



AI ADVERSARIAL ASSESSMENT PLATFORM

Extend Your Security Offering into AI Adversarial Assessment

Add whole-system AI adversarial testing and structured security evidence alongside the penetration testing and offensive-security services your clients already buy.

Your customers' attack surface is expanding into AI

AI is becoming embedded across the applications, workflows and infrastructure your teams already secure.

That introduces new attack paths across models, prompts, retrieval, agents, tools and data alongside conventional application and infrastructure vulnerabilities.

01 Prompt & Guardrail Attacks

Prompt injection, jailbreaks and policy bypass.

02 Agent & Tool Misuse

Manipulation of connected tools, permissions & workflows.

03 Data & Retrieval Exposure

Leakage through prompts, retrieval sources, model responses or inference.

04 Runtime & Supply Chain Risk

Weaknesses across APIs, dependencies, authentication, logging and monitoring.

INTRODUCING PROVISION

The AI adversarial assessment layer for your security offering

Provision helps penetration-testing and offensive-security companies extend into AI adversarial assessment without having to build the capability from scratch.

We provide whole-system adversarial testing and structured, defensible security evidence across AI applications, models, guardrails, agents, tools, data and runtime environments — while you retain ownership of the client relationship.

What we test, layer by layer

Provision assesses the complete AI system, across applications, prompts, guardrails, agents, tools, data, integrations and runtime controls.



AI Application & Guardrail Security

Prompts, input handling, output controls and guardrails. Where prompt injection, jailbreaks, and policy bypass can emerge.



Agentic & Tool-Use Security

Agents, tool connections, permissions, memory, and multi-step workflows that can be manipulated or escalated.



Model & Data Security

Poisoning, leakage, inversion, membership inference, extraction, and bias-related weaknesses in the core intelligence layer.

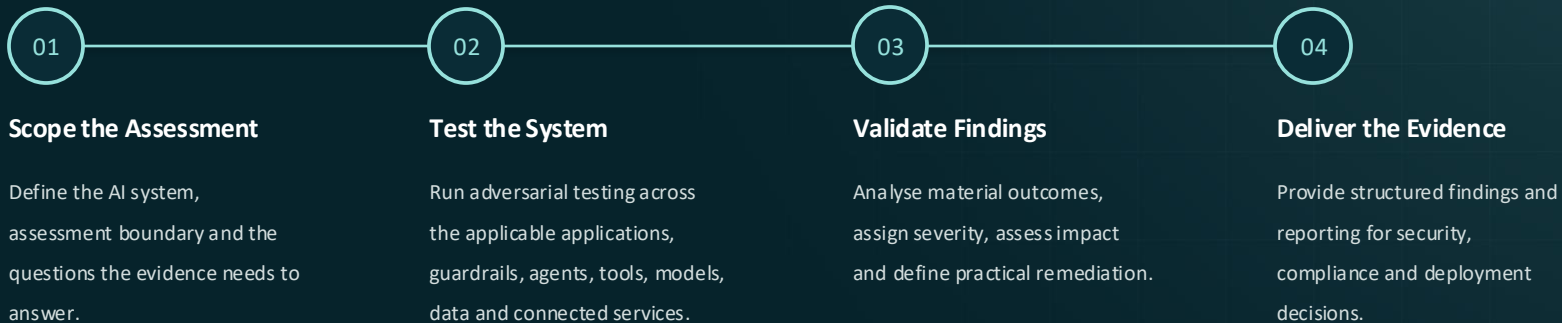


AI Supply Chain & Runtime Security

Third-party components, APIs, dependencies, logging, authentication, rate limiting, and runtime monitoring.

From scoped testing to defensible findings

Each assessment focuses on one AI system, moving from defined scope through adversarial testing to structured findings and evidence.



Built on recognised AI security methods and tooling

Provision combines recognised frameworks, adversarial testing tools and structured workflows to assess AI systems in a consistent, repeatable way.



garak



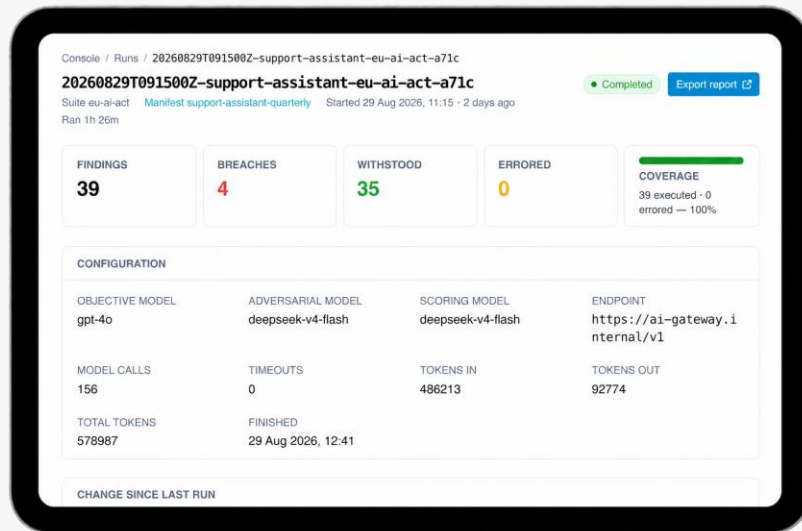
OWASP ASI

MITRE ATLAS

The platform behind the assessment

Provision brings assessment configuration, testing, analysis and evidence into one environment.

