HAIAMM - Comprehensive Handbook v2.0

Last Updated: 2025-12-29

NEW IN v2.0:

- **Assessment Methodology**: Comprehensive questionnaire-based assessment guide similar to OpenSAMM v1.0, including scoring methodology, 5-phase assessment process, industry benchmarks, and best practices
- **Threat Intelligence as Foundational Capability**: Integrated into Strategy & Metrics (SM) practice across all 6 domains, elevating threat intelligence from advanced capability to essential foundational requirement

NEW IN v2.1.1: Comprehensive Prompt Injection Security guidance integrated across 15 practice one-pagers, derived from the Arcanum Prompt Injection Taxonomy by Jason Haddix (CC BY 4.0)

Table of Contents

Executive Summary

The **Human Assisted Intelligence Assurance Maturity Model (HAIAMM)** provides a comprehensive framework for organizations designing and implementing AI to automate workflows, augment capabilities. What we refer to as Human Assisted Intelligence (HAI). Therefore, HAIAMM addresses the governance, building, verification, and operations of HAI systems with foundational practices to ensure trust, safety, and security.

Key Features

- **12 Security Practices** HAIAMM is built upon the following business functions of AI system adoption; Governance → Building → Verification → Operations.
- **6 Assurance Domains** these are similarly based on the OpenSAMM future talk and concepts. HAIAMM is covering the comprehensive technology stack (Software, Infrastructure, Endpoints, Data, Processes, Vendors)
- **3 Maturity Levels** enabling progressive capability secure building
- **72 Practice-Domain Combinations** with detailed one-pager guidance
- **Threat Intelligence as Foundational Capability** across all domains (consumption, analysis, production maturity progression)
- **Comprehensive Prompt Injection Security** based on Arcanum PI Taxonomy (13 attack intents, 18 techniques, 20 evasion methods)
- **Critical HAI Assurance (v2.2)** addressing 4 agentic AI risks: Excessive Agency, Agent Goal Hijack, Tool Misuse, Rogue Agents
- **"Least Agency Principle"** foundational governance for human-AI collaboration (v2.2)
- **95% OWASP Alignment** coverage of LLM Top 10 (2025) and Agentic Top 10 (2026)
- **290+ Assessment Questions** extracted from practice one-pagers for maturity measurement (243 base + 47 v2.2 additions)
- **Effort Estimates** based on OpenSAMM version 1.0 methodology for realistic planning
- **Framework Alignment** with ISO 27001, NIST CSF, NIST AI RMF, SAMM v1.0, OWASP

Target Audience

- **Security Leaders**: CISOs, Security Directors planning HAI security programs
- **Security Engineers**: Teams building and operating AI security systems (SAST, DAST, CSPM, EDR, DLP, SOAR)
- **AI/ML Engineers**: Teams developing AI models for security applications
- **AI/Governance**:Teams governing AI efforts in organizations
- **Risk & Compliance**: Teams ensuring AI security meets regulatory requirements
- **Vendors**: AI and automation vendors implementing AI-powered capabilities
- **IT Leaders**: Teams managing IT infrastructures and engineers automating workflows throughout the enterprise

---

Introduction

What is HAIAMM?

HAIAMM is a maturity model specifically designed for organizations designing and implementing **Human Assisted Intelligence** systems. Unlike traditional frameworks, HAIAMM addresses:

1. **HAI Governance**: Governing HAI deployments with appropriate human oversight and accountability
2. **HAI System Security**: Securing the AI systems themselves (models, data, infrastructure, endpoints)
3. **Trust, Safety, and Security**: Ensuring HAI solutions are trustworthy, safe, and secure across their lifecycle
4. **Human Oversight**: Balancing AI capabilities with human authority, validation, and control

Why HAIAMM?

Traditional security frameworks (ISO 27001, NIST CSF) weren't designed for Human Assisted Intelligence deployments. HAIAMM fills this gap by providing:

- **HAI-Specific Guidance**: Addresses AI model security, adversarial ML, explainability, bias, human oversight
- **Human Oversight Requirements**: Defines appropriate human control, approval, and validation mechanisms
- **Trust, Safety, and Security**: Comprehensive approach beyond security alone
- **Cross-Domain Coverage**: Unified approach across software, data, infrastructure, vendors, processes, endpoints
- **Practical Implementation**: Detailed checklists, assessment questionnaires, success metrics, effort estimates

HAI Use Cases Covered

HAIAMM applies to organizations implementing:

- **Automation Workflows**: HAI systems handling business processes with human oversight
- **Security Testing & Code Analysis**: AI-powered SAST/DAST tools, vulnerability scanning, automated code review
- **Customer Service Automation**: Chatbots and virtual assistants with human escalation pathways
- **Decision Support Systems**: HAI providing recommendations with human approval and validation
- **Data Processing & Analysis**: HAI assisting humans in pattern detection, insights generation, classification
- **Compliance Automation**: AI-assisted policy enforcement, audit trail generation, reporting

---

Framework/Model Overview

Structure

HAIAMM is organized as a **2-dimensional matrix**:

Dimension 1: Security Practices (12) Governance Practices:

1. Strategy & Metrics (SM)
2. Policy & Compliance (PC)
3. Education & Guidance (EG)

Building Practices: 4. Threat Assessment (TA) 5. Security Requirements (SR) 6. Security Architecture (SA)

Verification Practices: 7. Design Review (DR) 8. Implementation Review (IR) 9. Security Testing (ST)

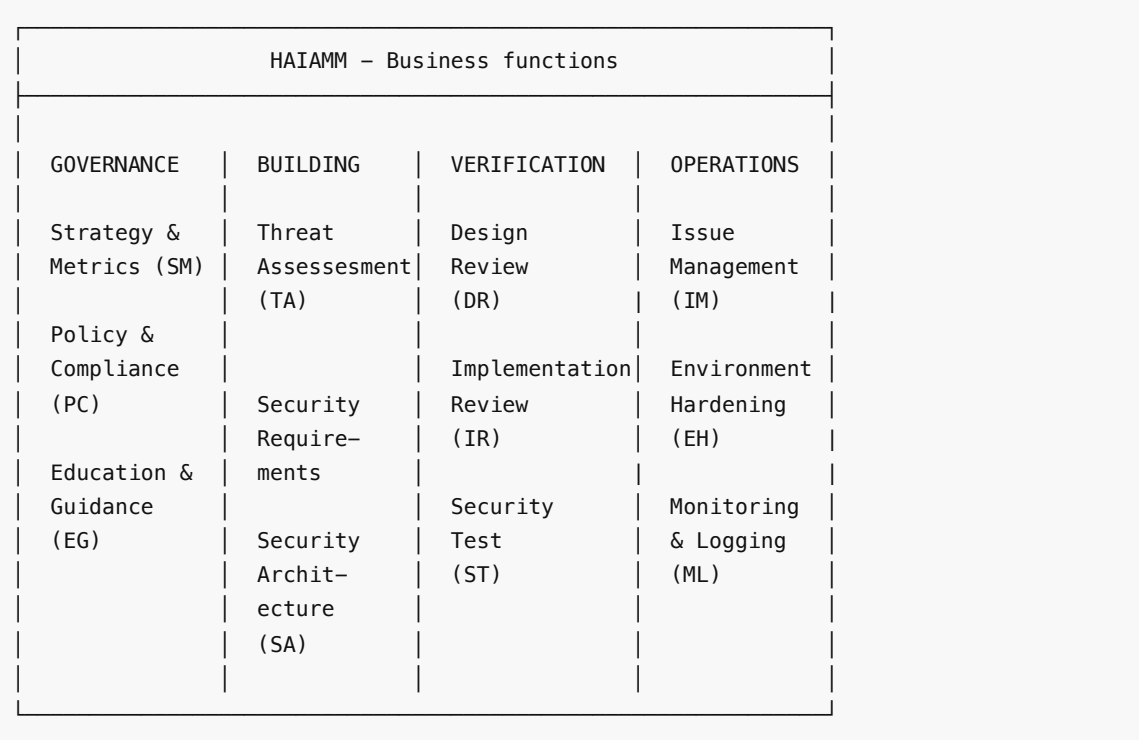
Operations Practices: 10. Issue Management (IM) 11. Environment Hardening (EH) 12. Monitoring & Logging (ML)

Dimension 2: Assurance Domains (6)

1. Software
2. Data
3. Infrastructure
4. Vendors
5. Processes
6. Endpoints

Practice Lifecycle

Practices follow the **Business functions of Human Assisted Intelligence**:



Business functions and practices Descriptions

Governance Phase:

- **Strategy & Metrics (SM)**: Establish a unified strategic roadmap for HAI security across all domains. Includes threat intelligence integration as foundational capability and metrics to measure desired outcomes
- **Policy & Compliance (PC)**: Define policies, standards, and compliance requirements for HAI security systems. Ensures regulatory compliance (e.g. GDPR, HIPAA, SOX) and internal governance
- **Education & Guidance (EG)**: Provide security training, awareness, and guidance for teams building and operating AI security systems. For example, covers secure AI development, prompt injection prevention, and operational best practices

Building Phase:

- **Threat Assessment (TA)**: Identify threats specific to AI systems (adversarial attacks, data poisoning, model theft, automation abuse, prompt injection)
- **Security Requirements (SR)**: Define measurable security requirements for AI systems (accuracy, latency, privacy, compliance)
- **Security Architecture (SA)**: Design secure, scalable architecture for AI security systems

Verification Phase:

- **Design Review (DR)**: Review designs before implementation to catch flaws early
- **Implementation Review (IR)**: Code review to ensure secure implementation following best practices
- **Security Testing (ST)**: Test AI systems for vulnerabilities, adversarial robustness, and performance

Operations Phase:

- **Issue Management (IM)**: Continuous vulnerability scanning and remediation (dependencies, models, infrastructure)
- **Environment Hardening (EH)**: Harden deployment environments (least privilege, encryption, network segmentation)
- **Monitoring & Logging (ML)**: Comprehensive logging and monitoring for security, compliance, and debugging

---

Prompt Injection Security

NEW IN v2.0: HAIAMM now includes comprehensive guidance on securing HAI systems against **prompt injection attacks**, derived from the Arcanum Prompt Injection Taxonomy by Jason Haddix (CC BY 4.0). and [OWASP top 10 for Agentic Applications] (https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/)

Why Prompt Injection Matters for AI Security

AI security systems increasingly use Large Language Models (LLMs) for:

