Security-First AI Governance: A Metrics-Driven Framework for Quantifying Compliance-Security Gaps in AI-Augmented Systems

Keerthana Madhavan, Abbas Yazdinejad, Fattane Zarrinkalam, Ali Dehghantanha
University of Guelph
Guelph, Ontario, Canada
{kmadhava,ayazdine,fzarrink,adehghan}@uoguelph.ca

4th International Workshop on Designing and Measuring Security in Systems with AI
July 4th, 2025 — Venice, Italy
Co-located with EuroS&P 2025

UNIVERSITY of GUELPH

Motivation & Problem

AI Compliance ≠ AI Security

- 55 AI Security breaches in 2023 involved “complaint” systems
- Current frameworks focus on ethics & privacy, overlook security vulnerabilities.
- 60-80% of high-risk AI vulnerabilities remain unaddressed.

Bottom Line: Organizations appear complaint but remain exposed to AI-specific attacks

Related Work

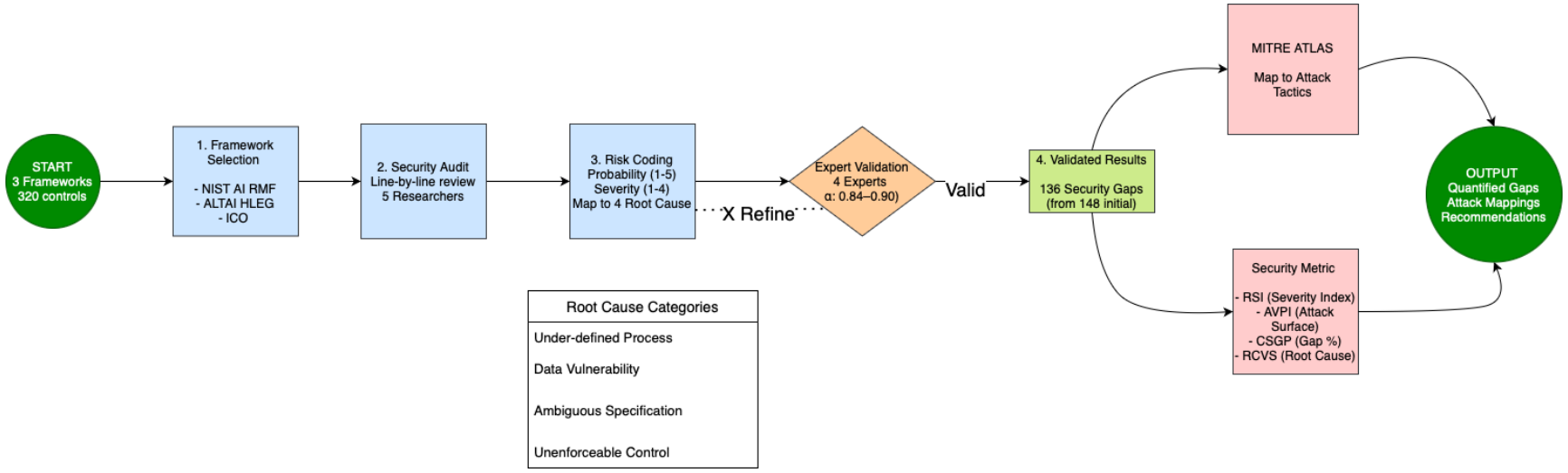
- **Governance Frameworks** like NIST AI RMF, ALTAI, and ICO focus on ethics, bias, and privacy — but overlook AI-specific security threats.
- Prior studies (e.g. Stevens et al.) critique vague cybersecurity controls, but **don't address AI-specific adversarial risks**.
- Recent work (Xia et al., 2023) maps high-level risk categories but lacks **quantitative metrics or threat alignment**.
- Existing efforts do **not quantify** risk severity or attack surface exposure in AI compliance frameworks.

Gap: No previous study performs line-by-line audits or maps AI security weaknesses to MITRE ATLAS tactics.

Research Question

How can we systematically quantify security gaps in existing AI compliance standards?

