ADAPTIVE AI RISK & ASSURANCE FRAMEWORK

AAIRAF™

Adaptive AI Risk & Assurance Framework (AAIRAF) for Cloud and Embedded Systems

Version 5.3

Author: Brian “SchleiF” Schleifer, MTSI

Contributors: Brenda Ivey, Credence Management Solutions Inc.

Technical Writer: Latasha T. Kelly, MTSI

AI Usage Disclosure:

AI Usage Disclosure: This document was created with assistance from AI tools and has been reviewed and edited by a human. Please note that AI-generated content may contain errors or inaccuracies. Users should exercise caution and conduct their own verification before relying on this information.

CONTENTS

ABBREVIATIONS AND ACRONYMS

Acronym	Definition
AAIRAF	Adaptive AI Risk & Assurance Framework
ABAC	Attribute-Based Access Control
AI	Artificial Intelligence
AI-BOM	Artificial Intelligence - Bill of Materials
AIRT	AI Incident Response Team
AI-SCM	AI Security Continuous Monitoring
AI-SCRM	AI Supply Chain Risk Management
API	Application Programming Interface
ATLAS	Adversarial Threat Landscape for Artificial Intelligence Systems
ATO	Authority to Operate
BOM	Bill of Material
cATO	Continuous Authority to Operate
CCPA	California Consumer Privacy Act
CDAO	Chief Digital and AI Office
CISO	Chief Information Security Officer
CMMC	Cybersecurity Maturity Model Certification
CSF	Cybersecurity Framework (NIST)
DAI	Detect AI Incidents
DHS	Department of Homeland Security
DIU	Defense Innovation Unit
DL	Deep Learning
DoD	Department of Defense
DoS	Denial of Service
EMI	Electromagnetic Interference
EU AI Act	European Artificial Intelligence Act
FDA	Food & Drug Administration
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GIG	Global Information Grid
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act
HITL	Human-in-the-loop
HMI	Human-Machine Interface
HOTL	Human-on-the-loop
HSM	Hardware Security Modules
IAM	Identify and Access Management
ICS	Industrial Control System
IGR	Identify and Govern AI Risk
ISMS	Information Security Management System
ISSM	Information System Security Manager
JSIG	Joint Special Access Program Implementation Guide
KPI	Key Performance Indicator
KRI	Key Risk Indicator
LLM	Large Language Model

M-BOM	Manufacturing - Bill of Materials
ML	Machine Learning
MLOps	Machine Learning Operations
NDAA	National Defense Authorization Act
NSG	Network Security Group
OT	Operational Technology
OTA	Over the Air
OWASP	Open Web Application Security Project
PAA	Protect AI Assets
PETs	Privacy-Enhancing Technologies
PLC	Programmable Logic Controller
RAG	Retrieval-Augmented Generation
RAI	Respond to AI Incidents
RBAC	Role-Based Access Control
REA	Recover and Evolve AI System
RMF	Risk Management Framework
RPA	Robotics Process Automation
S-BOM	Software - Bill of Materials
SCA	Security Control Assessor
SCADA	Supervisory Control and Data Acquisition
SIEM	Security Information and Event Management
TTP	Tactics, Techniques, and Procedures
USAF	United States Air Force
XAI	Explainable Artificial Intelligence

1 INTRODUCTION AND PURPOSE

1.1 Executive Summary

The increasing integration of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and emerging paradigms like Generative AI (GenAI) and Agentic AI into critical systems, such as embedded systems, critical infrastructure, and healthcare devices, introduces novel and complex cybersecurity, operational, and ethical risks. Traditional cybersecurity frameworks, while foundational, often lack the specificity to address vulnerabilities unique to AI models, data pipelines, and their operational environments. The Adaptive AI Risk & Assurance Framework (AAIRAF) is designed to bridge this gap, providing a structured, comprehensive, and adaptable approach to identifying, protecting, detecting, responding to, and recovering from AI-specific threats. AAIRAF enables organizations, particularly with critical systems and within national security contexts, to harness the transformative potential of AI securely and responsibly, ensuring trust, resilience, and compliance throughout the AI lifecycle while safeguarding national security and upholding ethical principles.

1.2 Scope and Applicability

AAIRAF applies to all organizations involved in the design, development, deployment, operation, and retirement of AI systems. For this framework, "AI Systems" encompass a broad range of technologies, including, but not limited to:

Audience

The AAIRAF framework serves multiple levels of the organization, with three primary target audiences:

1. Organizational Leadership – Executives, CISOs, CIOs, program/mission owners, and legal/ethics committees. They use AAIRAF to establish AI governance, define enterprise risk posture, and make strategic adoption decisions.
2. Assessors, auditors, ISSMs, SCAs, and independent reviewers. They apply the AAIRAF Assessment Plan and Checklist to validate risks, controls, and evidence, linking findings to the Traceability Matrix.
3. Implementors & Requestors – AI model requestors (business/mission owners) and implementors (developers, MLOps engineers, data stewards). They operationalize AAIRAF by producing required artifacts such as Model Cards, Bill of Materials (BOM), telemetry, lineage reports, and by embedding controls in the AI lifecycle.

AI Scope

1. Machine Learning (ML) Models: Supervised, unsupervised, reinforcement learning, deep learning (e.g., neural networks, transformers, large language models - LLMs).
2. Expert Systems & Knowledge-Based Systems: Rule-based AI.
3. Robotics Process Automation (RPA) with AI components.
4. Autonomous Systems: Vehicles, drones, industrial control systems, and especially weapon systems with AI decision-making.

The framework serves as the anchor document, covering the entire AI lifecycle, from initial concept and data acquisition through model training, validation, deployment, continuous monitoring, and eventual decommissioning. AAIRAF is applicable across diverse sectors (e.g., government, defense, financial services, healthcare, manufacturing, critical infrastructure) and can be tailored to the specific risk profile and operational context of an organization, distinguishing between general AI applications and those deemed "high-risk AI" (e.g., per EU AI Act criteria, or critical embedded systems). Sections of the AAIRAF explicitly address deployments in cloud environments and resource-constrained embedded systems, such as critical infrastructure and healthcare devices. In addition to the framework, a Risk Assessment Plan and Checklist utilize AAIRAF as the foundation, and a Traceability Matrix that traces other frameworks, security controls, and more to the corresponding AAIRAF risks.

1.3 Guiding Principles

AAIRAF is built upon the following foundational principles:

Human-Centric & Ethical AI: Prioritizing human well-being, safety, fairness, transparency, and accountability in the design and operation of AI systems. This includes mitigating algorithmic bias, ensuring human oversight, and maintaining human control, especially in lethal autonomous weapon systems, critical infrastructure, and healthcare devices.

Risk-Based Approach: Allocating resources and implementing controls commensurate with the likelihood and impact of AI-specific risks, ensuring the most critical assets and potential harms are addressed first.

Adaptability & Continuous Improvement: Recognizing the dynamic nature of AI technology, evolving threat landscapes, and emerging vulnerabilities. The framework promotes iterative assessment, learning, and adaptation.

Transparency & Explainability: Fostering understanding of AI decisions, limitations, and processes to build trust and enable effective governance and incident response, particularly crucial for auditable and accountable embedded systems.

Security-by-Design & Privacy-by-Design: Embedding cybersecurity and privacy considerations into every stage of the AI lifecycle, from initial conceptualization to deployment and beyond.

