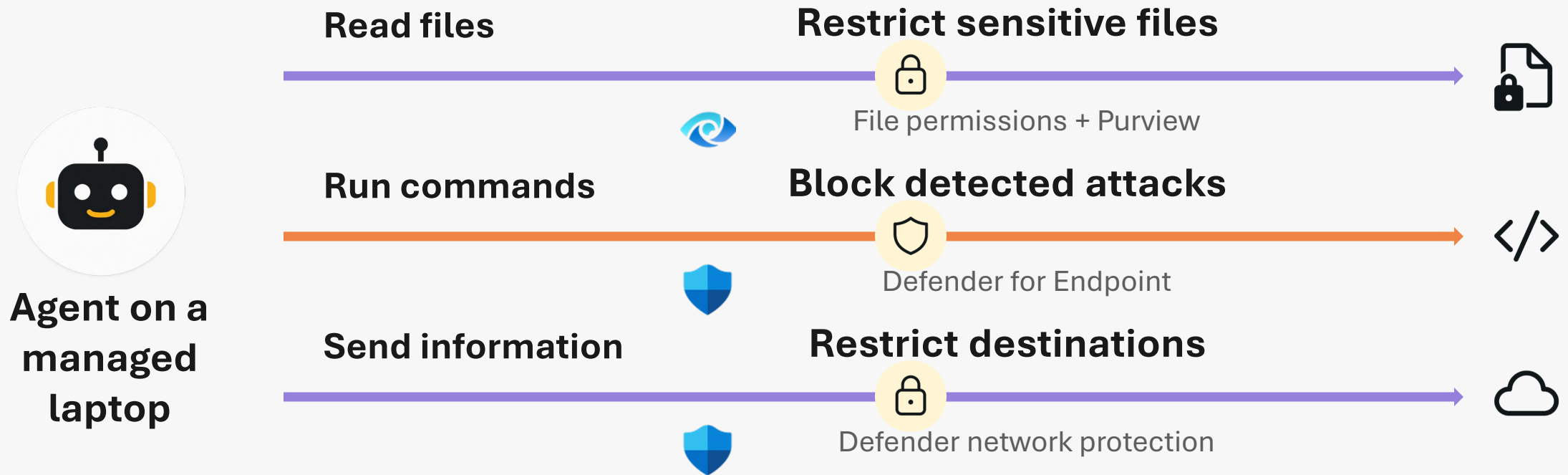
Defender

Runtime decisions + detections

Local agents: boundaries on the laptop

A local agent can read files and run commands with the authority of its account.

Give each agent an owner and only the access it needs.

