

# Governance Workbench

Cyber Risk Review - Report sperimentale

## Executive summary

Cyber Risk Review - Aurora Health Services. Il Workbench rende osservabili delega, accountability, supervisione umana e rischio sperimentale di Centauro Inverso. Nessuna raccomandazione sostituisce la decisione della persona accountable.

## Indicatori chiave

Delegation: 3 delegabili, 6 supportabili, 3 non delegabili.

Centaur Inverse Risk Score: 44.38/100 (MODERATE). Readiness: 67.75/100; stadio effettivo AI\_SUPPORTED.

## Stato della validazione

Gli indici sono rubriche deterministiche sperimentali, non probabilita. Il prototipo dimostra calcolo riproducibile, tracciabilita e gate, ma non validita predittiva, conformita, riduzione causale di tempi o rischi, sicurezza generale contro prompt injection o validazione su organizzazioni reali.

# 1. Progetto e perimetro

Aurora Health Services (synthetic) | Dominio: Cyber Risk Review | Criticita: HIGH | Stato: ACTIVE

Revisione controllata del rischio cyber per una piattaforma sanitaria completamente fittizia.

## Standard e riferimenti dichiarati

NIS2 | AI Act | ISO/IEC 42001 | NIST AI RMF

## Partecipanti

- H-01 - Andrea Conti - Project Manager (HUMAN)
- H-02 - Luca Ferri - Cyber Security Specialist (HUMAN)
- H-03 - Elena Riva - Risk Owner (HUMAN)
- H-04 - Sara Neri - Compliance Officer (HUMAN)
- H-05 - Marta Greco - Human Reviewer / Approver (HUMAN)
- H-06 - Paolo Serra - Observer / Researcher (HUMAN)
- A-01 - Evidence Collector - Agent Role (AGENT)
- A-02 - Cyber Risk Analyst - Agent Role (AGENT)
- A-03 - Compliance Mapper - Agent Role (AGENT)
- A-04 - Critical Reviewer - Agent Role (AGENT)
- A-05 - Orchestrator - Agent Role (AGENT)

## 2. Matrice di delega e accountability

**ACT-001 DELEGABLE 24.25**

Raccolta evidenze documentali | Accountable: Andrea Conti | Approver: Elena Riva | Agent: Evidence Collector | Controlli: CTRL-01, CTRL-04

**ACT-002 DELEGABLE 31.50**

Normalizzazione vulnerability report | Accountable: Luca Ferri | Approver: Andrea Conti | Agent: Evidence Collector | Controlli: CTRL-01, CTRL-05

**ACT-003 SUPPORTABLE 47.75**

Bozza del risk register | Accountable: Luca Ferri | Approver: Elena Riva | Agent: Cyber Risk Analyst | Controlli: CTRL-02, CTRL-06

**ACT-004 SUPPORTABLE 58.25**

Mappatura NIS2 e AI Act | Accountable: Sara Neri | Approver: Elena Riva | Agent: Compliance Mapper | Controlli: CTRL-02, CTRL-07

**ACT-005 SUPPORTABLE 49.00**

Critical review dei finding | Accountable: Marta Greco | Approver: Andrea Conti | Agent: Critical Reviewer | Controlli: CTRL-03, CTRL-08

**ACT-006 SUPPORTABLE 52.25**

Pianificazione remediation | Accountable: Andrea Conti | Approver: Elena Riva | Agent: Cyber Risk Analyst | Controlli: CTRL-02

**ACT-007 SUPPORTABLE 43.50**

Executive security reporting | Accountable: Andrea Conti | Approver: Elena Riva | Agent: Orchestrator | Controlli: CTRL-02, CTRL-09

**ACT-008 SUPPORTABLE 39.75**

Verifica copertura controlli | Accountable: Sara Neri | Approver: Marta Greco | Agent: Compliance Mapper | Controlli: CTRL-01, CTRL-07

**ACT-009 NON DELEGABLE 79.00**

Accettazione finale del rischio | Accountable: Elena Riva | Approver: Marta Greco | Agent: - | Controlli: CTRL-03, CTRL-10  
Hard gate: FINAL\_RISK\_ACCEPTANCE, LEGAL\_CONSEQUENCE