Collaboration & Information Sharing: Encouraging the exchange of threat intelligence, best practices, and lessons learned across organizations, industry sectors, and research communities, including defense alliances.

Resilience & Mission Assurance: Designing AI systems to withstand and recover from adverse events, including cyberattacks, data corruption, and system failures, to ensure continuity of critical military, infrastructure, and healthcare operations.

1.4 Relationship to other Frameworks and Policy Alignment

AAIRAF is designed as an overlay and specialization that complements and integrates with existing frameworks, rather than replacing them. It provides AI-specific depth and context to the broader cybersecurity and risk management landscape, with a strong emphasis on national and global security directives.

NIST Policy:

- NIST Cybersecurity Framework (CSF): AAIRAF adopts the CSF's core functions (Identify, Protect, Detect, Respond, Recover) as its high-level structure, enriching each with AI-specific categories and subcategories. This ensures a familiar and interoperable structure for security professionals.
- NIST Risk Management Framework (RMF): AAIRAF aligns with the RMF's seven steps (Prepare, Categorize, Select, Implement, Assess, Authorize, Monitor) by applying them specifically to AI systems. It guides the categorization of AI systems based on risk, selection of AI-specific controls, continuous monitoring of AI performance and security posture, and the authorization process for AI system deployment, including the DoD Authorization to Operate (ATO).
- NIST AI Risk Management Framework (AI RMF 1.0): AAIRAF heavily leverages and integrates the NIST AI RMF's four functions (Govern, Map, Measure, Manage) within its own structure. "Govern" and "Map" are central to AAIRAF's "Identify & Govern AI Risk" function; "Measure" informs "Detect AI Incidents" and "Recover & Evolve AI Systems"; and "Manage" is embedded across all protective and responsive functions. AAIRAF operationalizes the NIST AI RMF's principles within a cybersecurity context.
- DoD RMF / CMMC: For defense and government contexts, AAIRAF provides the necessary AI-specific controls and assessment criteria to ensure compliance with DoD RMF and CMMC requirements, particularly concerning supply chain security for AI components and data. It supports the rigorous testing and evaluation requirements for AI components in weapon systems, focusing on preventing adversarial attacks, ensuring reliability, safety, and maintaining human control.
- ISO/IEC 27001 / 27002 / 42001: AAIRAF aligns with the principles of an Information Security Management System (ISMS) by extending the scope of information security controls to cover AI-specific assets (models, training data, inference data) and processes (MLOps pipelines). It also incorporates principles from the emerging ISO/IEC 42001 (AI Management System).
- MITRE ATT&CK / ATLAS: AAIRAF incorporates the adversarial tactics and techniques identified in ATT&CK and ATLAS to inform threat modeling, vulnerability assessments, and the development of AI-specific detection and response strategies.
- OWASP Top 10 / OWASP Top 10 for LLMs: These lists directly inform the identification of common AI vulnerabilities and guide the implementation of protective measures during AI development and deployment.
- Privacy Frameworks (e.g., GDPR, CCPA, HIPAA): AAIRAF explicitly integrates privacy-by-design principles throughout the AI lifecycle, particularly in data management, model training, and explainability, ensuring compliance with relevant data protection regulations.

US National Policy & Directives:

- Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (October 2023): Broadly directs the US government to address AI safety and security risks.
- AI Safety Institute (under NIST): Focuses on AI safety and security testing and evaluation.
- National AI Initiative: Promotes AI research and development, including security aspects.
- Department of Homeland Security (DHS) AI Strategy: Outlines DHS's approach to using and securing AI technologies.

DoD AI Policy & Strategy:

- DoD AI Strategy: Focuses on accelerating the adoption of AI while ensuring its responsible and ethical use, including security considerations.
- NDAA (National Defense Authorization Act): Often includes provisions related to AI, such as requirements for testing and evaluation of AI systems used in defense applications.

Key U.S. Department of Defense AI Initiatives:

- Defense Innovation Unit (DIU)
  Collaborates with commercial AI companies to accelerate the adoption of innovative AI solutions across the Department of Defense, with an emphasis on security and reliability.
- Chief Digital and AI Office (CDAO)
  Provides leadership for DoD-wide AI strategy, ensuring responsible development, integration, and deployment of artificial intelligence capabilities.
- DoD AI Ethical Principles
  Establish a foundation for AI use within defense, emphasizing systems that are responsible, equitable, traceable, reliable, and governable.

2 CORE FUNCTIONS AND CATEGORIES

AAIRAF adopts the NIST CSF's five core functions, deeply enriching each with AI-specific considerations. Each function contains categories and subcategories that detail specific outcomes and activities.

2.1 Identify and Govern AI Risk (IGR)

Purpose: Develop an organizational understanding to manage AI-related cybersecurity, operational, and ethical risk to systems, data, assets, and capabilities. This function heavily integrates NIST AI RMF's "Govern" and "Map" functions.

2.1.1 AI Asset Management

2.1.1.1 Inventory of AI Systems and Components

Maintain an up-to-date inventory of all AI models, datasets, algorithms, libraries, frameworks, MLOps pipelines, and supporting infrastructure (e.g., GPUs, cloud services, embedded hardware). For weapon systems, this includes targeting algorithms, navigation systems, and

control interfaces. For industrial control and operational technology (ICS/OT), the inventory must capture PLCs, SCADA integrations, and protocol-specific dependencies (Modbus, DNP3, OPC UA). For healthcare AI-enabled devices, include diagnostic imaging models, infusion pumps, and implantable systems subject to FDA and IEC 62304/80001 guidance. For IoT environments, include device firmware, OTA update pipelines, and large-scale fleet configurations subject to NISTIR 8259/8425.

2.1.1.2 AI System Criticality and Impact Classification

Classify AI systems based on their potential impact (e.g., safety, financial, privacy, ethical, operational, mission-criticality for defense applications) and criticality to organizational missions. Differentiate between "high-risk AI" and lower-risk applications.

2.1.1.3 AI Data Lineage and Provenance

Document the origin, collection methods, transformations, and usage of all data throughout the AI lifecycle to ensure trustworthiness and support forensic analysis, especially for sensor data in embedded systems.

2.1.1.4 AI Dependency Mapping

AI Dependency Mapping: Identify and document internal and external dependencies for each AI system, including third-party models, Application Programming Interface (API)s, data sources, and software libraries (e.g., PyTorch, TensorFlow), considering supply chain risks.

2.1.2 AI Governance and Risk Management Strategy

Program Governance Enhancements:

- Runtime Assurance Controllers: Deterministic fallback (e.g., Simplex architecture) for mission/safety-critical contexts.
- Supply Chain Assurance: AI-BOM, S-BOM, and M-BOM required to trace all models, datasets, and code dependencies.
- Defense-Specific Step 0 (Prepare): Explicit scoping, categorization, and supply chain mapping.
- Continuous Authorization (cATO): Authorization decisions must be tied to live telemetry (model confidence, drift, bias, anomaly signals).

2.1.2.1 AI Risk Management Policies

Establish formal policies for managing AI-specific cybersecurity, privacy, ethical, and operational risks, integrated with enterprise and mission risk management.