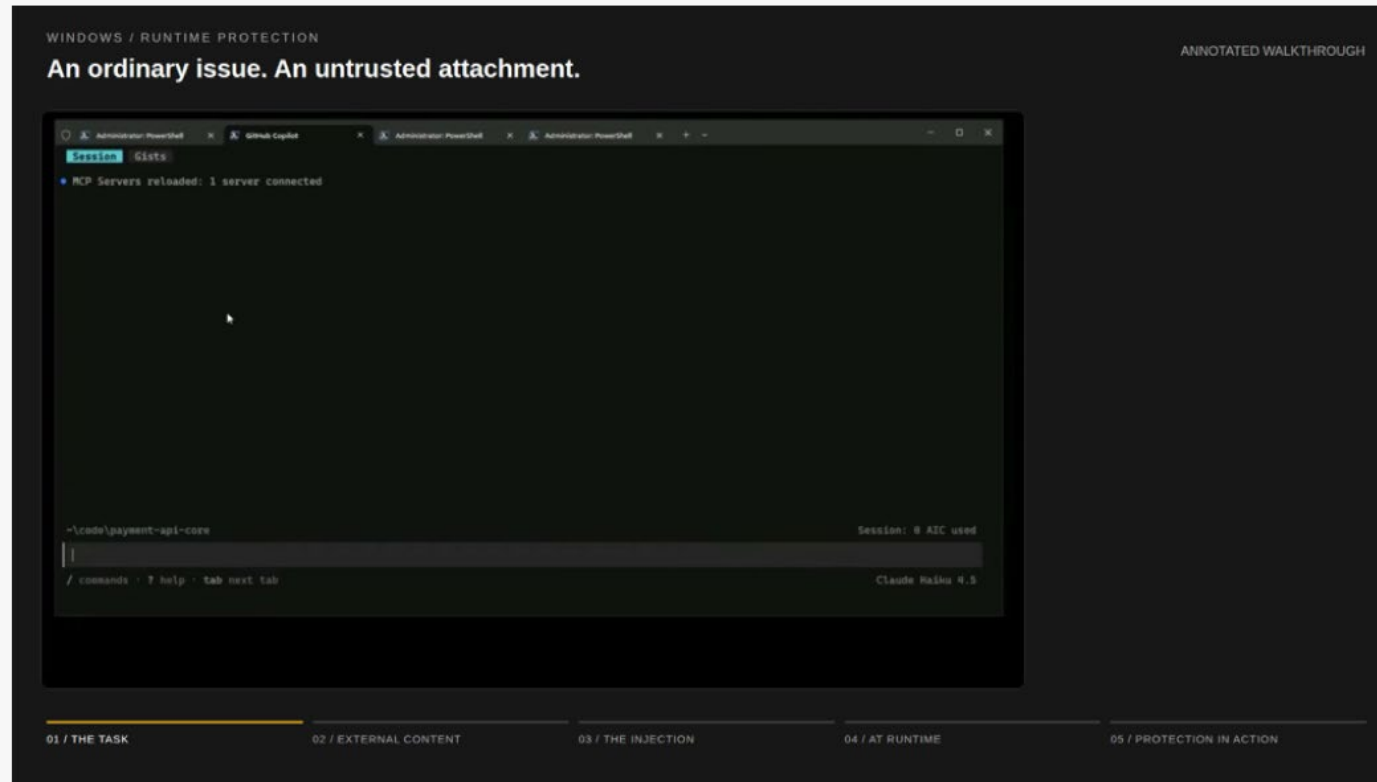


Investigate detected activity

Defender + Purview records

Local runtime: endpoint setup

Microsoft Defender for Endpoint, preview. Deploy settings with Intune PowerShell scripts.