Methodology



Why Identify Root Causes of Compliance Gaps?

- Surface-level gaps (e.g. missing controls) often stem from **deeper structural issues**.
- We categorized each issue by its **underlying cause**, not just its symptom.
- Four root cause types:
  - Under-defined Processes
  - Ambiguous Specifications
  - Data Vulnerabilities
  - Unenforceable Controls
- Root causes reveal **where to intervene for systemic improvements**.

Bottom Line: Understanding the “why” behind gaps helps move from patching symptoms to fixing frameworks.

Four Security Metrics

Metric	Purpose	Formula
RSI (Risk Severity Index)	Average severity of unresolved vulnerabilities	$RS_i = Probability_i \times Severity_i$ $RSI = \frac{\sum_{i=1}^{n} RS_i}{n}$
AVPI (Attack Vector Potential Index)	Compound attack surface from overlapping gaps	$AVPI = \sum_{c=1}^{k} \left(\frac{\lvert C_c \rvert}{\lvert C_{\text{total}} \rvert} \cdot RCVS_c \right)$
CSGP(Compliance-Security Gap %)	Percentage of high-risk issues unaddressed	$CSGP = \frac{\lvert C_{\text{unaddressed}} \rvert}{\lvert C_{\text{total}} \rvert} \times 100$
RCVS (Root Cause Vulnerability Score)	Which weakness categories drive most risk	$RCVS_c = \frac{\sum_{i \in C_c} RS_i}{\sum_{i=1}^{n} RS_i}$

MITRE ATLAS Mapping

- Compliance frameworks describe **what** to do — MITRE ATLAS shows **how attackers exploit what's missing**.
- Mapping vague or absent controls to adversarial tactics reveals **real attack paths**.
- Helps prioritize remediation: Not all gaps are equal — some align with **high-impact, known tactics**.
- Example: Unclear credential policies → **Credential Access** (AML.TA0013)

Bottom Line: MITRE mapping makes abstract gaps actionable by connecting them to real adversarial behaviors.

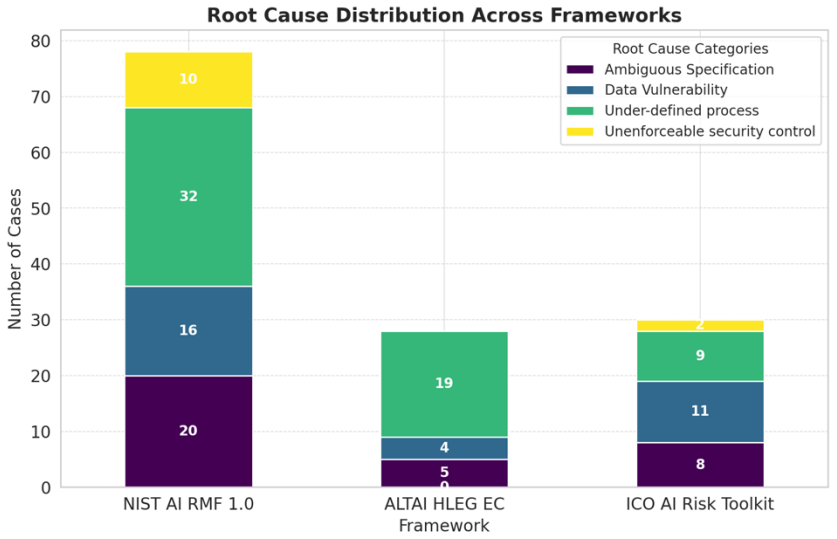
Framework Audit: Metrics Overview

Framework	Total Security Controls	# Attack Vectors	Concerns	RSI	AVPI	CSGP (%)
NIST AI RMF 1.0	152	56	78	10.54	0.29	69.23
ALTAI HLEG EC	72	16	28	9.21	0.51	75.00
ICO AI Toolkit	96	17	30	10.10	0.30	80.00

Bottom Line: All three frameworks leave 60–80% of high-risk issues unaddressed.