2.1.2.2 AI Roles and Responsibilities

Define clear roles, responsibilities, and accountability for AI risk management (e.g., Chief AI Officer, AI Ethics Committee, MLSecOps team, Data Stewards, military chain of command for AI decisions).

2.1.2.3 AI Risk Tolerance and Acceptance Criteria

Define the organization's acceptable level of risk for different categories of AI systems, considering potential harms, benefits, and mission requirements. For embedded systems—

including those in medical devices and industrial control systems—this involves establishing acceptable thresholds for accuracy, reliability, autonomy, patient safety, and clinical effectiveness in alignment with FDA guidance and IEC 62304. It also requires ensuring data integrity and cybersecurity in accordance with HIPAA and ISO/IEC 27001, as well as maintaining operational continuity and resilience against cascading failures in line with NIST SP 800-82 and ISA/IEC 62443 for ICS environments.

2.1.2.4 AI Supply Chain Risk Management (AI-SCRM) Strategy

Develop a robust strategy for managing risks associated with third-party AI models, open-source components, data providers, and AI development tools, aligned with CMMC requirements.

Overlay and Tailoring Integration
Overlay and tailoring artifacts, including the *NIST Securing-AI Overlays* and the *AI-Cyber Tailoring Guide*, shall be developed, reviewed, and stored in conjunction with the AI-BOM, S-BOM, and M-BOM. This ensures that context-specific control selection and tailoring decisions are traceable and auditable within the framework, supporting consistent evidence generation across systems and environments.

OWASP LLM Top 10 Alignment
The AAIRAF aligns its generative and LLM-specific risk coverage with the OWASP Top 10 for Large Language Models to strengthen control design and assurance testing. The mapping below provides assessors with direct linkages between OWASP categories and AAIRAF sections:

- **LLM01 Prompt Injection** → 2.1.3.4 (AI Red Teaming & Adversarial Testing) and 2.2.* (input validation, tool/plugin sandboxing).
- **LLM02 Insecure Output Handling** → 2.2.* (content filtering, output binding and type-safety), 2.4.* (runtime policy enforcement).
- **LLM03 Training Data Poisoning** → 2.2.* (data lineage/curation, signed datasets, poisoning detection).
- **LLM04 Model Denial of Service** → 2.3.* (capacity protection, rate-limiting) and 2.4.* (telemetry-driven throttling).
- **LLM05 Supply Chain Vulnerabilities** → 2.2.* (AI-BOM/S-BOM/M-BOM, provenance and signature verification).
- **LLM06 Sensitive Information Disclosure** → 2.2.* (PII redaction, retrieval scoping), 2.4.* (DLP at inference).
- **LLM07 Insecure Plugin/Tooling** → 2.2.* (least-privilege tool adapters, capability scoping).
- **LLM08 Excessive Agency** → 2.2.* (guardrails, constrained decoding, policy-aware tools).
- **LLM09 Overreliance** → 2.3.* (HITL/HOTL checkpoints, confidence-aware UX).
- **LLM10 Model Theft** → 2.2.* (model watermarking/fingerprinting, API abuse prevention).

Reference list: see 6.2.

2.1.3 AI Threat and Vulnerability Assessment

Embedded and sector-specific risks should be assessed alongside general AI vulnerabilities. Weapon systems face threats to targeting algorithms and navigation AI. ICS/OT environments must consider protocol fuzzing, AI-driven command injection, and unsafe state induction. Healthcare systems should account for adversarial manipulation of medical imaging or treatment models and tampering with life-critical devices. IoT ecosystems must address botnet recruitment, weak default credentials, firmware tampering, and large-scale compromise scenarios.

2A. Canonical Risk Categories and Areas

A. Data-Related Risks

- Data Poisoning & Backdoors
- Label/Feature Integrity & Schema Drift
- Data Leakage & Privacy (membership inference, inversion, re-identification)
- Dataset Bias & Representativeness

B. Model-Related Risks

- Model Poisoning/Manipulation
- Model Stealing/Extraction
- Adversarial Evasion
- Hallucination & Reliability
- Emergent Bias & Drift

C. Generative/LLM-Specific Risks

- Prompt Injection & Jailbreaks
- Retrieval-Augmented Generation (RAG) Poisoning
- Toxic/Unsafe Content
- Output Leakage

D. Infrastructure & Operational Risks

- Availability/Denial-of-Service
- Supply Chain Compromise
- Misconfiguration & Access Control
- Telemetry & Logging Gaps

E. Embedded System Risks

- Sensor/Data Spoofing & EMI/EMC
- Communications Jamming & Interception
- Safety & Human-Machine Interface (HMI)
- Physical Tampering & Side-Channels

Assessment Requirement: Each risk must be linked to mapped controls, adversarial test cases (MITRE ATLAS), runtime telemetry, and evidence provided in the Traceability Matrix.

2.1.3.1 AI-Specific Threat Identification

Identify and document AI-specific threat actors and their tactics, techniques, and procedures (TTPs) leveraging MITRE ATLAS. This includes:

Data-Related: Data Poisoning (Backdoor, Logic Corruption, Influence attacks), Data Integrity Issues (Missing Values, Incorrect Labels, Duplicate Entries, Data Entry Errors), Data Drift, Data Bias (Sampling, Measurement, Confirmation), Data Leakage (Membership Inference, Model Inversion).

Model-Related: Model Stealing, Model Manipulation, Adversarial Attacks (Evasion, Poisoning at Inference), Lack of Robustness/Stability, Overfitting/Underfitting, Model Bias (emergent), Model Degradation, Unintended Functionality.

Infrastructure & Operational: Denial of Service (DoS) attacks, API Security vulnerabilities, Infrastructure Vulnerabilities (servers, databases, pipelines), Software Supply Chain Vulnerabilities (libraries, dependencies), Misconfiguration and Access Control Issues.

Embedded System Specific: Compromised Sensors, Data Spoofing, GPS jamming, communication interception, control system compromise.

2.1.3.2 Adversarial Risk Profiling

Conduct tailored risk assessments that consider the potential for adversarial attacks against AI systems, including the attacker's capabilities, motivations, and access, particularly for high-stakes embedded systems (weapon, ICS, IoT, healthcare).

2.1.3.3 AI Pipeline Vulnerability Assessment

Regularly assess vulnerabilities within the entire MLOps pipeline, including data ingestion, feature engineering, model training, deployment, and inference environments, for both cloud and embedded contexts.

2.1.3.4 AI Red Teaming and Adversarial Testing

Conduct specialized red teaming exercises and adversarial robustness testing against deployed or pre-production AI systems to identify weaknesses.

2.1.4 AI Regulatory & Ethical Compliance

2.1.4.1 AI Regulatory Landscape Monitoring

Continuously monitor and interpret evolving AI-specific regulations (e.g., EU AI Act, proposed US AI legislation, NDAA provisions) and industry standards.

2.1.4.2 Algorithmic Bias and Fairness Assessment

Implement methodologies (e.g., fairness metrics, counterfactual explanations) to assess and mitigate unintended biases in AI models and their outputs, particularly for systems impacting personnel or critical decisions.

2.1.4.3 Data Privacy and Ethical Data Usage

Ensure compliance with data privacy regulations (e.g., GDPR, CCPA, HIPAA) for all data used in AI systems, including ethical considerations for data collection and consent.