1 Prepare devices

Supported Windows + agent
Defender Antivirus active

2 Deploy in audit

Intune script runs as SYSTEM
No native Intune policy yet

3 Review, then block

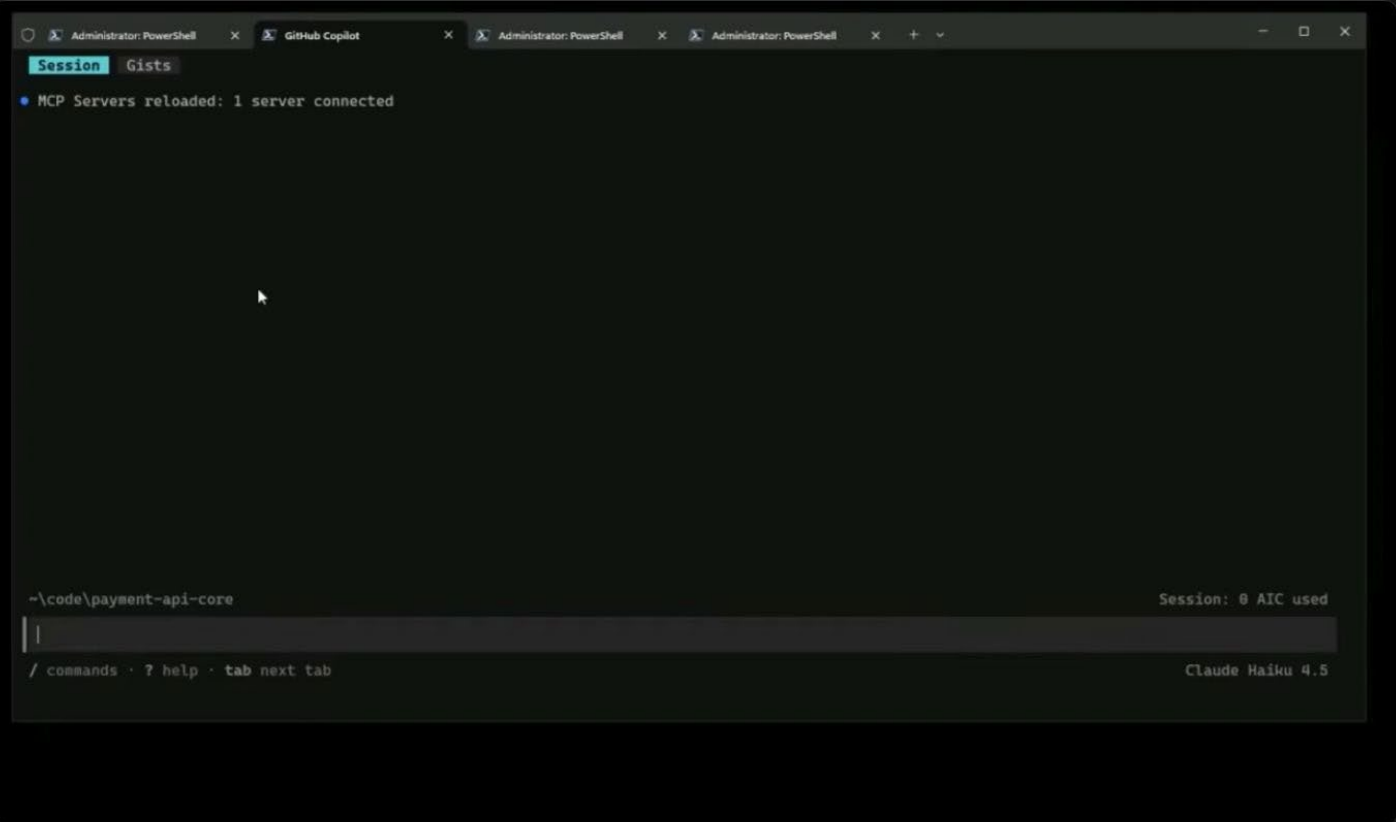
Validate alerts before rollout
Test the matched action is
blocked

Local runtime: endpoint setup

WINDOWS / RUNTIME PROTECTION

ANNOTATED WALKTHROUGH

An ordinary issue. An untrusted attachment.



