

AACF - Agentic AI Governance, Risk & Assurance Assessment

Fictional organisation: Atlas Services (fictional) | Date: 22 October 2026

Important: this report is based entirely on fictional scenarios, respondents and outcomes. It illustrates a possible reporting structure; no actual client engagement, independent audit, verification or validated results are claimed.

1. Executive summary

Illustrative context. Fictional Atlas Services is considering an AI agent to summarise case files and prepare draft workflow actions. In this scenario, the agent is not authorised for production use. All observations are scenario assumptions, not technical findings.

Area	Illustrative observation	Provisional interpretation
Authority	Read-only access to case files is proposed; write-enabled tools have not been bounded.	Design boundary open
Human approval	Approval for external actions is mentioned at concept level.	Implementation unknown
Prompt injection	External content could contain untrusted instructions.	Testing needed
Reconstructability	Completeness and integrity of decision/action logs are not evidenced.	Evidence gap
Intervention	No independently tested stop-and-recovery procedure in the scenario.	Evidence gap

2. Assessment question and technical boundaries

Under which permissions, controls and human decision rights could the agent safely support a limited workflow step? Scope: one fictional agent, two hypothetical tools and an example process. No production environment or vulnerability scan.

3. Risk hypotheses

- Untrusted source content attempts to induce a tool action beyond the agent's mandate.
- Excessive permissions allow an action that was not approved.
- An automated assessment is mistakenly treated as independent risk assurance.
- Insufficient logs prevent decisions and actions from being reconstructed.

4. Illustrative control improvements

Priority	Proposed control	Verification criterion
P1	Define explicit allowlists for tools, data and actions; assign human authorisation.	Unauthorised actions are demonstrably denied.
P2	Separate untrusted content from instructions and authorisation decisions.	Representative misuse cases pass agreed tests.



Priority	Proposed control	Verification criterion
P3	Implement independent logging, intervention and recovery.	Actions are reconstructable; stop is available outside the agent.

5. Test and evidence plan

The following are planned evaluations in a fictional scenario; they have not been performed. A real engagement requires access authorisation, mandate, test limits, data handling and disruption-risk arrangements.

Test area	Required evidence	Dependency
Identity & permissions	Allowed versus denied tool actions	IAM rules and owner authorisation
Untrusted content	Safe refusal of manipulative source instructions	Authorised red-team test fixtures
Human approval	Accessible decision and traceable approver	Implemented approval gate
Observability	Traceable request, decision and action data	Controlled log access
Containment	Independent stop and controlled recovery	Operations / security owner

6. Accountability and decision-making

- Business/process owner: intended process outcome and operational acceptance.
- Security, risk and privacy: specialist standards and escalation within their mandate.
- AI engineering and platform teams: implementation and execution of technical controls.
- Human decision-makers retain final accountability; the agent cannot be the sole assessor of its own risk.

7. Status and limits

No validated security assurance conclusion. This report contains no real system logs, audit trail, penetration test, risk acceptance, compliance attestation or independent opinion. AACF is a pilot-tested working framework according to the ModelForge register, not a certification standard in its own right. Legal compliance must be assessed against the laws actually applicable.

8. Recommended next step

Define the bounded AI agent, intended purpose and processing context. Agree owners, evidence requirements and authorised tests. Only then produce evidence-traceable findings, uncertainty statements, priorities and a remediation roadmap.