**ACT-010 NON DELEGABLE 86.25**

Modifica controllo in produzione | Accountable: Luca Ferri | Approver: Elena Riva | Agent: - | Controlli: CTRL-03  
Hard gate: PRODUCTION\_CHANGE, SECURITY\_CONTROL\_DISABLE

**ACT-011 NON DELEGABLE 71.50**

Approvazione compliance | Accountable: Sara Neri | Approver: Elena Riva | Agent: - | Controlli: CTRL-03, CTRL-07  
Hard gate: LEGAL\_OR\_COMPLIANCE\_APPROVAL

**ACT-012 DELEGABLE 29.50**

Controllo integrità delle fonti | Accountable: Marta Greco | Approver: Andrea Conti | Agent: Critical Reviewer | Controlli: CTRL-04, CTRL-08

### 3. Human Oversight e controlli

- CTRL-01 - Source allowlist e identificazione - PREVENTIVE - ACTIVE - Owner: Marta Greco - Evidenza: EV-001
- CTRL-02 - Evidence-linked recommendation - PREVENTIVE - ACTIVE - Owner: Andrea Conti - Evidenza: EV-004
- CTRL-03 - Independent human approval - APPROVAL - ACTIVE - Owner: Elena Riva - Evidenza: EV-009
- CTRL-04 - Integrity e provenance check - DETECTIVE - ACTIVE - Owner: Marta Greco - Evidenza: EV-002
- CTRL-05 - Schema validation - PREVENTIVE - ACTIVE - Owner: Luca Ferri - Evidenza: Test automatici
- CTRL-06 - Confidence threshold - DETECTIVE - ACTIVE - Owner: Luca Ferri - Evidenza: Policy POL-04
- CTRL-07 - Professional compliance review - APPROVAL - ACTIVE - Owner: Sara Neri - Evidenza: HG-02
- CTRL-08 - Prompt injection isolation - PREVENTIVE - ACTIVE - Owner: Marta Greco - Evidenza: EV-012
- CTRL-09 - Executive claim review - DETECTIVE - DRAFT - Owner: Andrea Conti - Evidenza: -
- CTRL-10 - Risk acceptance evidence pack - APPROVAL - MISSING - Owner: Elena Riva - Evidenza: -

# 4. Centauro Inverso e maturita

Score Centauro: 44.38/100 - fascia MODERATE - modello centaur-1.0.0-experimental

- Accettazione passiva delle raccomandazioni: 2/4
- Debolezza della verifica indipendente: 2/4
- Scarsa tracciabilità delle fonti: 1/4
- Limitata spiegabilità: 2/4
- Degradazione delle competenze: 2/4
- Ambiguità dell'accountability: 1/4
- Pressione temporale: 3/4
- Concentrazione dell'automazione: 2/4
- Assenza di fallback: 1/4
- Ridotta capacità di contestare l'agente: 2/4

## Maturity

Readiness 67.75/100 | potenziale AI\_SUPPORTED | effettivo AI\_SUPPORTED | modello maturity-1.0.0

- governance: superato
- securityCompliance: non superato
- humanOversight: superato
- criticalControls: non superato

## 5. Cyber Risk Review

Run RUN-001 | Stato AWAITING\_SOURCE\_APPROVAL | Provider MOCK

### Finding

RF-001 - HIGH - Copertura incompleta del controllo di risk acceptance - Evidenze EV-004, EV-011 - HUMAN\_REVIEW

RF-002 - HIGH - Privilegi eccessivi nel ruolo di service operator - Evidenze EV-005 - OPEN

RF-003 - MEDIUM - Restore test parziale - Evidenze EV-006 - ACCEPTED

RF-004 - MEDIUM - Supplier assurance da completare - Evidenze EV-007 - OPEN

RF-005 - CRITICAL - Indirect prompt injection nel contenuto vendor - Evidenze EV-012 - ACCEPTED

RF-006 - HIGH - Impatto legale non sufficientemente dimostrato - Evidenze EV-010 - REJECTED

### Security flags

EV-012 - INDIRECT\_PROMPT\_INJECTION: contenuto isolato e mai eseguito.