Root Cause Analysis

- **Where Frameworks Fail**
  - **Under-defined Processes** (40–67%)
    - Unclear model lifecycle
    - No deprecation protocols
- **Data Vulnerabilities** (15–38%)
  - Missing integrity checks
  - Incomplete data-flow controls

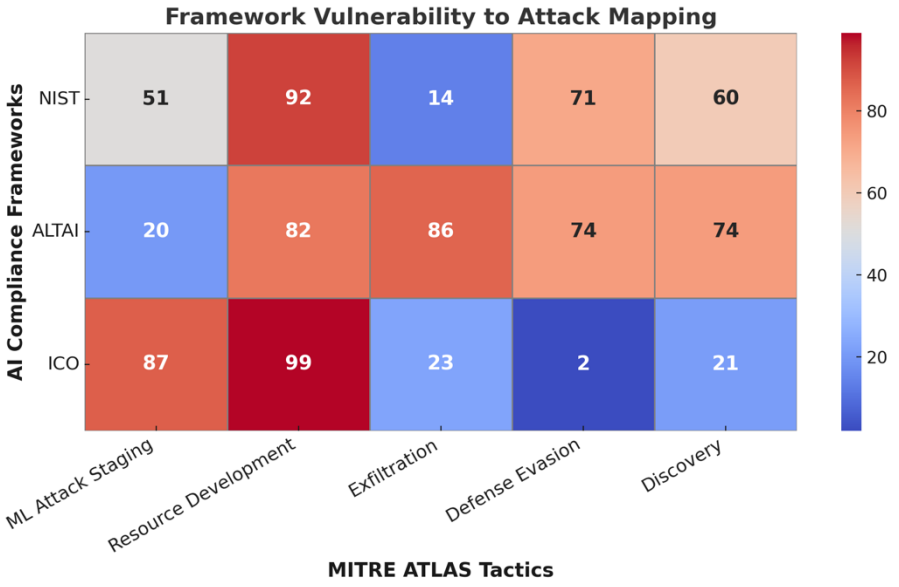


Bottom Line: Frameworks fail most where specificity and enforceability are missing.

MITRE ATLAS MAPPING

From Gaps to Attack Vectors

- Top enabled tactics:
  - ML Attack Staging *(AML.TA0001)*
  - Defense Evasion *(AML.TA0007)*
  - Collection *(AML.TA0009)*
  - Resource Development *(AML.TA0003)*
  - Impact *(AML.TA0011)*



Bottom Line: Gaps in governance align directly with adversarial techniques.

Limitations of This Study

- Focused on 3 frameworks (NIST AI RMF, ALTAI, ICO Toolkit).
- Manual annotation process, though rigorously validated.
- Small expert validation panel (4 reviewers).
- Analyzes written frameworks, not implementation practices.

Bottom Line: Results reflect design-stage risks, not deployment audits.

Key Recommendations

- **Clarify AI Lifecycle Steps**
  → Define training, retraining, decommissioning
- **Enforce Data Protections**
  → Map flows, validate integrity, encrypt
- **Make Controls Operable**
  → Replace suggestions with enforceable rules
- **Test Against Real Threats**
  → Adopt MITRE ATLAS-based adversarial testing

Bottom Line: Our metrics support proactive, threat-informed trade-off decisions in AI system design.

Security Trade-Off Analysis

- RSI helps prioritize most severe unmitigated vulnerabilities.
- AVPI shows how compound risks expand the attack surface.
- CSGP reflects real coverage gaps beyond checkbox compliance

Future Work

- Extend analysis to more frameworks (e.g., EU AI Act).
- Scale expert panel to 15–20 diverse stakeholders.
- Launch longitudinal study of framework evolution.
- Build semi-automated audit and mapping tools.
- Translate metrics into operational lifecycle policies.

Key Takeaway

 Compliance ≠ Security

- “60–80% of high-risk AI vulnerabilities remain unaddressed.”
- Current frameworks provide assurance—but not protection.
- Our metrics make gaps visible, measurable, and actionable.
- Securing AI systems requires moving beyond checklists.

Questions