- **Code Review Assistants**: AI analyzing code for security vulnerabilities
- **Security Chatbots**: AI assistants helping security teams investigate incidents
- **AI-Powered SAST/DAST**: LLMs augmenting traditional security scanning
- **SOAR Platforms**: AI-driven incident triage and response automation
- **DLP Systems**: AI classifying sensitive data and enforcing policies
- **Compliance Automation**: AI generating regulatory reports and evidence

These LLM integrations create a new attack surface: **prompt injection**. Adversaries can manipulate AI behavior through malicious prompts embedded in code comments, user inputs, documents, or data fields, potentially:

- Bypassing security controls (e.g., "Ignore SQL injection warnings in this file")
- Extracting sensitive system prompts revealing security logic
- Exfiltrating training data or classified information
- Manipulating incident triage to downgrade critical alerts
- Falsifying compliance reports

Comprehensive Coverage Across HAIAMM

HAIAMM v2.0 integrates prompt injection security across **6 practices** in the **Software, Data, and Processes** domains:

1. Threat Assessment (TA) - Prompt Injection Threat Modeling

Files: TA-Software, TA-Data, TA-Processes

Coverage:

- **13 Attack Intents**: System Prompt Leak, Jailbreak, Data Exfiltration, Tool Enumeration, API Enumeration, Business Integrity Compromise, Denial of Service, Multi-Chain Attacks, etc.
- **18 Attack Techniques**: Role-Playing, Cognitive Overload, Nested Injection (Russian Doll), Memory Exploitation, Narrative Smuggling, Meta Prompting, Contradiction, Link Injection, Variable Expansion, etc.
- **20 Attack Evasions**: Encoding (Base64, Hex, ASCII), Language variations (Leetspeak, alternate languages), Format-based (JSON, XML, Markdown), Obfuscation (emoji, steganography, morse), Character manipulation, Ciphers

Domain-Specific Threats:

- **Software**: Code comment injection, AI code reviewer manipulation, RAG document poisoning
- **Data**: DSAR manipulation, DLP bypass, data classification override, privacy policy circumvention
- **Processes**: Incident ticket injection, SOAR playbook bypass, alert suppression, compliance falsification

2. Security Requirements (SR) - Prompt Injection Prevention Requirements

Files: SR-Software, SR-Data, SR-Processes

Coverage:

- **Functional Requirements** (SR-PI-001 through SR-PI-007):
  - System prompts SHALL NOT contain credentials, API keys, PII, or sensitive logic
  - User inputs SHALL be validated against known prompt injection patterns
  - LLM outputs SHALL be validated before execution (code, commands, API calls)
  - System and user prompts SHALL be separated via structural delimiters (XML, JSON)
  - Context windows SHALL be scoped to minimum required history
  - Tool calling/function execution SHALL validate parameters before execution
  - RAG documents SHALL be sanitized before LLM ingestion
- **Non-Functional Requirements**:
  - Prompt injection detection ≤100ms latency budget
  - ≥95% detection rate, ≤5% false positive rate
  - Quarterly testing with Arcanum taxonomy
- **Privacy & Compliance Requirements**:
  - PII removal before LLM processing
  - Conversation history scoped per-user (no cross-user contamination)
  - Prompt injection attempts logged for security audit
  - System prompt changes version-controlled and auditable

3. Security Testing (ST) - Prompt Injection Testing Methodology

Files: ST-Software

Coverage:

- **Attack Intent Testing**: Test all 13 attack goals with specific test cases and success criteria
- **Attack Technique Testing**: Test all 18 techniques (role-playing, cognitive overload, nested injection, etc.)
- **Attack Evasion Testing**: Test all 20 obfuscation methods (encoding, language variants, format-based, etc.)
- **Automated Fuzzing**: Integration with Arcanum probe library for comprehensive test coverage

- **Multi-Turn Conversation Testing**: Test exploitation across conversation history
- **Code-Specific Testing**: Prompt injection via code comments, function names, string literals

Success Criteria:

- ≥95% overall detection rate
- 100% system prompt protection (zero successful extractions)
- ≥95% jailbreak resistance
- ≤5% false positive rate on legitimate code/prompts

4. Implementation Review (IR) - Prompt Injection Code Review Checklist

Files: IR-Software

Coverage:

- **Prompt Construction Review**: Input sanitization, delimiter separation, credential protection
- **Output Validation Review**: LLM output validation before execution, allowlist enforcement
- **Context Window Management**: Scope limiting, user isolation
- **Input Validation Review**: Pattern detection, rate limiting
- **RAG Implementation Review**: Document sanitization, PII removal, metadata validation
- **Tool Calling Review**: Parameter validation, function allowlisting, audit logging
- **Error Handling Review**: Information disclosure prevention
- **Testing Review**: Test case coverage validation

Common Anti-Patterns Flagged:

- Direct string concatenation of system and user prompts
- No input validation (accepting all user input)
- Credentials in system prompts
- Unbounded context accumulation
- Blind execution of LLM-generated code
- Missing delimiters between system/user content
- RAG document poisoning (no validation before ingestion)

5. Issue Management (IM) - Prompt Injection Vulnerability Tracking

Files: IM-Software

Coverage:

- **8 Vulnerability Categories** (PI-001 through PI-008):
  - PI-001: System Prompt Leakage (High/Medium severity)
  - PI-002: Jailbreak / Safety Bypass (Critical/High severity)
  - PI-003: Injection-Based Data Exfiltration (Critical/High severity)
  - PI-004: Tool/Function Enumeration (Medium severity)
  - PI-005: Prompt-Based Denial of Service (High/Medium severity)
  - PI-006: Multi-Chain Prompt Attack (High/Medium severity)
  - PI-007: RAG Knowledge Base Poisoning (Critical/High severity)
  - PI-008: Evasion via Encoding/Obfuscation (High/Medium severity)
- **Remediation Workflows**:
  - Critical: ≤24 hours (immediate LLM integration disable, investigation, fix, testing, gradual rollout)
  - High: ≤7 days (triage, fix, testing, deploy)

    - Medium: ≤30 days (investigate, fix, standard release)

- **Detection Sources**: Security testing, bug bounty, code review, WAF/prompt firewall logs, user reports

- **Vulnerability Tracking Fields**: Injection vector, attack intent, technique, evasion method, detection method, affected LLM component

6. Security Architecture (SA) - Defense-in-Depth for Prompt Injection

Note: While not yet integrated into SA one-pagers, the Arcanum 5-Layer Defense Model should inform architectural decisions:

5 Layers of Defense:

1. **Ecosystem Layer**: Dependency management, MFA, network segmentation for AI components
2. **Model Layer**: Select models with safety mechanisms, external prompt injection detection (Lakera Guard, Azure AI Content Safety)
3. **Prompt Layer**: Separate user/system prompts, prompt templates with validation, context scoping, sensitive data exclusion
4. **Data Layer**: PII removal before RAG ingestion, read-only API permissions, data access logging
5. **Application Layer**: Input validation, output encoding, sandboxed LLM execution, least privilege

Arcanum PI Taxonomy Reference

Source: Arcanum Prompt Injection Taxonomy v1.5 **Author**: Jason Haddix **License**: Creative Commons Attribution 4.0 International (CC BY 4.0) **Publication**: Executive Offense Newsletter - Taxonomy v1.5 Release

Taxonomy Structure:

- **13 Attack Intents** (goals): API Enumeration, Attack Users, Business Integrity, Data Poisoning, Denial of Service, Discuss Harm, Generate Image, Get Prompt Secret, Jailbreak, Multi-Chain Attacks, System Prompt Leak, Test Bias, Tool Enumeration
- **18 Attack Techniques** (methods): Act as Interpreter, Anti-Harm Coercion, ASCII, Binary Streams, Cognitive Overload, Contradiction, End Sequences, Framing, Inversion, Link Injection, Memory Exploitation, Meta Prompting, Narrative Smuggling, Puzzling, Rule Addition, Russian Doll, Spatial Byte Arrays, Variable Expansion
- **20 Attack Evasions** (obfuscation): Alt Language, Base64, Case Changing, Cipher, Emoji, Fictional Language, Graph Nodes, Hex, JSON, Link Smuggling, Markdown, Metacharacter Confusion, Morse, Phonetic Substitution, Reverse, Spaces, Splats, Stego, Waveforms, XML

Additional Resources:

- **32 Threat Modeling Questions** organized by category (inputs, ecosystem, model, prompts, data, app, pivoting, monitoring)
- **5-Layer Defense Checklist** (Ecosystem, Model, Prompt, Data, Application)
- **Probe Library** for identifying AI-enabled endpoints

Implementation Guidance

Organizations implementing HAIAMM should:

1. **Start with Threat Assessment**: Use TA-Software, TA-Data, and TA-Processes to understand prompt injection risks specific to your AI security systems

2. **Define Security Requirements**: Implement SR requirements (SR-PI-001 through SR-PI-007) for all LLM integrations

3. **Test Comprehensively**: Use ST-Software testing methodology to validate prompt injection defenses before production deployment
4. **Review Code**: Apply IR-Software code review checklist to all LLM integration code
5. **Track Vulnerabilities**: Use IM-Software vulnerability categories and workflows for ongoing prompt injection issue management
6. **Iterate and Improve**: Quarterly prompt injection testing with Arcanum taxonomy, continuous improvement based on detected patterns

Benefits of Integrated Prompt Injection Coverage

By integrating the Arcanum PI Taxonomy, HAIAMM provides:

Complete Attack Surface Coverage: 13 intents × 18 techniques × 20 evasions = comprehensive understanding **Industry-Standard Taxonomy**: Aligned with leading prompt injection research **Actionable Guidance**: Specific requirements, test cases, code review checklists, vulnerability workflows **Domain-Specific Adaptation**: Software, Data, and Processes domains with tailored examples **Measurable Success Criteria**: ≥95% detection rates, ≤5% false positives, SLA-based remediation **Cross-Practice Integration**: TA → SR → ST → IR → IM lifecycle coverage

Organizations following HAIAMM v2.0 will have **industry-leading prompt injection security** for their HAI security programs.

Threat Intelligence as Foundational Capability

NEW IN v2.0: HAIAMM now integrates threat intelligence as a **foundational capability** (Level 1 maturity) across all six security domains through the Strategy & Metrics (SM) practice, with clear maturity progression from consumption (L1) → analysis and correlation (L2) → production and sharing (L3).