## 6. Metriche calcolate e incertezza

Regola: una percentuale è pubblicata soltanto con numeratore e denominatore. Gli intervalli per le proporzioni sono Wilson al 95%. N/D indica che il dataset non consente una stima.

### KPI-01 - Tempo risparmiato: N/D

Stato NOT\_ESTIMABLE | Fonte Nessun dato osservazionale | Formula  $\text{media}(\text{tempo baseline}) - \text{media}(\text{tempo con AI})$

Numeratore N/D | Denominatore N/D | n N/D | IC95 N/D

Limite: Mancano timestamp comparabili e una baseline senza AI; il precedente valore dimostrativo è stato rimosso.

### KPI-02 - Tempo medio di escalation: N/D

Stato NOT\_ESTIMABLE | Fonte Nessuna coppia temporale completa | Formula  $\text{sum}(\text{timestamp risoluzione} - \text{timestamp apertura}) / \text{escalation risolte}$

Numeratore N/D | Denominatore N/D | n N/D | IC95 N/D

Limite: L'audit non registra ancora apertura e risoluzione abbinate per ogni escalation.

### KPI-03 - Copertura evidenze verificate: 75%

Stato MEASURED | Fonte Registro evidenze del progetto | Formula  $\text{evidenze VERIFIED} / \text{evidenze totali} \times 100$

Numeratore 9 | Denominatore 12 | n 12 | IC95 46.8-91.1%

Limite: Descrive lo stato del registro; non dimostra veridicità, qualità sostanziale o completezza delle fonti.

### KPI-04 - Accuratezza rispetto a un gold standard: N/D

Stato NOT\_ESTIMABLE | Fonte Gold standard assente | Formula  $\text{valutazioni corrette} / \text{valutazioni con etichetta indipendente} \times 100$

Numeratore N/D | Denominatore N/D | n N/D | IC95 N/D

Limite: Mancano etichette indipendenti, protocollo cieco e campione reale; i confidence score degli agenti non sono accuratezza.

### KPI-05 - Attività con override umano: 8.3%

Stato MEASURED | Fonte Attività e audit log | Formula  $\text{attività uniche con override} / \text{attività totali} \times 100$

Numeratore 1 | Denominatore 12 | n 12 | IC95 1.5-35.4%

Limite: Non indica se gli override fossero corretti o se abbiano ridotto il rischio.

### KPI-06 - Controlli critici attivi: 85.7%

Stato MEASURED | Fonte Control Registry | Formula  $\text{controlli critici ACTIVE} / \text{controlli critici totali} \times 100$

Numeratore 6 | Denominatore 7 | n 7 | IC95 48.7-97.4%

Limite: ACTIVE è uno stato applicativo: non equivale a test di efficacia né a certificazione di conformità.

### KPI-07 - Eventi bloccati registrati: 1

Stato MEASURED | Fonte Audit log | Formula  $\text{conteggio eventi con outcome BLOCKED}$

Numeratore 1 | Denominatore N/D | n 8 | IC95 N/D

Limite: È un conteggio descrittivo, non un tasso di efficacia e non include tentativi non osservati.

### KPI-08 - Canary di prompt injection isolato: 1/1

Stato SIMULATED | Fonte Un solo caso sintetico EV-012 | Formula  $\text{canary sintetici isolati} / \text{canary sintetici inseriti}$

Numeratore 1 | Denominatore 1 | n 1 | IC95 20.7-100%

Limite: n=1 sintetico: non generalizza ad altri attacchi, modelli, lingue o contesti

# APM-GF Governance Report - continuazione

e non dimostra che la prompt injection sia risolta.

**KPI-09 - Controlli di approvazione umana attivi: 66.7%**

Stato MEASURED | Fonte Control Registry | Formula controlli APPROVAL in stato ACTIVE  
/ controlli APPROVAL totali × 100

Numeratore 2 | Denominatore 3 | n 3 | IC95 20.8-93.9%

Limite: Non misura indipendenza, qualità o tempestività effettiva della revisione umana.