2.1.4.4 Human Oversight and Accountability

Establish robust mechanisms for human review, intervention, and ultimate accountability for decisions made by or supported by AI systems, especially for embedded systems and critical infrastructure, aligned with AI Ethical Principles (e.g UNESCO, OECD, and/or DoD).

2.2 Protect AI Assets

Purpose: Develop and implement appropriate safeguards to ensure the delivery of critical AI services.

2.2.1 Secure AI Data Management

2.2.1.1 Secure Data Collection and Ingestion

Implement secure protocols for data acquisition, including secure APIs, encrypted channels, and validation to prevent data poisoning at ingestion. For embedded systems, secure sensor data streams.

2.2.1.2 Data Anonymization and Privacy-Enhancing Technologies

Apply techniques such as differential privacy, homomorphic encryption, and secure multi-party computation to protect sensitive data during training and inference, especially for cloud-based data.

2.2.1.3 Data Integrity and Validation

Implement robust data validation, checksums, and cryptographic hashing to ensure the integrity of training, validation, and production data. Address issues like missing values, incorrect labels, and data entry errors.

2.2.1.4 Secure Data Storage and Access Control

Apply strict access controls (e.g., least privilege, role-based access control) and encryption at rest for all AI-related datasets, in cloud storage or embedded memory.

2.2.2 Secure AI Model Development and Deployment

Embedded and sector-specific development practices include: IoT device development must incorporate constraints of lightweight devices, with Over-The-Air (OTA) updates cryptographically signed and integrity-checked despite limited resources. ICS/OT deployment must require signed firmware, deterministic runtime assurance, and security overlays for industrial protocols (Modbus, DNP3, OPC UA). Healthcare AI deployments must validate model integrity in life-critical systems, aligning with FDA pre- and post-market guidance.

2.2.2.1 Secure MLOps Pipeline

Implement security controls across the entire MLOps pipeline, including secure CI/CD practices, vulnerability scanning of container images, and automated security testing, for both cloud and embedded deployment.

2.2.2.2 Adversarial Robustness Techniques

Incorporate techniques such as adversarial training, input sanitization, feature squeezing, and defensive distillation to enhance models' resilience to adversarial attacks and increase their robustness against minor input variations.

2.2.2.3 Model Integrity Verification

Implement cryptographic signing and hashing for AI models to verify their authenticity and detect unauthorized modifications during storage and deployment (e.g., secure boot for embedded systems).

2.2.2.4 Secure Model Serving and APIs

Secure model inference endpoints with authentication, authorization, rate limiting, input validation, and secure API gateway practices for cloud-based AI. For embedded systems, secure internal communication and Human Machine Interface (HMI).

Note: Baseline secure configs for AI stacks against applicable overlays; document deltas in the Traceability Matrix.

2.2.2.5 Secure Configuration Management

Maintain secure configurations for all AI development and production environments, including code, dependencies, and infrastructure, addressing misconfiguration and access control issues.

2.2.3 AI Access Control and Authorization

2.2.3.1 AI Environment Access Control

Implement strong authentication and granular authorization (e.g., RBAC, ABAC) for access to AI development environments, data lakes, model registries, and inference services, both in cloud and on-device.

2.2.3.2 Secure Human-in-the-Loop Interfaces

Ensure that interfaces for human oversight and intervention (e.g., in weapon systems, medical devices, IoT, etc.) are secure, validated, and resistant to manipulation.

2.2.4 AI System Resilience and Redundancy

2.2.4.1 Model and Data Backup/Recovery

Implement regular backups of AI models, training data, and metadata, along with tested recovery procedures. For embedded systems, consider limited storage and rapid recovery.

2.2.4.2 High Availability for Critical AI Services

Design and deploy critical AI systems with redundancy and failover mechanisms to ensure continuous operation, especially for embedded systems with real-time performance requirements.

2.2.4.3 Contingency Planning

Develop and test contingency plans for AI system failures, data corruption, or successful cyberattacks, including scenarios for resource-constrained environments.

2.2.5 AI Explainability & Interpretability Controls

2.2.5.1 Explainability Technique Implementation

Integrate explainable AI (XAI) techniques (e.g., Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive exPlanations (SHAP), feature importance) where appropriate to provide insights into model decisions, aiding in debugging, security analysis, and building trust. Address the "black box" nature.

2.2.5.2 Model Documentation and Lineage

Maintain comprehensive documentation of model architecture, training parameters, data sources, and version history to support auditing and understanding.

2.2.5.3 Audit Trails for AI Decisions

Implement robust logging and audit trails for AI system inputs, outputs, and internal decision-making processes to support post-incident analysis and accountability, especially for high-consequence decisions.

2.3 Detect AI Incidents

Purpose: Develop and implement appropriate activities to identify the occurrence of an AI cybersecurity event.

2.3.1 AI Anomaly and Threat Detection

2.3.1.1 Adversarial Attack Detection

Implement monitoring for indicators of adversarial attacks, such as unusual input patterns, sudden drops in model confidence, unexpected model outputs, or high rates of misclassification for specific input types. This includes detection of sensor malfunctions or data spoofing.

2.3.1.2 Data and Concept Drift Detection

Monitor for significant shifts in input data distributions (data drift) or changes in the relationship between input and output variables (concept drift), which can indicate data poisoning or environmental changes impacting model performance.

2.3.1.3 Model Tampering and Integrity Monitoring

Continuously verify the integrity of deployed models using cryptographic hashes or behavioral baselines to detect unauthorized modifications.

2.3.1.4 AI System Log and Telemetry Analysis

Collect and analyze logs from AI applications, infrastructure, and MLOps pipelines for suspicious activities, failed access attempts, or unusual resource consumption.

2.3.2 AI Performance and Bias Monitoring

2.3.2.1 Continuous Model Performance Monitoring

Monitor key performance metrics (e.g., accuracy, precision, recall, F1-score) of deployed AI models in real-time to detect degradation indicative of attacks or data issues, addressing "Inaccurate Performance" and "Limited Generalization."

2.3.2.2 Algorithmic Bias Monitoring

Continuously monitor for shifts in fairness metrics (e.g., demographic parity, equalized odds) across different sensitive attributes to detect emerging biases.

2.3.2.3 Output Validation and Sanity Checks

Implement automated checks on AI system outputs to identify anomalous or nonsensical predictions that could indicate compromise or malfunction.

2.3.3 AI Security Continuous Monitoring

2.3.3.1 Vulnerability Scanning of AI Dependencies

Regularly scan AI-specific libraries, frameworks, and underlying infrastructure for known vulnerabilities.

2.3.3.2 AI Threat Intelligence Integration

Subscribe to and integrate AI-specific threat intelligence feeds (e.g., new adversarial attack techniques, vulnerabilities in popular AI frameworks) into security operations.

2.3.3.3 Secure Configuration Auditing

Continuously audit AI development and production environments against established secure configuration baselines.

2.4 Respond to AI Incidents

Purpose: Develop and implement appropriate activities to act regarding a detected AI cybersecurity incident.

2.4.1 AI Incident Response Planning

Incident response planning must also address embedded and sector-specific contexts. For ICS/OT, unsafe physical states caused by AI-driven control failures require integration with OT safety recovery plans. Healthcare devices necessitate FDA-coordinated vulnerability disclosure and rapid device recall/remediation to maintain patient safety. IoT fleets demand fleet-scale containment strategies, including certificate revocation and emergency patch rollout to counter botnet propagation.