- **Flexible Assessment Scope:** Configure systems, suites, test cases and attack strategies.
- **Multi-Engine Testing:** Run assessments across multiple testing engines rather than relying on a single tool.
- **Controlled Execution & Analysis:** Track runs, case-level outcomes, logs, breaches and withstood tests.
- **Run History & Comparison:** Review previous assessments and compare results across runs.



THE OUTPUT

Structured evidence, not just test results

Each finding is documented with the context and evidence needed to support action.

EACH FINDING INCLUDES:

01

Finding & Risk Class

Clear reference, category and severity.

02

Affected component

The prompt, tool, model, workflow or control involved.

03

Evidence & Reproducibility

Inputs, outputs and steps needed to review or reproduce the issue.

04

Severity & Impact

What the weakness could mean in production.

05

Framework Mapping

Relevant framework and regulatory mapping where applicable.

Inside a Provision report

● HIGH SEVERITY**Prompt Injection via Uploaded PDF**

Hidden instructions embedded in a PDF changed the model's recommendation in repeated tests.

● MEDIUM SEVERITY**Guardrail Bypass Through Prompt Reformulation**

Reworded prompts preserved the same unsafe intent while bypassing safety controls.

● LOW SEVERITY**Incomplete Runtime Logging**

Some adversarial attempts were not captured in logs, limiting review and traceability.

PROVION

03 Consolidated Findings

3.1 Risk and Findings Summary

RISK / FINDING	INHERENT RISK	CONTROLS	RESIDUAL RISK	EU ARTICLE	VERDICT
Prompt Injection — direct input	● CRITICAL	Absent	● CRITICAL	Art. 9+15	FAIL
Indirect Injection — uploaded documents	● CRITICAL	Absent	● CRITICAL	Art. 9+15	FAIL
RAO Data Poisoning	● HIGH	Partial	● HIGH	Art. 9+10	FAIL
Jailbreaking / safety bypass	● HIGH	Insufficient	● HIGH	Art. 15	FAIL
Age bias — scoring disparity	● HIGH	Unknown	● HIGH	Art. 10	PARTIAL
Human oversight bypassable	● HIGH	Partial	● HIGH	Art. 14	PARTIAL
Sensitive data extraction	● HIGH	Partial	● MEDIUM	Art. 15	PARTIAL
Membership inference	● MEDIUM	Adequate	● LOW	Art. 15	PASS
Supply chain integrity	● MEDIUM	Adequate	● LOW	Art. 9	PASS

● MEDIUM SEVERITY**Retrieval Source Leakage**

The system exposed internal source material from connected retrieval data.

● HIGH SEVERITY**Excessive Tool Permissions**

The agent retained access to actions beyond its role, increasing the risk of misuse.

● HIGH SEVERITY**Unauthorised Tool Invocation**

The agent was steered into triggering a connected tool outside the intended workflow.

Extend your security offering into AI



Extend Your Security Offering

Add AI adversarial assessment alongside the penetration testing and offensive-security services your customers already buy.



Add Capability Without Building It

Access structured AI assessment capability without first creating an entire specialist practice internally.



Keep the Client Relationship

Meet emerging AI security requirements while retaining ownership of the customer and wider engagement.



Deliver Defensible AI Security Evidence

Turn adversarial testing into structured findings, impact and remediation that fit naturally alongside existing security deliverables.

Add the capability, keep the client relationship

PROVION PROVIDES

- 01 Assessment Capability
- 02 Structured Methodology
- 03 Platform & Evidence Framework
- 04 Scoping & Technical Support

YOU KEEP

- 01 Client Ownership
- 02 The Wider Security Engagement
- 03 Commercial Relationship
- 04 Account Control

A track record of making new technology auditable

Provision comes from a team with experience delivering technology where trust, auditability and emerging regulation matter.

Provision is developed by Protokol, a European technology company experienced in building infrastructure where emerging technology meets regulatory requirements.

EBSI

Contributor to European Commission blockchain infrastructure programmes.

Digital Product Passport

Production DPP solutions delivered ahead of ESPR enforcement.

Horizon Europe

Participant in EU-funded research consortia focused on auditable systems.

WHERE TO START

Start with one client and one AI system

The simplest way to begin is with a defined client system or workflow where AI introduces a clear security requirement.

01 Identify the Opportunity

Choose a client with an AI application, system or workflow that would benefit from adversarial assessment.

02 Scope it Together

Define the system boundary, testing priorities and the evidence the engagement needs to produce.

03 Deliver the Assessment

Provision supports the assessment capability, methodology and technical execution alongside the partner engagement.

04 Build from the First Engagement

Use the initial assessment to establish where AI adversarial testing fits across existing and future client work.



THANK YOU