

2025 Daniel P. Madden
License: CC BY-NC-ND 4.0

AI OSI Stack Condensed Risk Taxonomy

Author: Daniel P. Madden

Version: v4 Blueprint Integration

Date: November 2025

> ## Normative Language Notice

> This document uses normative language consistent with ISO/IEC 42010 and NIST conventions.

> SHALL denotes mandatory requirements, SHOULD denotes strong recommendations, and MAY denotes optional practices.

> Interpretations SHALL preserve authorial intent: layered accountability, epistemic integrity, and human dignity as binding design constraints.

1. Layer 1 Physical and Compute Substrate

- **Risk Themes:** Hardware tampering, supply chain disruption, resilience degradation.

- **Sentinel Indicators:** Facility access anomalies, unscheduled firmware changes, energy draw deviations.

- **Primary Controls:** Certified facility audits, redundant energy governance, tamper-evident custody logs.

- **Required Evidence:** GDS section L1, DRR infrastructure appendix, ILE temporal seals for facility inspections.

2. Layer 2 Data Stewardship

- **Risk Themes:** Data poisoning, consent erosion, epistemic contamination.

- **Sentinel Indicators:** Provenance gaps, unexplained distribution shifts, contested consent revocations.

- **Primary Controls:** Consent traceability ledgers, dataset review boards, epistemic hygiene playbooks.

- **Required Evidence:** ITP provenance bundles, DRR data stewardship summaries, AEIP lineage signatures.

3. Layer 3 Model Development

- **Risk Themes:** Bias amplification, evaluation blind spots, adversarial regressions.

- **Sentinel Indicators:** Diverging fairness metrics, untested safety scenarios, failed regression thresholds.

- **Primary Controls:** Composite evaluation harnesses, adversarial benchmarking cadences, persona-aligned regression suites.

- **Required Evidence:** OAM evaluation actions, GDS validation annex, DRR design decisions.

4. Layer 4 Instruction and Control

- **Risk Themes:** Prompt injection, persona drift, affect misalignment, refusal degradation.

- ****Sentinel Indicators:**** Persona mismatch alerts, refusal override logs, affect boundary excursions.
- ****Primary Controls:**** Persona mandate enforcement, refusal logic penetration testing, affect boundary audits.
- ****Required Evidence:**** ITP instruction traces, DRR persona matrices, GDS instruction governance statements.

5. Layer 5 Reasoning Exchange and Interface

- ****Risk Themes:**** Ledger desynchronization, packet replay, counter-signature gaps, handshake tampering.
- ****Sentinel Indicators:**** Hash mismatches, delayed acknowledgements, anomalous replay counts.
- ****Primary Controls:**** AEIP handshake validation, deterministic replay checkpoints, cross-node quorum review.
- ****Required Evidence:**** ILE handshake bundles, OAM remediation logs, AEIP transcript archives.

6. Layer 6 Deployment and Integration

- ****Risk Themes:**** Runtime drift, integration conflicts, incident opacity.
- ****Sentinel Indicators:**** Unauthorized configuration changes, failed deployment gates, unreported incidents.
- ****Primary Controls:**** Change gating with governance approvals, blue/green stewardship protocols, incident rehearsal cadences.
- ****Required Evidence:**** TRR release attestations, OAM incident records, ILE change approvals.

7. Layer 7 Governance Publication

- ****Risk Themes:**** Delayed disclosure, accountability gaps, legitimacy erosion.
- ****Sentinel Indicators:**** Missed publication cadences, stakeholder complaints, inconsistent disclosure metadata.
- ****Primary Controls:**** Public disclosure schedules, multi-stakeholder review councils, archival notarization.
- ****Required Evidence:**** GDS transparency packs, ILE publication proofs, custodial meeting minutes.

8. Cross-Cutting Dynamics

- ****Temporal Integrity:**** Validity windows SHALL be enforced with temporal seals and quarterly audits.
- ****Custodianship:**** Governance councils SHALL document every change request, remediation, and escalation in AEIP-compatible ledgers.
- ****Human Dignity:**** Affective constraint modules SHALL be monitored for erosion; violations trigger immediate OAM issuance and persona retraining.