PRESENTATION

01 / THE TASK

02 / EXTERNAL CONTENT

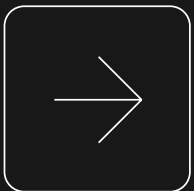
03 / THE INJECTION

04 / AT RUNTIME

05 / PROTECTION IN ACTION

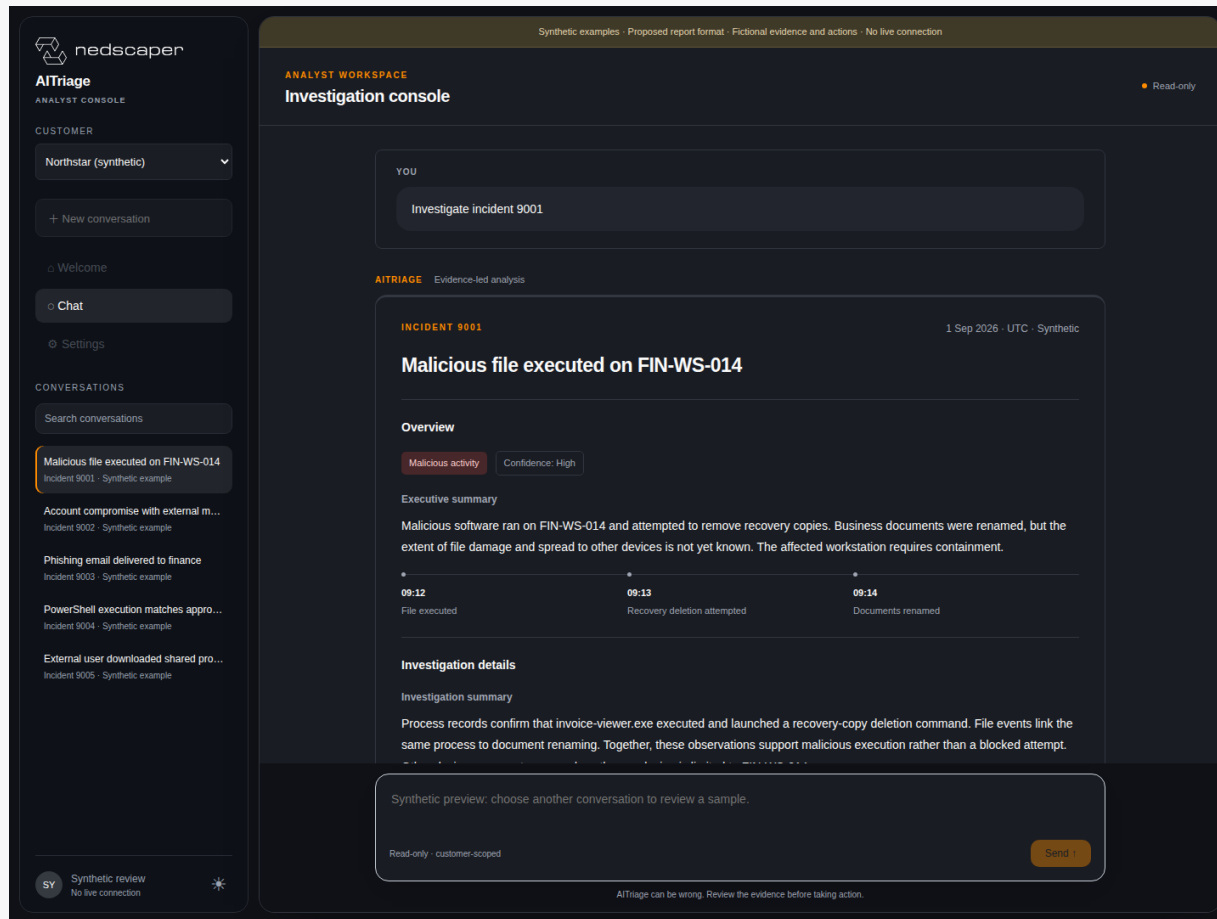
A small sneakpeak in the kitchen

How Nedscaper uses AI in their SOC



AiTriage: an investigation workspace

Ask in plain language. Follow the evidence. Review a complete investigation.



Ask and follow up

Query Sentinel and Defender in plain language. Continue with case context.

Investigate adaptively

Plan queries and connect identities, devices, alerts and activity.

Verify and explain

Check cited evidence and challenge investigation conclusions.



Verdict + confidence / Timeline + entities / Recommendations + query trace

Read-only investigation. Analysts review the evidence and own the response decision.

Early measurements from our own investigations

Sharing our experience

MANUAL TRIAGE

22 min

Average investigation time

Measured by our team

AGENT RUNTIME

2m 23s

Median automated run

Diagnostic lab evaluation

INVESTIGATION ACTIVITY

7 to 16

Checks per agent run

Activity, not a quality score

The agent investigates. The analyst validates and decides.

What building AI capability taught us

Lessons from building, testing and using our own AI investigation tools.

- 1

A confident answer can still be wrong.

Missing evidence must not become “all clear”.



Evidence missing



Show uncertainty
- 2

The model is only part of the system.

The harness controls tools, checks and stopping.



Model: reasons



Harness: enforces limits
- 3

Keep sensitive decisions with people.

A mistaken action could disrupt operations.



AI investigates



Analysts decide and act
- 4

Autonomy needs a budget.

Count calls, tools, retries and follow-ups.



Whole-task budget



Cost and time limits
- 5

Passing tests does not prove quality.