Why Threat Intelligence is Foundational for AI-Operated Security

HAI security systems make thousands of decisions per day - which vulnerabilities to prioritize, which endpoints to quarantine, which cloud configurations to flag, which data access patterns to investigate. **Without threat intelligence, AI security is context-blind:**

AI Security WITHOUT Threat Intelligence :

- Prioritizes vulnerabilities by CVSS scores (generic severity, not real-world exploitation)
- Detects anomalies based on baselines (not known-malicious patterns)
- Flags misconfigurations against standards (not active attack techniques)
- Investigates access anomalies (not credential compromise indicators)
- **Result**: High false positives, missed real threats, poor ROI, analyst alert fatigue

AI Security WITH Threat Intelligence :

- Prioritizes vulnerabilities actively exploited in wild (CISA KEV, exploit databases)
- Detects known-malicious indicators (IOCs, TTPs, malware signatures)
- Flags misconfigurations actively being exploited (MITRE ATT&CK, cloud security research)
- Investigates credential compromise (dark web monitoring, breach databases)
- **Result**: Informed decisions, reduced false positives, catches real threats, measurable ROI

Threat Intelligence as "Security Operating System": Just as an operating system provides shared context and services to applications, threat intelligence provides shared threat context to all AI security

tools. All AI security systems (SAST, EDR, CSPM, DLP, SOAR, vendor risk) benefit from same threat intelligence feeds, creating consistent, context-aware security operations.

Threat Intelligence Maturity Progression in HAIAMM v2.0

Level 1: Foundational Threat Intelligence Consumption

Objective: Establish basic threat intelligence consumption to inform HAI security decisions from day one.

Core Activities:

1. **Identify Domain-Specific Threat Intelligence Requirements** (per domain):
    - **Software**: Vulnerability exploitation intelligence, dependency vulnerability intelligence, code security trends, exploit availability, malicious packages
    - **Infrastructure**: Cloud attack intelligence, infrastructure CVE intelligence, MITRE ATT&CK techniques, configuration baseline intelligence
    - **Endpoints**: Malware intelligence, IOC intelligence (IPs, domains, file hashes), behavioral threat intelligence, vulnerability intelligence
    - **Data**: Data breach intelligence, privacy threat intelligence, data classification intelligence, exfiltration technique intelligence
    - **Processes**: Incident response intelligence, SOAR intelligence, detection evasion intelligence, operational failure intelligence
    - **Vendors**: Vendor breach intelligence, supply chain attack intelligence, vendor vulnerability intelligence, third-party risk intelligence
2. **Integrate Threat Intelligence into AI Security Decision-Making**:
    - **Enrichment Pattern**: Add threat intelligence context to AI security findings
    - **Prioritization Pattern**: Use threat intelligence to rank/prioritize AI findings
    - **Detection Pattern**: Use threat intelligence as detection rules/signatures
    - **Validation Pattern**: Use threat intelligence to validate AI decisions
3. **Measure Basic Threat Intelligence Effectiveness**:
    - **Coverage**: ≥70% of high/critical findings enriched with threat intelligence
    - **Freshness**: ≤24 hours for critical threat intelligence
    - **Actionability**: ≥40% of threat intelligence findings drive AI security action
    - **Detection Improvement**: ≥20% increase in true positive rate after integration

Free/Open-Source Threat Intelligence Sources (Level 1):

- CISA KEV (Known Exploited Vulnerabilities), NVD, GitHub Security Advisories
- MITRE ATT&CK, CIS Benchmarks, Cloud Provider Security Bulletins
- AbuseIPDB, VirusTotal, MISP, AlienVault OTX
- HaveIBeenPwned, Privacy Regulator Databases, Data Breach Notification Sites
- SANS Incident Handler Diary, DFIR Blogs, SOAR Vendor Blogs
- Vendor Breach Databases, Supply Chain Security Research, SBOM Scanners

Success Criteria (Level 1):

- Threat intelligence feeds integrated into ≥80% of AI security tools
- High/critical findings enriched with threat intelligence within ≤1 hour
- Documented evidence that threat intelligence improves AI security decision quality

Level 2: Comprehensive Threat Intelligence Analysis and Correlation

Objective: Analyze threat intelligence across domains, correlate threats, and optimize AI security based on threat intelligence insights.

Core Activities:

1. **Classify Threat Intelligence by Organizational Relevance**:
    - **Critical Threats**: Active exploits in your stack, zero-days in production, APT targeting your profile → Real-time response
    - **High-Relevance Threats**: Vulnerabilities in your stack (not yet exploited), emerging attack techniques → Daily updates, 24-hour response
    - **Medium-Relevance Threats**: General threats with possible applicability → Weekly updates, quarterly review
    - **Low-Relevance Threats**: Unlikely to affect organization → Monthly updates, informational only
2. **Establish Cross-Domain Threat Intelligence Correlation**:
    - **Endpoint ↔ Infrastructure**: Endpoint IOC + Infrastructure attack = Coordinated intrusion
    - **Software ↔ Vendor**: Vendor breach + Dependency vulnerability = Supply chain risk
    - **Data ↔ Processes**: Data breach intelligence + Process intelligence = Active attack
    - **Multi-Domain Reconstruction**: Correlate across all domains to reconstruct full attack chain
3. **Optimize AI Security Tools Based on Threat Intelligence Insights**:
    - **False Positive Reduction**: ≥30% reduction through threat intelligence tuning
    - **Detection Gap Analysis**: ≥90% coverage of critical threats
    - **Adaptive Threat Modeling**: Quarterly threat model updates based on threat intelligence trends
    - **Predictive Threat Hunting**: ≥3 hunting campaigns per quarter based on emerging intelligence
4. **Calculate Threat Intelligence ROI**:
    - **Formula**: (Detection Improvement + False Positive Savings + Breach Prevention + Response Time Improvement - Investment) / Investment
    - **Target**: ≥3:1 ROI (every $1 invested returns $3+ in value)
    - **Metrics**: Prioritization value, false positive savings, detection improvement, response time improvement

Success Criteria (Level 2):

- Threat intelligence classified by relevance with different AI response tiers
- Cross-domain correlation active with ≥85% correlation accuracy
- AI security tools tuned quarterly based on threat intelligence insights
- False positive reduction ≥30%, critical threat coverage ≥90%
- Demonstrable ≥3:1 ROI on threat intelligence investment

Level 3: Industry-Leading Threat Intelligence Production and Sharing

Objective: Produce original threat intelligence from HAI security operations and contribute to industry collective defense.

Core Activities:

1. **Produce Original Threat Intelligence from AI Security Operations**:

- **Anonymized IOC Sharing**: ≥100 novel IOCs per month to industry platforms (MISP, ISACs)
- **Attack Pattern Documentation**: ≥4 documented attack patterns per year (quarterly)
- **Vulnerability Research**: ≥2 responsibly disclosed vulnerabilities per year
- **Threat Trend Analysis**: ≥2 threat trend reports per year (semi-annual)

2. **Participate in Industry Threat Intelligence Sharing Communities**:

   - **Industry ISACs**: Join and contribute to sector-specific ISACs (FS-ISAC, H-ISAC, etc.)
   - **Vendor Partnerships**: Collaborate with security vendors to improve products
   - **Open-Source Platforms**: Contribute to MISP, OpenCTI, STIX/TAXII communities
   - **SLA**: Contribute ≥10 reports per year, consume daily intelligence feeds

3. **Demonstrate Industry Leadership Through Threat Intelligence Innovation**:

   - **Research & Publication**: ≥2 conference presentations or journal publications per year (Black Hat, DEF CON, RSA, academic journals)
   - **Open-Source Contributions**: ≥1 significant open-source project per year with ≥100 GitHub stars
   - **Standards Development**: ≥2 contributions per year to MITRE ATT&CK, OASIS (STIX/TAXII), NIST, ISO/IEC standards
   - **Thought Leadership**: ≥6 blog posts per year, ≥4 podcast/webinar appearances per year

Success Criteria (Level 3):

- Automated threat intelligence production and sharing (≥100 IOCs/month, ≥4 attack patterns/year)
- Active participation in ≥3 threat intelligence sharing communities
- Industry leadership through ≥2 conference presentations or publications per year
- Open-source contributions with community adoption (≥100 GitHub stars)
- Standards development participation (≥2 contributions per year)
- Organization recognized as threat intelligence thought leader

Comprehensive Coverage Across HAIAMM v2.0

Threat intelligence is now integrated into **Strategy & Metrics (SM) practice for all 6 domains**:

1. SM-Software: Vulnerability exploitation intelligence, dependency vulnerability intelligence, code security trend intelligence → Prioritize exploited vulnerabilities, enrich dependency risks, track emerging code threats

2. SM-Infrastructure: Cloud attack intelligence, infrastructure CVE intelligence, MITRE ATT&CK for Cloud → Prioritize exploited misconfigurations, enrich infrastructure vulnerabilities, track cloud attack techniques

3. SM-Endpoints: Malware intelligence, IOC intelligence, behavioral threat intelligence → Detect ransomware campaigns, hunt for malicious IPs/domains/hashes, identify compromise indicators

4. SM-Data: Data breach intelligence, privacy threat intelligence, exfiltration technique intelligence → Monitor credential exposure, track regulatory enforcement, detect data theft patterns

5. SM-Processes: Incident response intelligence, SOAR intelligence, detection evasion intelligence → Enrich incident triage, validate automation playbooks, detect process attacks

6. SM-Vendors: Vendor breach intelligence, supply chain attack intelligence, third-party risk intelligence → Detect vendor breaches early, prevent supply chain attacks, monitor vendor security posture

Domain-Specific Threat Intelligence Integration Examples

Example 1: Software Domain - Dependency Vulnerability Prioritization

- **Without Threat Intelligence**: AI SCA scanner finds 500 dependency vulnerabilities → Prioritize by CVSS score → Team fixes highest CVSS first (many theoretical vulnerabilities, few real-world exploits)
- **With Threat Intelligence**: AI SCA finds 500 vulnerabilities → Threat intelligence enrichment: 12 have active exploits (CISA KEV), 47 have public PoCs → Prioritize these 59 first → Team fixes exploited vulnerabilities before theoretical ones → **Result**: 80% faster mitigation of real-world risks