**KPI-10 - Fallback operativo verificato: N/D**

Stato NOT\_ESTIMABLE | Fonte Test di fallback assente | Formula prove di fallback superate / prove pianificate × 100

Numeratore N/D | Denominatore N/D | n N/D | IC95 N/D

Limite: Il registro contiene controlli dichiarati, ma nessun test eseguito con esito e criteri di accettazione.



## 7. Statistiche descrittive

Score di delega:  $n=12$ ; media=51.04; mediana=48.38; deviazione standard campionaria=19.67; min-max=24.25-86.25.

Delegabili: 3/12 (25%), IC95 8.9-53.2%.

Supportabili: 6/12 (50%), IC95 25.4-74.6%.

Non delegabili: 3/12 (25%), IC95 8.9-53.2%.

Queste statistiche descrivono solo lo scenario sintetico corrente e non dimostrano generalizzabilità o validità predittiva.

## 8. Protocollo di validazione

### VAL-01 - Gli score predicono comportamenti o incidenti

Stato: NOT\_TESTED. Evidenza corrente: Gli score descrivono risposte e regole del prototipo; non esistono outcome longitudinali associati.

Studio richiesto: Campione reale prospettico, outcome predefiniti, validazione temporale ed esterna.

Metriche: discriminazione (AUC/c-statistic); calibrazione e calibration slope; Brier score; sensibilità e specificità; utilità decisionale

### VAL-02 - L'organizzazione è conforme a norme o standard

Stato: DESCRIPTIVE\_ONLY. Evidenza corrente: Il prototipo registra controlli, owner ed evidenze dichiarate; non svolge una valutazione formale di conformità.

Studio richiesto: Mappatura requisito-per-requisito, prove verificabili, scope definito e revisione da professionisti qualificati o organismo competente.

Metriche: requisiti applicabili; requisiti coperti; gap aperti; qualità e validità delle evidenze; esito della verifica indipendente

### VAL-03 - L'AI riduce realmente tempi o rischi

Stato: NOT\_TESTED. Evidenza corrente: Non sono presenti baseline, gruppo di confronto o osservazioni appaiate.

Studio richiesto: Confronto predefinito tra workflow baseline e AI-assisted, idealmente randomizzato o appaiato, con campione adeguato.

Metriche: differenza media o mediana dei tempi; effect size; intervallo di confidenza 95%; errori e incidenti per unità di esposizione; carico di revisione umana

### VAL-04 - La prompt injection è risolta in generale

Stato: SIMULATED\_ONLY. Evidenza corrente: È isolato un singolo canary sintetico noto (EV-012).

Studio richiesto: Benchmark ripetibile con attacchi diversificati e input benigni, su modelli, lingue e template diversi.

Metriche: attack success rate; recall/true-positive rate; precision; false-positive rate; intervalli di confidenza; regressioni tra versioni

### VAL-05 - Il modello è validato su organizzazioni reali

Stato: NOT\_TESTED. Evidenza corrente: Tutti i record, i nomi e gli outcome inclusi sono sintetici.

Studio richiesto: Pilota multi-organizzazione con valutatori indipendenti, casi eterogenei, protocollo preregistrato e analisi di trasferibilità.

Metriche: content validity; Cohen kappa pesato per le classi; ICC per gli score continui; test-retest; sensibilità a pesi e soglie; validazione esterna

## 9. Motori di score

### Delegation Engine - delegation-1.0.0

Formula:  $100 \times \text{sum}(\text{valore criterio } 0-4 \times \text{peso}) / \text{sum}(4 \times \text{peso})$

Pesi: impact=15; irreversibility=12; dataSensitivity=12;  
regulatoryResponsibility=15; ambiguity=10; verificationDifficulty=12;  
traceabilityWeakness=8; privileges=10; fallbackWeakness=6

Soglie: <35 Delegabile; 35-<65 Supportabile; >=65 Non delegabile; gli hard gate prevalgono sullo score