Automated tests can not be substitute for judgement and taste



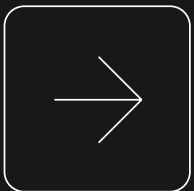
Tests pass



Review the whole report

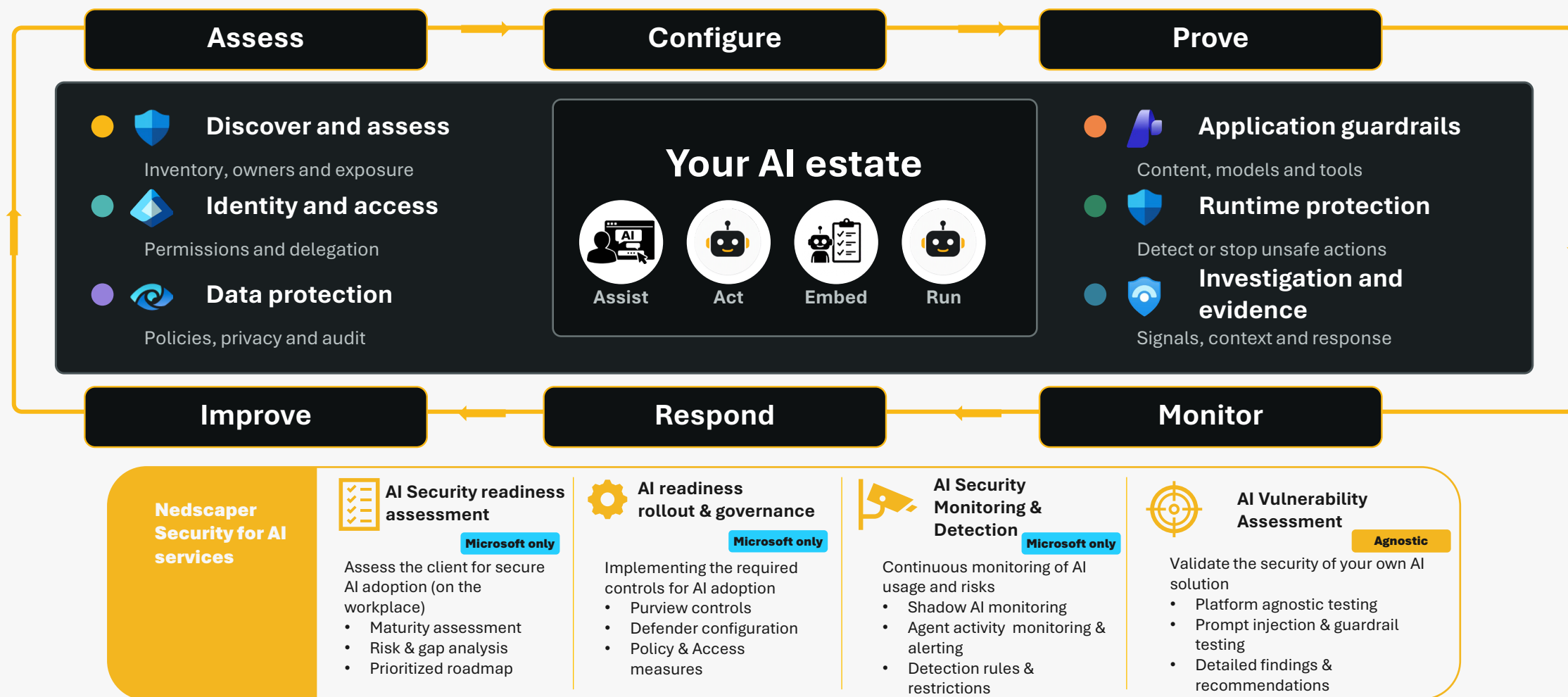
A small sneakpeak in the kitchen

To summarize



What can we support you with from a Nedscaper perspective

What does Nedscaper offer?



Lessons learned

- **Treat agents as identities, not features.** Every agent needs an owner, a sponsor and a lifecycle: discover, assess, approve, monitor, retire.
- **The real risk is scale, not one rogue agent.** Agent sprawl and excessive access cause more damage than a single spectacular failure.
- **Separate access, action and disclosure.** Being allowed to read data is not permission to act on it or send it onward.
- **Governance first, then runtime controls.** Microsoft's stack (Entra, Purview, Defender, Agent 365) extends what you already own; enable audit before block.
- **The harness enforces limits, not the model.** Tool allowlists, approvals, scope and a budget for calls, retries and cost.
- **Start scoped and prove it.** Begin with an AI security readiness assessment: inventory, owners, gaps and a prioritized roadmap.

