

AACF - Agentic AI Governance, Risk & Assurance Assessment

Fictieve organisatie: Atlas Services (fictief) | Datum: 22 oktober 2026

Belangrijk: dit rapport gebruikt volledig verzonnen scenario's, respondenten en uitkomsten. Het is uitsluitend een voorbeeld van mogelijke rapportagestructuur, zonder echte opdrachtgever, audit, verificatie of gevalideerd effect.

1. Managementsamenvatting

Demonstratiecontext. De fictieve organisatie Atlas Services overweegt een AI-agent die klantdossiers samenvat en conceptacties in een workflow voorbereidt. De agent heeft geen productieautorisatie in dit voorbeeld. Alle bevindingen zijn scenario-aannames, geen technische observaties.

Aspect	Illustratieve observatie	Voorlopige duiding
Bevoegdheden	Toegang tot read-only dossierinformatie is voorgesteld; write-tools zijn nog niet afgebakend.	Ontwerpgrens open
Menselijke goedkeuring	Goedkeuring voor externe acties is conceptueel benoemd.	Implementatie onbekend
Prompt injection	Externe teksten kunnen onvertrouwde instructies bevatten.	Testplan nodig
Reconstructie	Volledigheid en integriteit van actie- en besluitlogs niet aangetoond.	Evidence gap
Ingrijpen	Geen onafhankelijk geteste stop- en herstelprocedure in het scenario.	Evidence gap

2. Onderzoeksvraag en technische grens

Onder welke bevoegdheden, controles en menselijke besluitrechten kan de agent veilig een beperkt workflowonderdeel ondersteunen? Scope: één fictieve agent, twee hypothetische tools en een voorbeeldproces. Geen productieomgeving, geen kwetsbaarheidsscan.

3. Risicohypothesen

- Onbetrouwbare broninhoud probeert de agent aan te zetten tot een toolactie buiten zijn opdracht.
- Te ruime rechten maken een niet-goedgekeurde actie mogelijk.
- Een geautomatiseerde beoordeling wordt ten onrechte als onafhankelijk risicobewijs gezien.
- Ontbrekende logdetails verhinderen het reconstrueren van beslissingen en acties.

4. Voorbeeldverbeteringen

Prioriteit	Voorgestelde maatregel	Verificatiecriterium
P1	Leg expliciete allowlists vast voor tools, gegevens en acties; bepaal menselijke autorisatie.	Ongeautoriseerde acties worden aantoonbaar geweigerd.
P2	Scheid onvertrouwde input van instructie- en autorisatiebesluiten.	Representatieve misbruikscenario's doorlopen afgesproken tests.



Prioriteit	Voorgestelde maatregel	Verificatiecriterium
P3	Richt onafhankelijke logging, interventie en herstel in.	Acties zijn reconstrueerbaar; stop kan buiten de agent worden uitgevoerd.

5. Test- en evidenceplan

De teststappen hieronder zijn geplande evaluaties in een fictief scenario. Ze zijn niet uitgevoerd. Voor een echte opdracht zijn toegang, mandaat, testgrenzen, verwerking van data en afspraken over verstoringsrisico vereist.

Testfamilie	Gewenst bewijs	Afhankelijkheid
Identiteit & rechten	Toegestane versus geweigerde toolactie	IAM-beleid en eigenaarstoestemming
Untrusted content	Veilige weigering bij manipulerende broninput	Geautoriseerde red-teamfixtures
Menselijke goedkeuring	Beschikbaar besluit en traceerbare approver	Gerealiseerde approval gate
Observability	Herleidbare verzoek-, besluit- en actiegegevens	Afgeschermd logtoegang
Containment	Onafhankelijke stop en gecontroleerd herstel	Operations / security eigenaar

6. Verantwoordelijkheid en besluitvorming

- Business/process owner: beoogd procesresultaat en operationele acceptatie.
- Security / risk / privacy: specialistische normering en escalatie binnen eigen mandaat.
- AI engineering en platform: implementatie en technische controle-uitvoering.
- Menselijke besluitnemers blijven eindverantwoordelijk; een agent beoordeelt niet als enige zijn eigen risico.

7. Status en grenzen

Geen gevalideerd beveiligingsoordeel. Dit rapport bevat geen werkelijke systeemlogs, audittrail, penetratietest, risicoacceptatie, conformiteitsverklaring of onafhankelijk oordeel. AACF is volgens het ModelForge-register een pilotgetest working framework, niet zelfstandig een certificeringsstandaard. Juridische compliance vraagt beoordeling tegen werkelijk toepasselijke regelgeving.

8. Aanbevolen volgende stap

Definieer een afgebakende AI-agent, doel en verwerkingscontext. Stel eigenaars, bewijsvereisten en toegestane testen vast. Lever pas daarna een bevindingenrapport op basis van herleidbaar bewijs, met onzekerheden, prioriteiten en herstelroadmap.