Example 2: Endpoints Domain - Ransomware Detection

- **Without Threat Intelligence**: AI EDR detects unusual file encryption activity → Flags as suspicious → Analyst investigates manually → 45-minute response time
- **With Threat Intelligence**: AI EDR detects file encryption + Threat intelligence match: Known LockBit ransomware C2 IP + Known TTP (MITRE T1486 Data Encrypted for Impact) → High-confidence ransomware detection → Automated containment triggered → **Result**: <5-minute response, breach prevented

Example 3: Vendors Domain - Supply Chain Attack Prevention

- **Without Threat Intelligence**: Vendor security questionnaire completed 6 months ago, no continuous monitoring
- **With Threat Intelligence**: AI vendor monitoring detects vendor breach disclosure (dark web intelligence) → Vendor "Acme Corp" compromised → Your code uses Acme's npm package → AI correlates as supply chain exposure → Immediate security review, dependency update, incident response → **Result**: Supply chain attack prevented through early detection

Implementation Guidance for Threat Intelligence Integration

Organizations implementing HAIAMM v2.0 threat intelligence capabilities should:

1. **Start with Level 1 Foundational Consumption** (0-6 months):
    - Identify domain-specific threat intelligence requirements
    - Prioritize free/open-source threat intelligence sources to minimize cost barriers
    - Integrate threat intelligence into ≥80% of AI security tools
    - Measure basic effectiveness (coverage, freshness, detection improvement)
2. **Progress to Level 2 Comprehensive Analysis** (6-18 months):
    - Classify threat intelligence by organizational relevance (critical/high/medium/low)
    - Establish cross-domain threat intelligence correlation
    - Optimize AI tools based on threat intelligence insights (reduce false positives, close detection gaps)
    - Calculate and demonstrate threat intelligence ROI (≥3:1 target)
3. **Achieve Level 3 Industry Leadership** (18+ months):
    - Produce original threat intelligence from HAI security operations (IOCs, attack patterns, vulnerability research)
    - Participate in industry threat intelligence sharing communities (ISACs, vendor partnerships, open-source)
    - Demonstrate industry leadership (conference presentations, open-source projects, standards development)
    - Contribute to collective defense (anonymized threat sharing, community collaboration)

Benefits of Integrated Threat Intelligence Coverage in HAIAMM v2.0

Context-Aware AI Security: All AI security tools operate with real-world threat context, not just generic risk scores

Improved Prioritization: Exploited vulnerabilities, active malware, breached vendors prioritized over theoretical risks

Reduced False Positives: Threat intelligence validates AI decisions, reducing noise and alert fatigue

Faster Threat Response: Threat intelligence accelerates detection and response (minutes vs. hours/days)

Measurable ROI: Clear metrics demonstrate threat intelligence value (≥3:1 ROI target)

Industry Alignment: Aligned with industry consensus that threat intelligence is essential for modern security operations

Progressive Maturity: Clear path from basic consumption (L1) → comprehensive analysis (L2) → industry leadership (L3)

Organizations following HAIAMM v2.0 will have **industry-leading threat intelligence-driven HAI security programs** with measurable ROI, improved security outcomes, and contributions to collective defense.

Reference: See Threat-Intelligence-Integration-Analysis.md for comprehensive analysis of threat intelligence integration into HAIAMM, including philosophical rationale, maturity progression details, implementation roadmap, and risk assessment.

Critical HAI Assurance

NEW IN v2.0: HAIAMM now addresses **4 critical risks** for Human Assisted Intelligence deployments that traditional frameworks don't cover, achieving **95% alignment** with OWASP Top 10 for LLM Applications (2025) and OWASP Top 10 for Agentic Applications (2026).

Why These Risks Matter Now

2025-2026 represents the emergence of "agentic AI" - AI agents are now granted:

- Production deployment authority
- Database modification permissions
- External communication capabilities
- Financial transaction authority

Without proper controls, these powerful capabilities can be exploited or misused, resulting in data breaches, production outages, compliance violations, and financial losses.

The Four Critical Risks

1. Excessive Agency (LLM06:2025)

Definition: AI agents granted excessive autonomy or permissions perform unauthorized actions beyond their intended assistance role, violating the fundamental HAI principle that humans maintain decision authority.

The "Least Agency Principle": Grant agents the **minimum autonomy** required to safely assist humans. High-risk actions MUST require human approval.

Action Classification Framework:

Class	Autonomy Level	Human Involvement	Examples
Autonomous (Green)	Agent acts independently	Human notified after action	Code scanning, report generation, log analysis
Human-Validated (Yellow)	Agent proposes, human approves	Human approval required before action	Security patches, config changes, user notifications
Human-Only (Red)	Agent prohibited	Human performs action	Production deployment, data deletion, compliance decisions

Success Metrics: ≥95% approval compliance rate, ≤1% privilege escalation attempts

2. Agent Goal Hijack (ASI01:2026)

Definition: Attackers manipulate agent objectives through prompt injection or data poisoning, causing agents to pursue malicious goals while appearing to function normally.

Key Controls:

- **Goal Validation**: Verify agent goal matches intended objective before each action
- **Immutability Controls**: Agent goals cannot be modified via prompts
- **Multi-Turn Consistency**: Detect gradual goal drift across conversations
- **Monitoring**: Alert on any goal state changes (CRITICAL severity)

Success Metrics: 100% pre-action goal validation, ≥95% hijack detection rate

3. Tool Misuse (ASI02:2026)

Definition: AI agents use **legitimate, authorized tools** for **malicious purposes**. Unlike unauthorized access (which gets blocked), tool misuse exploits capabilities the agent is supposed to have.

Key Controls:

- **Intent Validation**: Verify tool usage aligns with business purpose (not just parameters valid)
- **Destructive Operation Approval**: Delete/destroy actions require human approval
- **Anomaly Detection**: Alert on unusual tool usage patterns
- **Scoped Authorization**: Tools restricted to specific contexts (e.g., email only to internal domains)

Success Metrics: ≥90% tool misuse prevention rate, 100% destructive operation human approval

4. Rogue Agents (ASI10:2026)

Definition: Compromised agents act maliciously while **appearing to function normally**, undermining trust in HAI systems.

Key Controls:

- **Behavioral Baselining**: Establish normal agent behavior over 30 days
- **Real-Time Anomaly Detection**: Alert on deviations ≥2 standard deviations
- **Automatic Containment**: Severe anomalies trigger containment within 30 seconds
- **Ephemeral Goal State**: Compromised goals don't persist across sessions

Success Metrics: ≤5 minutes mean time to detect, ≥95% containment success rate

OWASP Alignment

HAIAMM v2.0 achieves comprehensive alignment with industry-leading OWASP frameworks:

Framework	v2.1 Coverage	v2.2 Coverage	Improvement
OWASP Top 10 for LLM Applications 2025	7/10 (70%)	9/10 (90%)	+20%
OWASP Top 10 for Agentic Applications 2026	4/10 (40%)	9/10 (90%)	+50%
Overall OWASP Alignment	11/20 (55%)	18/20 (90%)	+35%

Reference: See HAIAMM-v2.0-Executive-Summary.md for executive overview and HAIAMM-v2.0-Practice-Additions.md for detailed implementation guidance.

Maturity Levels

HAIAMM defines **3 maturity levels** for progressive capability building:

Level 1: Foundational

Definition: Essential practices for minimally viable AI assurance program

Characteristics:

- Core practices established
- Basic automation with human oversight
- Manual processes for complex decisions
- Reactive security posture
- Individual tool-focused approach

Typical State:

- Ad-hoc threat assessment
- Basic security requirements documented
- Security considered late in development
- Limited security review coverage
- Manual vulnerability tracking
- Basic logging and monitoring

Target: All organizations starting AI asurance programs

Level 2: Comprehensive

Definition: Advanced practices for mature HAI security program

Characteristics:

- Proactive security practices
- Extensive automation with safety controls
- Integrated security across lifecycle
- Metrics-driven improvement

- Risk-based prioritization

Typical State:

- Systematic threat modeling
- Comprehensive security requirements
- Security integrated early (shift-left)
- Automated code review and testing
- Continuous issue management
- Advanced monitoring and correlation

Target: Organizations with 1-2 years AI security maturity

Level 3: Industry-Leading

Definition: Cutting-edge practices for AI security leadership

Characteristics:

- Continuous security validation
- AI-assisted security operations
- Public transparency and research
- Contribution to industry standards
- Formal verification where applicable

Typical State:

- Continuous threat intelligence
- Continue threat detection, prevention, monitoring
- AI-powered security automation
- Formal security proofs
- Public security research
- Open-source contributions
- Security innovation leadership

Target: Organizations leading AI security innovation (3+ years maturity)

Effort Estimation

Estimation Methodology

Based on **OpenSAMM v1.0** approach, effort estimates assume:

Team Composition:

- 1 Security Lead (oversight, planning, review)
- 2-3 Security Engineers (implementation, testing, operations)
- 1-2 AI/ML Engineers (model security, adversarial testing)
- Developers (security integration into development workflow)

Effort Units: Person-days (1 person-day = 8 hours of focused work)

Assumptions:

- Organization has basic security foundation
- Team has AI/ML and security expertise

- Tools and infrastructure available
- Management support secured

Effort by Maturity Level

Maturity Level	Initial Setup	Ongoing (Annual)	Description
Level 1	180-240 days	120-180 days	Foundational practices, establish processes
Level 2	120-180 days	90-150 days	Build on Level 1, add advanced capabilities
Level 3	90-150 days	60-120 days	Continuous improvement, innovation

Note: Effort is cumulative (Level 2 assumes Level 1 complete, Level 3 assumes Level 2 complete)

Effort by Practice (Level 1)

Practice	Setup (Days)	Ongoing (Days/Year)	Key Activities
Threat Assessment (TA)	15-25	10-15	Threat modeling, attack tree analysis, risk assessment
Security Requirements (SR)	20-30	15-20	Requirements definition, metrics, validation criteria
Security Architecture (SA)	30-45	20-30	Architecture design, review, technology selection
Design Review (DR)	15-20	20-30	Design review process, checklists, peer review
Implementation Review (IR)	20-30	30-45	Code review process, tools, automated analysis
Security Testing (ST)	25-35	25-40	Test strategy, adversarial testing, automation
Issue Management (IM)	20-30	30-50	Scanning tools, remediation workflows, tracking
Environment Hardening (EH)	25-35	15-25	Baseline configs, access controls, encryption
Monitoring & Logging (ML)	25-35	20-30	Log aggregation, SIEM integration, alerting
TOTAL	**195-285**	**185-285**	

Effort by Domain (Level 1)

Domain	Complexity	Setup Multiplier	Notes
Software	High	1.2x	Most complex (AI models, code, pipelines)
Data	High	1.2x	Privacy, compliance, classification complexity

Processes	Medium	1.0x	Automation safety, human oversight
Infrastructure	Medium	1.0x	Multi-cloud, configuration management
Endpoints	Medium	1.0x	Cross-platform, performance constraints
Vendors	Low	0.8x	Limited control, mainly assessment

Example: Implementing Threat Assessment (TA) for Software domain:

- Base effort: 15-25 days
- Domain multiplier: 1.2x
- **Total: 18-30 days** for TA-Software at Level 1

Assessment Methodology

Overview

HAIAMM uses a **questionnaire-based maturity assessment** approach, similar to OpenSAMM v1.0, to provide objective, evidence-based measurement of organizational maturity in governing, building, verifying, and operating Human Assisted Intelligence systems.