2.4.1.1 AI-Specific Incident Response Playbooks

Develop and regularly test playbooks for common AI incidents (e.g., data poisoning, model evasion, prompt injection, model theft, bias incidents, sensor spoofing, model manipulation in embedded systems).

2.4.1.2 AI Incident Response Team

Establish a dedicated or cross-functional team with expertise in AI, cybersecurity, legal, and communications for incident response, including military operational personnel.

2.4.1.3 Communication Protocols for AI Incidents

Define clear internal and external communication plans for AI-related incidents, including regulatory reporting requirements and military command structures.

2.4.2 AI Incident Containment & Eradication

2.4.2.1 Model Quarantine and Isolation

Implement procedures to quickly quarantine or isolate compromised AI models or data pipelines to prevent further damage. For embedded systems, this may involve physical isolation or secure shutdown.

2.4.2.2 Data Source Isolation and Cleaning

Isolate suspicious data sources and implement procedures for identifying and removing poisoned or malicious data from training and production datasets.

2.4.2.3 Model Rollback and Versioning

Utilize model versioning systems to enable rapid rollback to a known good, uncompromised model state. For embedded systems, this requires robust (OTA) or secure physical update mechanisms.

2.4.3 AI Incident Analysis & Forensics

2.4.3.1 AI-Specific Forensic Data Collection

Collect and preserve forensic artifacts unique to AI incidents (e.g., compromised model weights, adversarial examples, training logs, inference requests, input/output pairs), including data from embedded system sensors and internal logs.

2.4.3.2 Root Cause Analysis for AI Incidents

Conduct thorough investigations to determine the root cause of AI incidents, including identifying attack vectors, vulnerabilities exploited, and data provenance issues.

2.4.3.3 Explainability for Anomalous Behavior

Utilize XAI techniques to understand why an AI system behaved anomalously during an incident, aiding in diagnosis and remediation.

2.4.4 AI Remediation & Recovery Coordination

2.4.4.1 Model Retraining and Validation

Implement procedures for retraining compromised models with clean, validated data and rigorous re-validation before re-deployment.

2.4.4.2 AI System Patching and Hardening

Apply patches, update configurations, and implement additional hardening measures to address vulnerabilities exploited during an incident. For embedded systems, this includes secure firmware updates.

2.4.4.3 Secure Re-deployment

Ensure that re-deployment of AI systems follows secure MLOps practices and includes post-incident verification.

2.5 Recover and Evolve AI Systems

Purpose: Develop and implement appropriate activities to restore any capabilities or services that were impaired due to an AI cybersecurity incident, and to continuously improve the AI security posture.

2.5.1 AI Recovery Planning and Implementation

2.5.1.1 Full AI System Restoration

Execute comprehensive plans to restore all affected AI systems, data, and infrastructure to full operational capability, prioritizing mission-critical functions for embedded systems.

2.5.1.2 Validation of Recovered AI Assets

Rigorously validate the integrity, performance, and security of all recovered AI models and datasets before returning them to production, including real-world testing for embedded systems.

2.5.1.3 Data Integrity Restoration

Implement processes to ensure the integrity and trustworthiness of all data used by AI systems post-recovery.

2.5.2 AI Communications

2.5.2.1 Stakeholder Communication

Communicate recovery status and lessons learned to relevant internal and external stakeholders, including military command and allied forces.

2.5.2.2 Transparency and Disclosure (as required)

Adhere to legal and ethical obligations for transparency and public disclosure regarding AI incidents, particularly those affecting user safety, privacy, or fairness, or national security.

2.5.3 AI Post-Incident Review & Improvement

2.5.3.1 AI Incident Lessons Learned

Conduct thorough post-incident reviews to identify root causes, effectiveness of response, and areas for improvement in AI security controls and processes.

2.5.3.2 Framework and Control Updates

Update AI risk assessments, security policies, and control implementations based on lessons learned from incidents and emerging threat intelligence.

2.5.3.3 Continuous AI Security Training

Provide ongoing training and awareness programs for AI development, MLOps, and security teams on new AI threats, vulnerabilities, and mitigation strategies.

3 CROSS-CUTTING CAPABILITIES AND MANAGEMENT

These are overarching elements that apply across all core functions and are crucial for the successful implementation and maintenance of AAIRAF.

3.1 AI Security Program Management

3.1.1 AI Security Leadership and Governance

Establish clear leadership for AI security (e.g., a dedicated AI Security Lead or team, integrated with Chief Information Security Officer) and integrate AI security governance into the broader enterprise and mission security program.

3.1.2 Resource Allocation

Ensure adequate financial, personnel, and technological resources are allocated to support AI security initiatives.

3.1.3 AI Security Policy and Procedure Development

Develop, disseminate, and enforce specific policies and procedures governing secure AI development, deployment, and operation, aligned with military standards.

3.2 AI Supply Chain Risk Management (AI-SCRM)

3.2.1 Third-Party AI Assessment

Conduct due diligence and continuous monitoring of third-party AI models, data providers, open-source components, and AI service vendors for security and ethical risks, aligned with Global, National, Governmental, and CMMC supply chain requirements.

3.2.2 Contractual Security Requirements

Include stringent AI-specific security, privacy, and incident response requirements in contracts with AI-related third parties.

3.2.3 Component Provenance and Integrity

Verify the provenance and integrity of all AI components (datasets, pre-trained models, libraries) acquired from external sources.

3.3 AI Security Training and Awareness

3.3.1 Secure AI Development Training

Provide specialized training for AI/ML engineers, data scientists, and MLOps teams on secure coding practices for AI, adversarial robustness, privacy-preserving AI, and secure prompt engineering.

3.3.2 AI Risk Awareness

Develop awareness programs for all employees on the risks associated with AI systems, social engineering targeting AI, and the importance of ethical AI use in all applicable contexts.

3.4 AI Security Measurement and Metrics

3.4.1 Key Performance Indicators (KPIs) for AI Security

Define and track KPIs related to the effectiveness of AI security controls (e.g., number of vulnerabilities remediated, time to detect adversarial attacks, robustness score improvements, system uptime for critical AI).

3.4.2 Key Risk Indicators (KRIs) for AI Risks

Establish KRIs for AI-specific risks (e.g., data drift magnitude, bias detection frequency, adversarial attack attempts, model degradation rates).

3.4.3 AI Security Reporting

Implement regular reporting mechanisms to communicate the AI security posture and risk landscape to leadership, military command, and relevant stakeholders.

3.5 Legal, Regulatory, and Ethical Alignment

3.5.1 AI Ethics Committee/Review Board

Establish or leverage an existing committee to review AI projects for ethical implications, bias, and compliance with internal, national, and international guidelines (e.g., DoD AI Ethical Principles).

3.5.2 Legal Counsel Engagement

Involve legal counsel in the review of AI systems, especially those dealing with sensitive data, high-risk applications (like weapon systems, healthcare devices, ICS, etc.), or potential for significant societal or military impact.

3.6 Human-AI Teaming and Accountability

3.6.1 Defined Human-in-the-Loop Processes

Clearly define where and how human oversight, intervention, and override mechanisms are integrated into AI workflows, particularly for critical decisions in embedded systems, ensuring human control.