Evidenza: Ipotesi di design, non modello predittivo validato

### Centauro Inverso - centaur-1.0.0-experimental

Formula: media dei segnali per fattore; poi  $100 \times \text{sum}(\text{fattore } 0-4 \times \text{peso}) / \text{sum}(4 \times \text{peso})$

Pesi: passiveAcceptance=20; independentVerification=20;  
traceabilityExplainability=15; skillDegradation=15; ambiguousAccountability=10;  
pressureDependency=10; missingFallback=10

Soglie: <25 Basso; 25-<50 Moderato; 50-<75 Alto; >=75 Critico

Evidenza: Indice sperimentale descrittivo; non inferisce comportamenti o incidenti

### Maturity Engine - maturity-1.0.0

Formula:  $100 \times \text{sum}(\text{dimensione } 0-4 \times \text{peso}) / \text{sum}(4 \times \text{peso})$ , poi applicazione dei maturity gate

Pesi: governance=20; peopleSkills=15; processes=20; technologyData=15;  
securityCompliance=20; measurement=10

Soglie: <40 Human-Designed; 40-<75 AI-Supported; >=75 potenziale LLM-First, consentito solo con tutti i gate

Evidenza: Rubrica di readiness; non certifica maturità o conformità

# 10. Roadmap

RM-01 - HUMAN\_DESIGNED - HIGH - Formalizzare le policy di delega - Owner Andrea Conti - DONE

RM-02 - HUMAN\_DESIGNED - HIGH - Completare l'inventario delle evidenze - Owner Marta Greco - IN\_PROGRESS

RM-03 - AI\_SUPPORTED - MEDIUM - Istituire esercizi human-only - Owner Luca Ferri - PLANNED

RM-04 - AI\_SUPPORTED - HIGH - Attivare approval gate indipendenti - Owner Elena Riva - IN\_PROGRESS

RM-05 - AGENT\_ENABLED - HIGH - Versionare prompt e output schema - Owner Marta Greco - PLANNED

RM-06 - AGENT\_ENABLED - HIGH - Eseguire red-team sulla prompt injection - Owner Luca Ferri - PLANNED

RM-07 - LLM\_FIRST - HIGH - Validare gli indici con esperti - Owner Paolo Serra - PLANNED

RM-08 - LLM\_FIRST - HIGH - Definire il fallback operativo - Owner Andrea Conti - PLANNED

# 11. Decisioni e audit essenziale

2026-08-18T14:32:00Z | Elena Riva (HUMAN) | DECISION\_REJECTED | REJECTED | Impatto non supportato da evidenze indipendenti | CORR-CRR-001

2026-08-18T14:26:00Z | Critical Reviewer (AGENT) | CRITICAL\_REVIEW\_COMPLETED | SUCCESS | Due security flag e tre unknown rilevati | CORR-CRR-001

2026-08-18T14:18:00Z | Sistema (SYSTEM) | PROMPT\_INJECTION\_ISOLATED | BLOCKED | Istruzione nel documento trattata come contenuto non attendibile | CORR-CRR-001

2026-08-18T14:11:00Z | Andrea Conti (HUMAN) | SOURCE\_GATE\_APPROVED | SUCCESS | Fonti verificate e dataset sintetico confermato | CORR-CRR-001

2026-08-18T13:55:00Z | Sara Neri (HUMAN) | DELEGATION\_OVERRIDE | SUCCESS | Revisione umana rafforzata per requisito NIS2 | CORR-ASS-004

2026-08-18T13:42:00Z | Evidence Collector (AGENT) | EVIDENCE\_INDEXED | SUCCESS | Indicizzazione completata senza tool esterni | CORR-CRR-001

2026-08-18T12:20:00Z | Sistema (SYSTEM) | MATURITY\_SNAPSHOT\_CREATED | SUCCESS | Modello maturity-1.0.0 | CORR-MAT-001

2026-08-18T11:00:00Z | Andrea Conti (HUMAN) | PROJECT\_CREATED | SUCCESS | Avvio scenario sintetico di ricerca | CORR-PRJ-001