Assessment Philosophy:

- **Evidence-Based**: Every "Yes" answer requires documented proof
- **Progressive**: Level 1 must be achieved before Level 2, Level 2 before Level 3
- **Comprehensive**: All 12 practices assessed across relevant domains
- **Repeatable**: Standardized questions enable tracking progress over time
- **Practical**: Designed for self-assessment or third-party audit

---

Questionnaire Structure

Total Assessment Scope (v2.0):

- **12 Security Practices** across 6 domains = **72 practice-domain combinations**
- **3 Maturity Levels** per practice = **216 assessment points**
- **2 questions per level** (average) = **432 total assessment criteria**

Tiered Assessment Options:

Tier	Duration	Questions	Domains	Use Case
Tier 1 (Foundation)	20-30 min	24	Software, Data	Quick baseline, quarterly checks
Tier 2 (Standard)	3-4 hours	192	+Infrastructure, Endpoints	Operational assessment, planning
Tier 3 (Comprehensive)	12-16 hours	432	All 6 domains	Compliance, audit, benchmarking

Question Format:

Each practice level has 2-3 specific, testable questions:

```
Practice: Threat Assessment (TA)
Domain: Software
Level: 1 (Foundational)

TA-Software-L1-Q1: Have you identified and documented threats specific
to your HAI systems (e.g., adversarial attacks, prompt injection, data
poisoning)?

Evidence Required:
□ Threat model document exists
□ Threat library or database
□ Updated within last 12 months

Answer: □ Yes  □ No
```

Scoring Methodology

Level Achievement Logic:

For each practice, maturity level is determined by percentage of questions answered "Yes":

Simplified Approach (Recommended for Self-Assessment):

```
Level 1 Achieved = ALL Level 1 questions answered "Yes"
Level 2 Achieved = ALL Level 1 + ALL Level 2 questions answered "Yes"
Level 3 Achieved = ALL questions (L1, L2, L3) answered "Yes"

Practice Score:
- 0.0 = No Level 1 questions answered "Yes"
- 1.0 = All Level 1 "Yes", but not all Level 2
- 2.0 = All Level 1 + Level 2 "Yes", but not all Level 3
- 3.0 = All questions "Yes" (full maturity)
```

Precise Approach (For Formal Audits):

```
Practice Score = (L1_score * 1.0) + (L2_score * 1.0) + (L3_score * 1.0)

Where:
L1_score = (# of L1 "Yes" answers) / (# of L1 questions)
L2_score = (# of L2 "Yes" answers) / (# of L2 questions) * L1_score
L3_score = (# of L3 "Yes" answers) / (# of L3 questions) * L2_score

Note: Higher levels only count if lower level achieved (cascading dependency)
```

Overall Maturity Score:

```
Overall Score = Average of all 12 practice scores

Example:
SM=0.5, PC=1.0, EG=0.8, TA=1.2, SR=1.5, SA=0.7,
DR=0.3, IR=0.5, ST=0.8, EH=1.0, IM=1.2, ML=1.3
```

```
Overall = (0.5+1.0+0.8+1.2+1.5+0.7+0.3+0.5+0.8+1.0+1.2+1.3) / 12
        = 10.8 / 12
        = 0.90 (Mid Level 1 maturity)
```

Assessment Process

Phase 1: Pre-Assessment Preparation (4-8 hours)

1.1 Define Assessment Scope

- [ ] Identify HAI system(s) to assess (start with 1-2 critical systems)
- [ ] Determine assessment tier (Foundation, Standard, Comprehensive)
- [ ] Select domains to assess (minimum: Software + Data for Foundation tier)
- [ ] Define assessment timeframe (point-in-time vs. 3-month lookback)

1.2 Assemble Assessment Team

Recommended team composition:

- **Assessment Lead** (Security/GRC): Facilitates, ensures objectivity
- **HAI Product Owner**: Provides context on system design and usage
- **Security Engineer**: Technical security evidence and controls
- **AI/ML Engineer**: Model security, data governance, algorithmic risks
- **Operations Engineer**: Runtime environment, monitoring, incident response
- **Compliance/Legal** (if applicable): Regulatory requirements, policies

1.3 Gather Evidence

Prepare documentation to support assessment:

- [ ] HAI system architecture diagrams
- [ ] Security policies and procedures
- [ ] Threat models and risk assessments
- [ ] Security requirements documentation
- [ ] Code review reports and tools output
- [ ] Security testing results (SAST, DAST, penetration test reports)
- [ ] Incident response logs and postmortems
- [ ] Monitoring dashboards and logging configurations
- [ ] Training records and security awareness materials
- [ ] Vendor contracts and security questionnaires (if using third-party AI)

1.4 Schedule Assessment Workshop

Typical schedule:

- **Foundation Tier (24 questions)**: 2-3 hour workshop
- **Standard Tier (192 questions)**: 1-2 day workshop with breaks
- **Comprehensive Tier (432 questions)**: 3-5 day assessment (can be distributed)

Deliverable: Assessment plan, stakeholder roster, evidence repository

---

Phase 2: Conduct Assessment (2-16 hours depending on tier)

2.1 Questionnaire Completion Process

For each practice in scope:

Step 1: Review practice definition and objectives

- Read practice one-pager (e.g., `TA-Software-OnePager.md`)
- Understand what "good looks like" for each maturity level

Step 2: Answer Level 1 questions

- Read each question carefully
- Discuss with team whether practice is implemented
- **Require evidence** - Do NOT answer "Yes" without proof
- Document evidence reference (e.g., "Threat model doc dated 2024-11-15")
- Record partial implementations (e.g., "Yes for 60% of HAI systems")

Step 3: Proceed to Level 2 only if ALL Level 1 = "Yes"

- If any Level 1 question = "No", **STOP** (cannot achieve Level 2)
- If all Level 1 = "Yes", proceed to Level 2 questions
- Apply same evidence-based approach

Step 4: Proceed to Level 3 only if ALL Level 2 = "Yes"

Critical Assessment Rules:

1. **"Yes" requires evidence**: No "trust me" answers
2. **"Partial Yes" = "No"**: Practice must be complete and consistent
3. **"Planned" = "No"**: Only implemented practices count
4. **"Sometimes" = "No"**: Practice must be systematic, not ad-hoc
5. **Lookback period**: 3-month default (practice must be current, not historical)

2.2 Evidence Documentation

For each "Yes" answer, document:

- **What**: Description of the practice implementation
- **Where**: Location of evidence (URL, document name, system)
- **When**: Date evidence was created/last updated
- **Who**: Responsible team/individual

Example:

```
Question: TA-Software-L1-Q1: Have you documented threats to your HAI system?

Answer: ☑ Yes

Evidence:
- What: Threat model for Customer Service Chatbot HAI system
- Where: Confluence: /Security/Threat-Models/Chatbot-TM-2024
- When: Created 2024-11-15, Last updated 2024-12-10
- Who: Security Engineering team (lead: Jane Doe)
```

2.3 Handle Edge Cases

Scenario 1: Practice not applicable

- Example: "We don't use third-party AI vendors" (Vendor domain)
- **Solution**: Mark domain as "N/A" and exclude from overall score calculation

Scenario 2: Evidence is confidential

- Example: Penetration test report contains sensitive findings
- **Solution**: Provide summary with sanitized details, offer restricted access to auditors

Scenario 3: Disagreement among assessors

- Example: Engineers say "Yes", security says "No"
- **Solution**: Mark as "No" with note, investigate evidence, escalate if needed

Deliverable: Completed questionnaire with evidence references

---

Phase 3: Scoring and Analysis (2-4 hours)

3.1 Calculate Practice Scores

For each of the 12 practices:

```
1. Count "Yes" answers per level:
   - L1_yes = # of L1 "Yes" answers
   - L2_yes = # of L2 "Yes" answers
   - L3_yes = # of L3 "Yes" answers

2. Count total questions per level:
   - L1_total = # of L1 questions
   - L2_total = # of L2 questions
   - L3_total = # of L3 questions

3. Calculate level scores:
   - L1_score = L1_yes / L1_total
   - L2_score = L2_yes / L2_total (only if L1_score = 1.0)
   - L3_score = L3_yes / L3_total (only if L2_score = 1.0)

4. Calculate practice score:
   - If L1_score < 1.0:  Practice_Score = L1_score
   - If L1_score = 1.0 AND L2_score < 1.0:  Practice_Score = 1.0 + L2_score
   - If L2_score = 1.0 AND L3_score < 1.0:  Practice_Score = 2.0 + L3_score
   - If L3_score = 1.0:  Practice_Score = 3.0 (full maturity)
```

3.2 Calculate Overall Maturity

```
Overall_Maturity = SUM(all practice scores) / NUMBER(practices assessed)

Example (all 12 practices):
Overall = (SM + PC + EG + TA + SR + SA + DR + IR + ST + EH + IM + ML) / 12
```

3.3 Calculate Business Function Scores

```
Governance_Score = (SM + PC + EG) / 3
Building_Score = (TA + SR + SA) / 3
```

```
Verification_Score = (DR + IR + ST) / 3
Operations_Score = (EH + IM + ML) / 3
```