3.6.2 Accountability Framework

Establish a clear accountability framework for decisions and outcomes involving AI systems, ensuring that human responsibility is maintained, and clear lines of authority exist.

3.6.3 Decision Logging & Auditability

Ensure that AI systems log decisions, human interventions, and relevant context to enable post-hoc analysis and auditing, critical for post-mission review and accountability.

4 Implementation Guidance and Assessment Methodology

Assessment Methodology Enhancements

Traceability Rule: Every identified risk must trace to (a) mapped control(s), (b) at least one adversarial test case (MITRE ATLAS-aligned), (c) runtime telemetry evidence, and (d) outcome/impact metrics.

Lifecycle Checkpoints (capAI/EU AI Act): Document conformity checkpoints across design, development, evaluation, operation, and retirement.

This section provides practical guidance for adopting AAIRAF and details the structured assessment methodology for AI systems, integrating the NIST RMF steps.

4.1 Getting Started

4.1.1 Executive Buy-in & Sponsorship

Secure strong leadership support (e.g., C-Suite, organizational command, etc.) to drive the adoption and integration of AAIRAF.

4.1.2 Pilot Program

Begin with a pilot implementation on a non-critical AI system or a specific phase of the AI lifecycle to gain experience and demonstrate value.

4.1.3 Gap Analysis

Conduct an initial assessment of existing cybersecurity practices against AAIRAF to identify gaps and prioritize initial efforts.

4.1.4 Phased Implementation

Adopt a phased approach, focusing on high-impact areas first, such as AI asset inventory and basic protection controls for critical systems.

4.2 Maturity Model

Organizations can assess their current AI security posture using the following maturity levels:

Level 1: Initial/Ad Hoc: AI security practices are informal, reactive, and inconsistent. Limited awareness of AI-specific risks.

Level 2: Developing: Some AI security practices are defined, but implementation is inconsistent. Basic inventory and some ad-hoc protections.

Level 3: Defined: AI security policies and procedures are documented and partially implemented across relevant AI systems. Roles and responsibilities are clearer.

Level 4: Managed: AI security controls are systematically implemented, measured, and monitored. Performance and risk metrics are tracked, and incident response plans are tested.

Integration Note: Use NIST Securing-AI Overlays and the AI Cybersecurity RMF Tailoring Guide as normative tailoring inputs when deriving AAIRAF-aligned controls and assessment methods. Maintain pointers to these in the Traceability Matrix.

Level 5: Optimized: AI security is integrated into the organizational culture, continuously improved through lessons learned, threat intelligence, and proactive research. AI security is a strategic enabler.

4.3 Integration with Existing Frameworks

AAIRAF is designed for seamless integration. Organizations should:

Map Controls: Cross-reference AAIRAF categories and subcategories with existing controls from NIST CSF, NIST RMF, ISO 27002, or CMMC. For example, a control like "2.2.1.3 Data Integrity & Validation" in AAIRAF maps directly to "PR.DS-1: Data at rest is protected" in NIST CSF, but with AI-specific validation techniques.

Leverage Existing Processes: Integrate AI risk assessments into existing enterprise risk management processes. Incorporate AI-specific incident response playbooks into the overall incident response plan.

Utilize Existing Tools: Extend the use of current security tools (e.g., SIEM, vulnerability scanners) to cover AI-specific logs, environments, and dependencies where possible, and invest in specialized AI security tools as needed.

4.4 Tailoring and Customization

AAIRAF should be tailored to an organization's unique context, especially concerning deployment environments.

Risk Profile: Organizations dealing with "high-risk AI" (e.g., autonomous weapon systems, medical AI) will need to implement a more stringent set of controls than those using AI for internal administrative tasks.

Industry Sector: Specific industry regulations (e.g., healthcare data privacy in HIPAA, financial sector regulations) will dictate additional requirements. For defense, adherence to DoD directives is paramount.

AI Use Cases: The specific type of AI (e.g., LLMs versus traditional ML) will influence the most relevant threats and protective measures (e.g., prompt injection for LLMs).

Organizational Size and Resources: Smaller organizations may need to prioritize the most critical controls and leverage cloud provider security features for AI services.

4.5 AI System Assessment Methodology (NIST RMF Steps Applied to AI)

This framework provides a structured approach to assessing the security of AI systems in both cloud and embedded environments, aligned with the NIST RMF.

4.5.1 Step 1: Categorize the System (NIST SP 800-37 Step 1)

Purpose: Define the mission and operational context of the AI component within the system (e.g., weapon system, medical device, etc.). Determine the potential impact if the AI system were compromised.

Detailed Actions:

Mission Definition: Clearly articulate the system's mission (e.g., air defense, ground attack, surveillance). Document the specific role of the AI (e.g., target recognition, autonomous navigation, threat assessment).

Impact Analysis (NIST AI RMF - Govern Function): Conduct a preliminary impact analysis considering safety, mission, financial, reputational, legal/regulatory consequences of a breach. For healthcare, critical infrastructure, and weapon systems, safety and mission impacts are paramount.

Security Categorization (NIST SP 800-60): Assign a security category (High, Moderate, Low) based on the potential impact of a security breach on confidentiality, integrity, and availability. For critical embedded systems, "High" is often the default.

4.5.2 Step 2: Select Security Controls (NIST SP 800-37 Step 2, NIST SP 800-53)

Purpose: Choose a baseline set of security controls (from AAIRAF Section 2) appropriate for the system's security category and specific threats. Reference appropriate organization guidance if differs from traditional RMF guidance.

Detailed Actions:

Baseline Control Selection: Select a baseline set of security controls from AAIRAF Section 2 and NIST SP 800-53. Prioritize controls relevant to embedded systems and AI.

Tailoring: Tailor controls to the specific requirements of the system, considering embedded system constraints (resource limitations, real-time requirements, physical security), AI-specific threats (ATLAS), and system architecture.

Control Enhancements: Add security control enhancements to address specific risks (e.g., Hardware Security Modules (HSMs), secure boot, code signing, memory protection, intrusion detection for embedded systems).

Document Control Selection: Document selected controls, tailoring decisions, and enhancements.

4.5.3 Step 3: Implement Security Controls (NIST SP 800-37 Step 3)

Purpose: Implement the selected security controls in the system and its AI components.

Detailed Actions:

Develop Implementation Plan: Create a detailed plan outlining steps to implement each control.

Configure Security Settings: Configure security settings on all components (hardware, software, data).

Integrate Security Tools: Integrate tools like intrusion detection systems and vulnerability scanners.

Develop Secure Coding Practices: Use secure coding for AI system's code.

Document Implementation: Document configuration settings, integration details, and secure coding practices.

4.5.4 Step 4: Assess Security Controls (NIST SP 800-37 Step 4, NIST AI RMF - Measure and Monitor Function)

Purpose: Determine if implemented controls are operating as intended and are effective in mitigating risks. This is critical for AI.

Detailed Actions:

Develop Assessment Plan: Outline scope, objectives, and methodology.

Conduct Vulnerability Scanning and Penetration Testing: Identify known vulnerabilities and simulate real-world attacks.

Conduct Adversarial Testing (NIST AI RMF - Govern Function): Specifically test AI's resilience to adversarial attacks, including:

Generating Adversarial Examples (FGSM, PGD, C&W, Transferability Attacks).

Performing Fuzzing.

Simulating Model Extraction Attacks.

Simulating Data Poisoning Attacks.

Review Security Logs: Analyze logs for suspicious activity.

Interview Personnel: Interview those responsible for operation and maintenance.

Document Assessment Results: Document identified vulnerabilities, weaknesses, and areas for improvement.

4.5.5 Step 5: Authorize the System (NIST SP 800-37 Step 5)

Purpose: Obtain formal authorization (e.g., Authorization to Operate - ATO) from a designated official.

Detailed Actions:

Prepare Authorization Package: Include System Security Plan, Security Assessment Report, Remediation Plan, Risk Assessment Report.

Obtain Authorization Decision: Secure a decision indicating whether the system is authorized to operate and any conditions.

4.5.6 Step 6: Monitor the System (NIST SP 800-37 Step 6, NIST AI RMF - Measure & Monitor Function)

Purpose: Continuously monitor the security posture and adjust. Crucial for AI as models can drift and new vulnerabilities emerge.

Detailed Actions:

Implement Continuous Monitoring: Include security log monitoring, intrusion detection, vulnerability scanning, performance monitoring.

Conduct Regular Security Assessments: Identify and address new threats/vulnerabilities.

Update Security Controls: Adjust controls based on new threats.

Retrain AI Models: Regularly retrain with fresh data, considering adversarial retraining.

Report Security Incidents: Report to appropriate authorities.

Review and Update Documentation: Regularly update system security documentation.

4.6 Conducting a Risk Assessment

4.6.1 Identify Assets

List all relevant tangible and intangible assets (hardware, software, data, personnel, facilities, documentation, intellectual property), including weapon system-specific assets (targeting algorithms, navigation systems).

4.6.2 Identify Threats

Identify potential threats (internal, external, intentional, unintentional) using MITRE ATLAS. Examples:

External: Adversarial Attacks (Evasion, Poisoning, Model Extraction, Model Inversion), Cyberattacks (Malware, Phishing, DoS), Physical Attacks, Supply Chain Attacks.

Internal: Insider Threats, Accidental Errors.

Environmental: Natural disasters, power outages.

AI-Specific (from ATLAS): Reconnaissance, Resource Subversion, Impair Integrity.

Embedded System-Specific: Sensor spoofing, GPS jamming, communication interception, control system compromise.

4.6.3 Identify Vulnerabilities

Identify weaknesses (software, hardware, data, personnel, AI-specific, weapon system-specific) that could be exploited. Examples: Lack of adversarial training, bias in training data, lack of explainability, insecure communication protocols.

4.6.4 Analyze Likelihood

Determine the likelihood a threat will exploit a vulnerability considering attacker capabilities, motivation, resources, control effectiveness, and threat intelligence. Rate as High, Moderate, Low.

4.6.5 Analyze Impact

Determine the potential impact (confidentiality, integrity, availability, financial, reputational, safety, mission) if a threat exploits a vulnerability. Rate as High, Moderate, Low.

4.6.6 Determine Risk Level

Combine likelihood and impact using a risk matrix (e.g., Critical, High, Moderate, Low, Very Low).

4.6.7 Develop Mitigation Strategies

Develop strategies (avoidance, reduction, transfer, acceptance) to address identified risks. Examples: Stronger access controls, patching, IDS, security awareness, adversarial training, data validation, physical security, human override for embedded systems.

4.6.8 Document and Communicate Risks

Document findings clearly and concisely in a Security Assessment Report and communicate to relevant stakeholders.

4.6.9 Monitor and Review Risks

Continuously monitor risks, review mitigation effectiveness, and update the assessment based on changes in the system, threat environment, or security posture.

Standards Overlay: Apply cloud-relevant controls from the NIST Securing-AI Overlays and Tailoring Guide when implementing 2.2.*, 2.3.*, and 2.4.* in cloud pipelines.

5 CONTEXT-SPECIFIC CONSIDERATIONS

This section details specific considerations for implementing AAIRAF in cloud and embedded (weapon, aviation, healthcare, critical infrastructure system) environments, highlighting key differences.

5.1 Cloud Environment Specifics

Environment-Specific Overlays

• Cloud AI – FedRAMP Moderate/High, CSA AI guidance, zero-trust principles.

• Embedded System AI – DoD RMF overlays, mission assurance and human override.

• Generative AI – Retrieval integrity (RAG), hallucination monitoring, jailbreak defenses.

• Predictive AI – Drift detection, subgroup bias testing, retraining thresholds.

Connectivity: Cloud systems are inherently connected to traditional IT networks, increasing attack surfaces for cyberattacks and data breaches. Implement network segmentation techniques for isolation as applicable.

Data Risks:

Data Poisoning: Malicious actors injecting bad data into training sets accessible via the cloud.

Data Integrity: Accidental corruption or errors in large cloud-stored datasets.

Data Leakage: Unauthorized access to sensitive data stored in the cloud.

Data Bias: Amplified biases in large cloud-based datasets used for training.

Model Risks:

Model Theft: Stealing trained models hosted in the cloud.

Model Manipulation: Unauthorized modification of cloud-hosted models.

Model Drift: Changes in data patterns affecting cloud-deployed model performance over time.

Deployment/Operational Risks:

Denial of Service (DoS) attacks: Disrupting access to cloud-based AI services.

API Security: Vulnerabilities in APIs used to access cloud-based AI.

Dependence on Cloud Provider: Vendor lock-in and potential service disruptions.

Scalability and Resource Management: Challenges managing resources for large-scale AI in the cloud.

Security Measures:

Cloud Provider Security: Assess the security posture of the cloud provider (SOC 2, ISO 27001 compliance).

Cloud-Specific Threats: Address data breaches, misconfiguration of cloud resources, and DoS attacks.

Standards Overlay: Apply embedded system overlays (DoD RMF/Securing-AI Overlays) and Tailoring Guide outputs for 2.2.*, 2.3.*, 2.4.* in constrained or safety-critical contexts.

Cloud Security Controls: Leverage cloud-specific security controls (IAM, NSGs, data encryption).

Cloud Logging and Monitoring: Ensure the configuration of AI-specific logging and monitoring for detecting attacks, unauthorized use, and other security incidents.

Data Residency: Consider data residency requirements.

Container Security: Implement robust container security practices (vulnerability scanning, secure image repositories).

Cloud Access: Cloud access keys/credentials must be securely managed and rotated.

5.2 Embedded Systems Specifics

This section applies broadly to embedded, OT, IoT, and medical Devices. ICS/OT devices require security for PLCs, safety instrumented systems, and deterministic AI recovery, mapped to NERC CIP and ISA/IEC 62443. Healthcare devices must comply with FDA cybersecurity guidance, IEC 62304 (the software lifecycle standard), and IEC 80001 (health IT risk management). IoT devices must integrate NISTIR 8259/8425 requirements, including consumer labeling, secure OTA updates, and lifecycle management at fleet scale. Additional risks include adversarial sensing attacks, such as spoofing or EMI, communication jamming, and unsafe HMI manipulations, which should be tied to MITRE ATLAS-informed test cases.

Connectivity: Embedded systems, especially those in autonomous systems (weapon), can be designed for isolated operation, which reduces some network risks but requires robust offline update mechanisms. When connected (e.g., to platforms, ground support, DoD GIG), they face

traditional network risks. Secure communication protocols (e.g., encryption) must be in place to prevent interception and manipulation.