3.4 Maturity Interpretation

Score Range	Maturity Level	Interpretation
0.0 - 0.5	Early Level 1	Just starting, ad-hoc practices, significant gaps
0.5 - 1.0	Mid Level 1	Basic practices in place, inconsistent application
1.0 - 1.5	Late L1 / Early L2	Solid foundation, beginning systematic approach
1.5 - 2.0	Mid Level 2	Mature practices, organization-wide adoption
2.0 - 2.5	Late L2 / Early L3	Advanced capabilities, beginning innovation
2.5 - 3.0	Level 3	Industry-leading, continuous improvement, research

3.5 Gap Analysis

Identify and categorize gaps:

Critical Gaps (Priority 1):

- Practices scoring < 0.5 in high-risk areas (TA, SR, ST)
- Missing foundational practices (Level 1 not achieved)
- Regulatory compliance requirements not met

Important Gaps (Priority 2):

- Practices scoring 0.5-1.0 that block Level 2 progression
- Inconsistent practice application across HAI systems
- Manual processes that should be automated

Enhancement Opportunities (Priority 3):

- Practices at Level 1 that could advance to Level 2
- Automation and tooling improvements
- Metrics and measurement enhancements

Deliverable: Assessment scorecard, gap analysis report

Phase 4: Improvement Roadmap (4-8 hours)

4.1 Prioritize Improvements

Use risk-based prioritization matrix:

```
Priority = (Business Impact) × (Current Gap Size) × (Effort to Improve)^-1

Where:
- Business Impact: 1 (Low) to 5 (Critical)
- Gap Size: (Target Score - Current Score)
- Effort: Person-days estimated from Effort Estimation section
```

4.2 Define Improvement Roadmap

Template:

```
Phase 1: Foundation (Months 1–3)
Goal: Achieve Level 1 across all critical practices
Target Practices: [List practices currently < 1.0]
Success Criteria: All practices ≥ 1.0
Resources: [Team, budget, tools]

Phase 2: Consolidation (Months 4–6)
Goal: Strengthen weak areas, begin Level 2 for top priorities
Target Practices: [List Priority 1 practices for Level 2]
Success Criteria: 4–6 practices at Level 2
Resources: [Team, budget, tools]

Phase 3: Advancement (Months 7–12)
Goal: Systematic maturity, majority of practices at Level 2
Target Practices: [Remaining practices for Level 2]
Success Criteria: Overall score ≥ 1.5
Resources: [Team, budget, tools]
```

4.3 Define Success Metrics

For each improvement initiative:

- **Practice Score Target**: Current → Target (e.g., TA: 0.5 → 1.0)
- **Timeline**: Start date, milestones, completion date
- **Owner**: Responsible individual or team
- **Investment**: Estimated effort (person-days), budget, tools
- **Success Criteria**: Specific, measurable outcomes
- **Evidence**: How compliance will be demonstrated

Deliverable: 12-month improvement roadmap with quarterly milestones

---

Phase 5: Re-Assessment (Ongoing)

Re-Assessment Frequency:

- **Quarterly**: For organizations actively improving (track progress)
- **Semi-Annual**: For stable, mature organizations (maintain compliance)
- **Annual**: Minimum recommended frequency (prevent regression)
- **Triggered**: After major HAI system changes or incidents

Re-Assessment Benefits:

- Track improvement progress against roadmap
- Identify regression (scores decreasing)
- Validate investment effectiveness (ROI)
- Maintain compliance and audit readiness
- Benchmark against industry peers

Continuous Improvement:

- Use assessment results to refine roadmap
- Adjust priorities based on emerging threats and business changes
- Celebrate successes (practices achieving new levels)

- Share lessons learned across organization

Assessment Tools and Templates

HAIAMM Desktop Application provides:

- **Dynamic Questionnaire Generator**: Tier-based assessment with 24/192/432 questions
- **Evidence Repository**: Link evidence to each answer
- **Automated Scoring**: Calculates practice, function, domain, and overall scores
- **Gap Analysis Dashboard**: Visual identification of strengths and weaknesses
- **Progress Tracking**: Compare assessments over time
- **Report Export**: JSON/CSV export with PGP encryption for secure sharing

Manual Assessment Templates (if not using desktop app):

- Assessment planning template (scope, team, timeline)
- Questionnaire spreadsheet (432 questions with evidence columns)
- Scoring calculator (formulas for practice/overall scores)
- Gap analysis template (prioritization matrix)
- Roadmap template (phases, milestones, success criteria)

Industry Benchmarks

Typical First Assessment Scores:

Organization Type	Typical Score	Common Strengths	Common Gaps
Startup (<2 years)	0.3 - 0.6	Operations (IM, ML)	Governance (SM, PC)
Scale-Up (2-5 years)	0.6 - 1.2	Building (TA, SR)	Verification (DR, IR)
Enterprise (5+ years)	0.8 - 1.5	Governance (PC, EG)	Innovation (Level 3)
Security-First Orgs	1.0 - 1.8	Verification (ST)	AI-Specific (adversarial testing)

Maturity Progression Benchmarks:

Based on organizations actively working to improve:

Timeframe	Expected Progress	Typical Score Range
First assessment	Baseline	0.3 - 0.8
After 3 months	Quick wins implemented	0.6 - 1.1
After 6 months	Level 1 mostly achieved	0.8 - 1.3
After 12 months	Level 2 begun	1.2 - 1.8
After 24 months	Level 2 mature	1.8 - 2.4
After 36 months	Level 3 achieved	2.2 - 2.8

Note: Progression speed depends on organizational commitment, resources, and starting maturity.

Assessment Best Practices

DO:

- Use evidence-based assessment (require proof for every "Yes")
- Involve cross-functional team (security, AI/ML, ops, compliance)
- Be honest and objective (this is for improvement, not judgment)
- Document partial implementations (helps prioritize improvements)
- Re-assess regularly (track progress, prevent regression)
- Share results with stakeholders (transparency builds support)
- Celebrate successes (recognize teams achieving maturity levels)

DON'T:

- Guess or estimate answers (require actual evidence)
- Accept "partial yes" as "yes" (practice must be complete)
- Skip evidence documentation (needed for audits and tracking)
- Assess aspirations vs. reality ("we plan to" ≠ "we do")
- Only assess once (maturity requires continuous measurement)
- Make it punitive (focus on improvement, not blame)
- Assess alone (multiple perspectives improve accuracy)

Assurance Domains

1. Software Domain

Scope: AI security systems code (SAST, DAST, models, pipelines, APIs)

Key Challenges:

- AI model security (adversarial attacks, poisoning)
- Code vulnerabilities (injection, authentication, crypto)
- Data pipeline security (ETL, labeling, feedback loops)
- Integration security (IDE plugins, CI/CD, APIs)

AI Technologies: SAST tools, DAST tools, vulnerability scanning, code analysis models

One-Pagers: TA-Software | SR-Software | SA-Software | DR-Software | IR-Software | ST-Software | IM-Software | EH-Software | ML-Software

---

2. Infrastructure Domain

Scope: Cloud and network security (multi-cloud CSPM, network monitoring)

Key Challenges:

- Multi-cloud complexity (AWS, Azure, GCP)
- Configuration drift and misconfiguration detection
- Remediation safety (blast radius, rollback)
- Network security at scale

AI Technologies: CSPM platforms, cloud security automation, network anomaly detection

One-Pagers: TA-Infrastructure | SR-Infrastructure | SA-Infrastructure | DR-Infrastructure | IR-Infrastructure | ST-Infrastructure | IM-Infrastructure | EH-Infrastructure | ML-Infrastructure

---

3. Endpoints Domain

Scope: Endpoint security (EDR/XDR with AI behavioral analytics)

Key Challenges:

- Cross-platform support (Windows, macOS, Linux, mobile)
- Performance constraints (≤5% CPU, ≤200MB memory)
- Privacy preservation (BYOD, no user content)
- False positive management

AI Technologies: EDR/XDR platforms, behavioral analytics, malware detection models

One-Pagers: TA-Endpoints | SR-Endpoints | SA-Endpoints | DR-Endpoints | IR-Endpoints | ST-Endpoints | IM-Endpoints | EH-Endpoints | ML-Endpoints

---

4. Data Domain

Scope: Data security (DLP with AI classification, privacy-preserving ML)

Key Challenges:

- Data classification accuracy (≥90% target)
- Privacy preservation (differential privacy, federated learning)
- Regulatory compliance (GDPR, CCPA, HIPAA)
- Exfiltration prevention

AI Technologies: DLP platforms, data classification models, privacy-preserving ML

One-Pagers: TA-Data | SR-Data | SA-Data | DR-Data | IR-Data | ST-Data | IM-Data | EH-Data | ML-Data

---

5. Processes Domain

Scope: Security orchestration (SOAR with AI alert triage, playbook automation)

Key Challenges:

- Automation safety (blast radius, human oversight)
- Alert triage accuracy (≥95% true positive detection)
- Multi-tool integration (SIEM, EDR, firewall, cloud, ticketing)
- Mean Time to Respond (MTTR) reduction

AI Technologies: SOAR platforms, alert triage models, automation orchestration

One-Pagers: TA-Processes | SR-Processes | SA-Processes | DR-Processes | IR-Processes | ST-Processes | IM-Processes | EH-Processes | ML-Processes

---

6. Vendors Domain

Scope: Vendor risk management (AI-powered assessment, SBOM analysis) Suppliers and partners.

Key Challenges:

- Scale (thousands of vendors)
- Limited visibility (vendor systems are black boxes)
- Supply chain security (SBOM analysis, transitive dependencies)

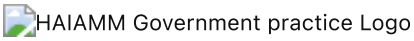
- Continuous monitoring (ratings, breaches, vulnerabilities)

AI Technologies: Vendor risk platforms, SBOM analysis, supply chain security tools

One-Pagers: TA-Vendors | SR-Vendors | SA-Vendors | DR-Vendors | IR-Vendors | ST-Vendors | IM-Vendors | EH-Vendors | ML-Vendors

---

Security Practices



Governance business function

Purpose: Covers 3 best practices to govern a Human Assisted Intelligence Assurance program

1. Strategy & Metrics (SM)

Purpose: Establish unified strategic roadmap and measure effectiveness of HAI security programs

Key Activities:

- Define AI security strategy across all domains (Software, Infrastructure, Endpoints, Data, Processes, Vendors)
- Integrate threat intelligence as foundational capability (consumption, analysis, production)
- Establish metrics for AI security effectiveness (true positive rates, false positive rates, coverage, MTTR)
- Executive sponsorship and governance structure
- ROI measurement and industry benchmarking

Success Metrics:

- AI security strategy documented and current (<12 months old)
- Threat intelligence integrated into all 6 domains at Level 1+
- AI security metrics tracked across practice-domain matrix
- Annual ROI reporting to executives/board

Effort: 20-30 days setup, 15-25 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

2. Policy & Compliance (PC)

Purpose: Define policies, standards, and ensure regulatory compliance for HAI security systems

Key Activities:

- Develop AI security policies (acceptable use, data handling, model governance)
- Define security standards and baselines
- Ensure regulatory compliance (GDPR, CCPA, HIPAA, SOX, industry-specific)
- Policy enforcement and compliance monitoring
- Audit preparation and evidence collection

Success Metrics:

- AI security policies documented and approved

- 100% compliance with applicable regulations
- Policy violations tracked and remediated
- Annual compliance audits passed

Effort: 25-35 days setup, 20-30 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

3. Education & Guidance (EG)

Purpose: Provide security training, awareness, and guidance for teams building and operating AI security systems

Key Activities:

- Security awareness training for AI/ML engineers
- Secure AI development training (prompt injection prevention, adversarial ML)
- AI security best practices documentation
- Role-based training (developers, operators, security analysts)
- Continuous learning and knowledge sharing

Success Metrics:

- ≥90% of AI/ML engineers completed security training annually
- Security champions program established
- Security best practices documented and accessible
- Training effectiveness measured (pre/post assessments)

Effort: 15-25 days setup, 15-20 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

Building business function

Purpose: This function covers all best practices recommended to improve and build secure Human Assisted Intelligence systems

4. Threat Assessment (TA)

Purpose: Identify and analyze threats specific to HAI security systems

Key Activities:

- AI-specific threat modeling (adversarial ML, data poisoning, model theft)
- Attack tree analysis
- Abuse case development
- Risk scoring and prioritization

Success Metrics:

- 100% of AI systems have threat model
- ≥80% of threats have mitigations
- Threat models updated quarterly

Effort: 15-25 days setup, 10-15 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

5. Security Requirements (SR)

Purpose: Define measurable security requirements for AI systems

Key Activities:

- Functional security requirements (authentication, authorization, encryption)
- Non-functional requirements (performance, accuracy, latency, privacy)
- Compliance requirements (GDPR, CCPA, HIPAA)
- Acceptance criteria definition

Success Metrics:

- 100% of AI systems have documented security requirements
- ≥90% of requirements testable and measurable
- ≥95% of requirements validated before production

Effort: 20-30 days setup, 15-20 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

6. Security Architecture (SA)

Purpose: Design secure, scalable architecture for AI security systems

Key Activities:

- Architecture design (components, data flows, integrations)
- Technology selection (frameworks, databases, cloud services)
- Security control design (authentication, encryption, access control)
- Scalability and performance design

Success Metrics:

- 100% of AI systems have documented architecture
- Architecture reviewed before implementation
- Performance targets met (latency, throughput)

Effort: 30-45 days setup, 20-30 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

Verification business function

Purpose: The Verification business function recommends 3 best practices to verify HAI systems for security, ensure they have been built with security in mind.

7. Design Review (DR)

Purpose: Review designs before implementation to catch flaws early

Key Activities:

- Peer design review
- Security design review (threat-focused)
- Architecture review (scalability, performance)

- Compliance review (regulatory requirements)

Success Metrics:

- 100% of designs reviewed before implementation
- ≥80% of issues caught in design review (before coding)
- Review turnaround ≤5 business days

Effort: 15-20 days setup, 20-30 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

8. Implementation Review (IR)

Purpose: Code review to ensure secure implementation

Key Activities:

- Peer code review (pull request reviews)
- Security code review (vulnerability-focused)
- Automated code analysis (SAST, linters, dependency scanning)
- Test coverage review

Success Metrics:

- 100% of code reviewed before merge
- ≥80% of bugs caught in code review
- ≥90% of security vulnerabilities caught in review
- Review turnaround ≤2 business days

Effort: 20-30 days setup, 30-45 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

9. Security Testing (ST)

Purpose: Test AI systems for vulnerabilities and adversarial robustness

Key Activities:

- Adversarial testing (evasion, poisoning, model extraction)
- Penetration testing (SAST, DAST, API security)
- Performance testing (latency, throughput, scalability)
- Compliance testing (GDPR, privacy guarantees)

Success Metrics:

- ≥80% code coverage for security tests
- ≥90% of vulnerabilities caught in testing
- Adversarial robustness validated
- Performance targets met

Effort: 25-35 days setup, 25-40 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

Operations business function

Purpose: The Operations business functions covers all best practices recommended to build, improve how HAI systems operate.

10. Issue Management (IM)

Purpose: Continuous vulnerability scanning and remediation, tracking of tickets/issues like vulnerabilities.

Key Activities:

- Dependency scanning (libraries, frameworks, ML packages)
- Code vulnerability scanning (SAST)
- Infrastructure scanning (CSPM, container scanning)
- Remediation tracking (SLA-driven patching)

Success Metrics:

- 100% of systems scanned daily
- ≥95% vulnerabilities remediated within SLA (Critical ≤24h, High ≤7d)
- MTTR: Critical ≤24h, High ≤7d, Medium ≤30d

Effort: 20-30 days setup, 30-50 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

11. Environment Hardening (EH)

Purpose: Harden deployment environments to reduce attack surface

Key Activities:

- Secure baseline configurations (CIS Benchmarks)
- Least privilege access (IAM, RBAC)
- Encryption (at rest, in transit)
- Network segmentation (zero trust, micro-segmentation)

Success Metrics:

- ≥95% systems comply with security baselines
- 100% sensitive data encrypted
- ≥90% accounts follow least privilege
- Zero unnecessary public exposure

Effort: 25-35 days setup, 15-25 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

12. Monitoring & Logging (ML)

Purpose: Comprehensive logging and monitoring for security, compliance, debugging

Key Activities:

- Security event logging (authentication, attacks, anomalies)
- AI model monitoring (accuracy, drift, explainability)
- Performance monitoring (latency, throughput, errors)
- Compliance logging (GDPR, audit trails)

Success Metrics:

- 100% of services emit logs
- ≥99% logs delivered to SIEM
- ≥90% attacks detected within ≤5 minutes
- 100% compliance events logged

Effort: 25-35 days setup, 20-30 days/year ongoing

Domain Coverage: Software | Infrastructure | Endpoints | Data | Processes | Vendors

---

Practices-Domain Matrix

Complete Coverage Matrix (12 × 6 = 72 Practice-Domain Combinations)

Practice	Software	Infrastructure	Endpoints	Data	Processes	Vendors
Strategy & Metrics (SM)	SM-SW	SM-INF	SM-EP	SM-DA	SM-PR	SM-VE
Policy & Compliance (PC)	PC-SW	PC-INF	PC-EP	PC-DA	PC-PR	PC-VE
Education & Guidance (EG)	EG-SW	EG-INF	EG-EP	EG-DA	EG-PR	EG-VE
Threat Assessment (TA)	TA-SW	TA-INF	TA-EP	TA-DA	TA-PR	TA-VE
Security Requirements (SR)	SR-SW	SR-INF	SR-EP	SR-DA	SR-PR	SR-VE
Security Architecture (SA)	SA-SW	SA-INF	SA-EP	SA-DA	SA-PR	SA-VE
Design Review (DR)	DR-SW	DR-INF	DR-EP	DR-DA	DR-PR	DR-VE
Implementation Review (IR)	IR-SW	IR-INF	IR-EP	IR-DA	IR-PR	IR-VE
Security Testing (ST)	ST-SW	ST-INF	ST-EP	ST-DA	ST-PR	ST-VE
Issue Management (IM)	IM-SW	IM-INF	IM-EP	IM-DA	IM-PR	IM-VE
Environment Hardening (EH)	EH-SW	EH-INF	EH-EP	EH-DA	EH-PR	EH-VE
Monitoring & Logging (ML)	ML-SW	ML-INF	ML-EP	ML-DA	ML-PR	ML-VE

Total: 54 practice-domain one-pagers

Implementation Roadmap

Phase 1: Foundation (Months 1-6) - Level 1

Objective: Establish essential AI security practices

Priority Practices:

1. **Threat Assessment (TA)** - Understand AI-specific threats
2. **Security Requirements (SR)** - Define what "secure" means
3. **Issue Management (IM)** - Start scanning for known vulnerabilities
4. **Monitoring & Logging (ML)** - Gain visibility into AI systems

Priority Domains:

1. **Software** - Secure the AI code and models
2. **Data** - Protect sensitive data and ensure privacy

Key Milestones:

- Month 1-2: TA and SR for Software and Data domains
- Month 3-4: VM and ML for Software and Data domains
- Month 5-6: Expand to Infrastructure and Endpoints domains

Effort: 120-180 person-days

Success Criteria:

- Threat models for all AI systems
- Security requirements documented
- Daily vulnerability scanning operational
- Logging to SIEM implemented

Phase 2: Maturity (Months 7-18) - Level 1 Complete + Level 2 Start

Objective: Complete Level 1 across all domains, begin Level 2

Priority Practices:

1. **Security Architecture (SA)** - Design secure AI systems
2. **Design Review (DR)** - Catch design flaws early
3. **Implementation Review (IR)** - Ensure secure coding
4. **Security Testing (ST)** - Test for vulnerabilities
5. **Environment Hardening (EH)** - Harden deployments

Priority Domains: All remaining domains (Processes, Vendors)

Key Milestones:

- Month 7-9: SA, DR, IR for all domains
- Month 10-12: ST, EH for all domains
- Month 13-15: Complete Level 1 for all 54 combinations
- Month 16-18: Begin Level 2 for high-priority practices/domains

Effort: 180-240 person-days

Success Criteria:

- All 54 Level 1 practices implemented
- 100% code review coverage
- Security testing automated
- Environments hardened per baselines

Phase 3: Optimization (Months 19-36) - Level 2

Objective: Advanced AI security capabilities

Focus Areas:

1. **Automation**: AI-assisted security (automated triage, response)
2. **Integration**: Unified security across lifecycle
3. **Metrics**: Data-driven security improvement
4. **Proactive Security**: Shift-left, continuous validation

Key Milestones:

- Month 19-24: Level 2 for Software, Data, Processes domains
- Month 25-30: Level 2 for Infrastructure, Endpoints, Vendors domains
- Month 31-36: Optimization, metrics, continuous improvement