Real-time Constraints: Embedded systems often operate under stringent real-time constraints, making robust and predictable performance crucial. Security controls must not interfere with these requirements.

Resource Constraints: AI resource constraints introduce systemic risk by limiting computational redundancy, model retraining frequency, and defensive capacity—conditions that can amplify vulnerability exposure and reduce resilience across safety-critical operations. In embedded environments, such constraints may include limited GPU or TPU throughput, restricted memory and power budgets, low-bandwidth communication links, and real-time processing deadlines that inhibit the deployment of comprehensive AI assurance, monitoring, and adversarial defense mechanisms.

Safety Criticality: The consequences of failure are significantly higher in embedded critical infrastructure, aviation, and healthcare systems, making safety and reliability paramount.

Data Risks:

Data Integrity: Sensor malfunctions or adversarial data manipulation affecting input to embedded AI.

Data Spoofing: Presenting false sensor data to mislead the AI system.

Model Risks:

Model Robustness: Ensuring reliable operation in unpredictable real-world conditions.

Unintended Behavior: Unexpected actions by the AI system in unforeseen scenarios.

Adversarial Attacks: Targeted manipulation of sensor inputs to exploit model vulnerabilities.

Deployment/Operational Risks:

Safety and Reliability: Critical importance of fail-safe mechanisms and error handling.

Security Breaches: Unauthorized access and control of the embedded system.

Hardware Failures: Malfunctioning hardware components impacting AI system operation.

Physical Security: Embedded systems are often deployed in physically insecure environments, vulnerable to tampering and theft.

Limited Connectivity: Difficult to monitor for security events and apply patches.

Aviation/Weapon System Unique Threats:

Compromised Sensors: Adversarial manipulation of sensor data (e.g., radar jamming, spoofing GPS signals).

Model Manipulation: Direct access to the embedded AI to alter model parameters or inject malicious code.

Human-Machine Interface (HMI) Exploitation: Targeting vulnerabilities in the HMI to manipulate the AI system's behavior.

Security Measures:

Hardware Security Modules (HSMs) & Trusted Platform Modules (TPMs): Protect cryptographic keys and verify system integrity.

Secure Boot & Code Signing: Prevent unauthorized code execution and ensure integrity.

Memory Protection: Prevent attackers from accessing sensitive data.

Anomaly Detection: Identify unusual data patterns (e.g., sensor data).

Physical Security: Implement measures against tampering and theft.

Redundancy and Fail-Safe Mechanisms: Ensure safe operation even if components are compromised.

Strict Access Control: Implement rigorous access control to the AI model.

Model Obfuscation: Employ techniques to make reverse engineering difficult.

Regular Updates: Establish secure mechanisms for updating AI models and system software, even in limited connectivity environments.

Human Override: Ensure a reliable human override capability in case of AI malfunction or compromise.

5.3 Key Considerations for ML/DL Specific Security (Applicable to both environments)

Data Provenance: Understand the origin and integrity of training data; track data lineage.

Model Bias: Assess and mitigate bias (sampling, measurement, confirmation) to ensure fairness.

Explainability: Strive for explainability (e.g., SHAP, LIME) to understand model decisions, identify vulnerabilities, and build trust.

Regularization: Use techniques to prevent overfitting and improve generalization.

Adversarial Training: Train models with adversarial examples to enhance robustness.

Differential Privacy: Protect privacy of training data.

Federated Learning: Consider for decentralized data without direct sharing.

6 GLOSSARY AND REFERENCES

6.1 Glossary of Terms

AI System: A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

Adversarial Example: An input to an AI model specifically crafted to cause the model to make an incorrect prediction or classification.

Adversarial Training: A defense technique where an AI model is trained on both legitimate and adversarial examples to improve its robustness.

Algorithmic Bias: Systematic and repeatable errors in a computer system that create unfair outcomes, such as favoring or disfavoring particular groups of people.

Backdoor Attacks: Creating hidden triggers that cause the model to misbehave in specific circumstances.

Concept Drift: A change in the relationship between the input data and the target variable over time, leading to reduced model performance.

Data Drift: A change in the distribution of input data over time, which can lead to model degradation.

Data Poisoning: An attack where malicious data is injected into an AI model's training dataset to compromise its integrity or performance.

Explainable AI (XAI): Techniques that allow humans to understand the output of AI models.

Inference: The process of using a trained AI model to make predictions or decisions on new, unseen data.

Large Language Model (LLM): A type of AI model trained on vast amounts of text data, capable of generating human-like text, translating languages, and performing other language-related tasks.

MLOps (Machine Learning Operations): A set of practices that aims to deploy and maintain ML models in production reliably and efficiently.

Membership Inference Attacks: Determining if specific data points were used in training.

Model Evasion: An attack where an adversary crafts inputs to bypass a deployed AI model's detection capabilities.

Model Inversion Attacks: Reconstructing training data from model outputs.

Model Registry: A centralized repository for managing and versioning AI models.

Prompt Injection: An attack specific to LLMs where malicious instructions are inserted into a prompt to manipulate the model's behavior.

Provenance (Data/Model): The documented history of an asset, including its origin, transformations, and usage.

Privacy-Enhancing Technologies (PETs): Technologies that protect personal data when it is used, stored, or transferred (e.g., differential privacy, homomorphic encryption).

Robustness (AI): The ability of an AI model to maintain its performance and integrity when faced with perturbed or adversarial inputs.

6.2 References

National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF)

National Institute of Standards and Technology (NIST) Risk Management Framework (RMF)

National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF 1.0)

ISO/IEC 27001:2022 - Information security, cybersecurity and privacy protection — Information security management systems — Requirements

ISO/IEC 27002:2022 - Information security, cybersecurity and privacy protection — Information security controls

ISO/IEC 42001:2023 - Information technology — Artificial intelligence — Management system

MITRE ATT&CK Framework

MITRE ATLAS (Adversarial Threat Landscape for Artificial-intelligence Systems)

OWASP Top 10 for Large Language Models (LLMs)

European Union Artificial Intelligence Act (EU AI Act)

U.S. Department of Defense (DoD) Risk Management Framework (RMF)

Cybersecurity Maturity Model Certification (CMMC)

U.S. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (October 2023)

DoD AI Ethical Principles

7 CONCLUSION

Securing AI systems, especially ML/DL in cloud and embedded environments like critical weapon systems, infrastructure, aviation, medical devices, and critical infrastructure, is an ongoing and complex process. The Adaptive AI Risk & Assurance Framework (AAIRAF) offers a comprehensive, structured, and adaptable approach to managing the unique cybersecurity, operational, and ethical risks associated with AI. By building on established cybersecurity principles, integrating AI-specific considerations across the entire lifecycle, and incorporating a robust assessment methodology, AAIRAF aims to foster the secure, responsible, and mission-effective development and deployment of artificial intelligence within defense and other critical sectors. It is crucial to stay informed about the latest threats and vulnerabilities and to continuously improve security controls, adopting a risk-based, layered approach tailored to the specific characteristics of the AI system, its environment, and the potential impact of a security breach. Attachment A: AAIRAF Assessor Checklist

Contact Information:

Brian Schleifer

E-mail: brian.schleifer@mtsi-va.com