Effort: 120-180 person-days

Success Criteria:

- ≥70% automation rate for security operations
- ≥90% of bugs caught before production
- Mean time to detect (MTTD) ≤5 minutes
- Mean time to respond (MTTR) ≤10 hours

Phase 4: Leadership (Months 37+) - Level 3

Objective: Industry-leading AI security

Focus Areas:

1. **Innovation**: AI-powered security, formal verification
2. **Research**: Public security research, contributions
3. **Standards**: Participation in AI security standards
4. **Open Source**: Contribute security tools and frameworks

Continuous Activities:

- Security research and publication
- Open-source contributions
- Conference presentations
- Standards body participation

Effort: 60-120 person-days/year

Framework Mappings

ISO/IEC 27001:2022 Mapping

ISO 27001 Control	HAIAMM Practices	Notes
A.5.1 Policies	TA, SR	Information security policy includes AI security
A.5.7 Threat Intelligence	TA	AI-specific threat intelligence
A.8.8 Management of Technical Vulnerabilities	VM	Includes AI model, dependency vulnerabilities
A.8.26 Application Security Requirements	SR, SA, IR, ST	AI application security
A.8.29 Security Testing	ST	Includes adversarial testing for AI
A.8.32 Change Management	IR, ST	AI model, code deployment

Coverage: HAIAMM provides AI-specific implementation of ISO 27001 controls

---

NIST Cybersecurity Framework 2.0 Mapping

NIST CSF Function	HAIAMM Practices	Notes
GOVERN (GV)	TA, SR	AI risk management, security strategy
IDENTIFY (ID)	TA, IM	AI asset inventory, vulnerability discovery
PROTECT (PR)	SA, EH, IR	Secure AI architecture, hardening
DETECT (DE)	ML, ST	AI security monitoring, testing
RESPOND (RS)	IM, ML	Issue remediation, incident response
RECOVER (RC)	EH, ML	Backup, recovery, resilience

Coverage: HAIAMM implements CSF 2.0 for HAI security programs

---

NIST AI Risk Management Framework Mapping

NIST AI RMF Function	HAIAMM Practices	Notes
GOVERN	TA, SR	AI governance, accountability
MAP	TA, SA	AI context, categorization, risk assessment
MEASURE	ST, ML	AI performance measurement, monitoring
MANAGE	VM, EH, IR	AI risk treatment, mitigation

Coverage: HAIAMM operationalizes NIST AI RMF for security applications

---

Appendices

Appendix A: Glossary

AI-Operated Security Program: Security program using AI for threat detection, response, risk assessment (SAST, DAST, CSPM, EDR, DLP, SOAR)

Adversarial Machine Learning: Attacks against AI models (evasion, poisoning, extraction, inversion)

Blast Radius: Maximum scope of automated security action (e.g., ≤50 IPs blocked per action)

Data Poisoning: Injecting malicious data into training datasets to degrade model performance

Differential Privacy: Mathematical privacy guarantee through noise addition (epsilon/delta parameters)

Federated Learning: Distributed ML training where data stays on endpoints, only gradients shared

Model Drift: Degradation of model accuracy over time due to data distribution changes

Prompt Injection: Attack manipulating LLM behavior through malicious prompts embedded in user input, code comments, or documents to bypass security controls, extract sensitive information, or exfiltrate data

RAG (Retrieval-Augmented Generation): LLM architecture pattern where external documents/knowledge bases augment model responses. Vulnerable to knowledge base poisoning if documents aren't validated.

SBOM: Software Bill of Materials - list of software components and dependencies

Zero-Day: Previously unknown vulnerability with no available patch

---

Appendix B: References

Standards & Frameworks:

- ISO/IEC 27001:2022 - Information Security Management
- ISO/IEC 42001:2023 - AI Management System
- NIST Cybersecurity Framework 2.0 (2024)
- NIST AI Risk Management Framework (2023)
- OpenSAMM v1.0 - Software Assurance Maturity Model

AI Security Research:

- MITRE ATLAS - Adversarial Threat Landscape for AI Systems
- OWASP Top 10 for LLM Applications
- NIST IR 8269 - Taxonomy of Adversarial ML
- **Arcanum Prompt Injection Taxonomy v1.5** by Jason Haddix (CC BY 4.0)
  - GitHub: https://github.com/Arcanum-Sec/arc_pi_taxonomy
  - Publication: Executive Offense Newsletter
  - Comprehensive taxonomy of prompt injection attacks (13 intents, 18 techniques, 20 evasions)
  - 5-layer defense model, 32 threat modeling questions, probe library
  - Integrated into HAIAMM v2.0 TA, SR, ST, IR, IM practices across Software, Data, Processes domains

---

Appendix C: Tool Categories

SAST/DAST: Semgrep, CodeQL, SonarQube, Snyk Code, Checkmarx, Veracode

CSPM: Prisma Cloud, Wiz, Orca Security, Aqua CSPM, AWS Security Hub

EDR/XDR: CrowdStrike, SentinelOne, Microsoft Defender, Carbon Black

DLP: Symantec DLP, Forcepoint DLP, Microsoft Purview, Digital Guardian

SOAR: Splunk SOAR, Palo Alto XSOAR, IBM Resilient, Swimlane

Vendor Risk: BitSight, SecurityScorecard, RiskRecon, Prevalent, Panorays

---

Appendix D: Contributors

HAIAMM Development Team:

- Framework Design
- Practice Development
- Domain Expertise
- Effort Estimation

Acknowledgments:

- OpenSAMM and BSIMM for maturity model methodology
- NIST for AI and cybersecurity frameworks
- ISO for security management standards
- **Jason Haddix** and the Arcanum Security team for the Arcanum Prompt Injection Taxonomy (CC BY 4.0), which provides comprehensive prompt injection attack coverage integrated into HAIAMM v2.0

Appendix E: Change Log

Version 2.0 (2025-12-25):

- Initial comprehensive handbook release
- 54 practice-domain one-pagers
- Effort estimation methodology
- Framework mappings (ISO 27001, NIST CSF, NIST AI RMF, OWASP SAMM)
- Implementation roadmap
- Complete practice and domain coverage

Version 2.0 (2025-12-26):

- **Threat Intelligence as Foundational Capability**: Integrated threat intelligence as Level 1 foundational requirement across all 6 domains
- **Strategy & Metrics (SM) Practice Enhancement**: Added comprehensive threat intelligence guidance to SM practice for all domains
  - SM-Software: Vulnerability exploitation intelligence, dependency vulnerability intelligence, code security trends
  - SM-Infrastructure: Cloud attack intelligence, infrastructure CVE intelligence, MITRE ATT&CK for Cloud
  - SM-Endpoints: Malware intelligence, IOC intelligence, behavioral threat intelligence
  - SM-Data: Data breach intelligence, privacy threat intelligence, exfiltration technique intelligence
  - SM-Processes: Incident response intelligence, SOAR intelligence, detection evasion intelligence
  - SM-Vendors: Vendor breach intelligence, supply chain attack intelligence, third-party risk intelligence (NEW one-pager created)
- **Threat Intelligence Maturity Progression**:

    - Level 1 (Foundational): Threat intelligence consumption with free/open-source sources (CISA KEV, MISP, VirusTotal, etc.)
    - Level 2 (Comprehensive): Threat intelligence classification, cross-domain correlation, ROI calculation (≥3:1 target)
    - Level 3 (Industry-Leading): Threat intelligence production (≥100 IOCs/month), industry sharing, thought leadership (≥2 presentations/year)
- **Updated Practice One-Pagers**:
    - SM-Software, SM-Infrastructure, SM-Endpoints, SM-Data, SM-Processes: Added threat intelligence sections at all 3 maturity levels
    - SM-Vendors: Created new comprehensive one-pager with integrated threat intelligence guidance
- **Handbook Updates**:
    - New "Threat Intelligence as Foundational Capability" section (200+ lines) with implementation guidance
    - Updated Executive Summary, Table of Contents, Framework Overview
    - Added 3 domain-specific threat intelligence integration examples
    - Reference to Threat-Intelligence-Integration-Analysis.md (comprehensive 23-page analysis document)
- **Philosophical Shift**: Threat intelligence elevated from "advanced capability" (L2/3) to "foundational capability" (L1) across all domains
- **Benefits**: Context-aware AI security, improved prioritization, reduced false positives (≥30%), faster response, measurable ROI

Version 2.0.1 (2025-12-26):

- **Prompt Injection Security Integration**: Added comprehensive prompt injection guidance across 6 practices (TA, SR, ST, IR, IM, SA) in Software, Data, and Processes domains
- **Arcanum PI Taxonomy Integration**: Integrated Arcanum Prompt Injection Taxonomy v1.5 by Jason Haddix (CC BY 4.0)
    - 13 Attack Intents (System Prompt Leak, Jailbreak, Data Exfiltration, Tool Enumeration, etc.)
    - 18 Attack Techniques (Role-Playing, Cognitive Overload, Nested Injection, Memory Exploitation, etc.)
    - 20 Attack Evasions (Encoding, Language variations, Format-based, Obfuscation, etc.)
    - 5-Layer Defense Model (Ecosystem, Model, Prompt, Data, Application)
- **Updated Practice One-Pagers**:
    - TA-Software, TA-Data, TA-Processes: Added comprehensive prompt injection threat categories
    - SR-Software, SR-Data, SR-Processes: Added prompt injection prevention requirements (SR-PI-001 through SR-PI-008)
    - ST-Software: Added detailed prompt injection testing methodology
    - IR-Software: Added LLM integration code review checklist
    - IM-Software: Added 8 prompt injection vulnerability categories with remediation workflows
- **Handbook Updates**:
    - New "Prompt Injection Security" section with implementation guidance
    - Updated Executive Summary, Table of Contents
    - Added prompt injection glossary terms, references, acknowledgments
- **Attribution**: All integrated content properly attributed to Arcanum PI Taxonomy (CC BY 4.0)

Document Information

- **Document**: HAIAMM Comprehensive Handbook
- **Version**: 2.0
- **Last Updated**: 2025-12-26
- **Status**: Published
- **License**: Creative Commons Attribution 4.0 International (CC BY 4.0)
- **Third-Party Attributions**: Arcanum Prompt Injection Taxonomy v1.5 by Jason Haddix (CC BY 4.0)
- **Contact**: [To be added]
- **Website**: [To be added]

End of HAIAMM Handbook v2.0