
**Version 3.7**
September 2026
Document SCL-ARF-001

# AI Requirements Framework

## for Safety-Critical Systems

Requirements and Verification Standards for Artificial Intelligence in Safety-Critical Applications

**Safety Critical Labs**
AI CERTIFICATION AUTHORITY

# Table of Contents

# Section 1: Introduction

## 1.1 PURPOSE

This framework establishes the AI-specific requirements and verifications against which Safety Critical Labs conducts certification assessments for systems incorporating artificial intelligence or machine learning capabilities. It is built for safety-critical and mission-critical applications; the classification assigned under Section 1.2.2 determines which of its requirements apply to a given system and at what stringency.

This framework:

- Establishes requirements addressing AI and ML-specific failure modes not covered by traditional software verification practices
- Defines verification methods and evidence standards for data-driven, probabilistic systems
- Provides AI-specific requirements for bias, drift, hallucination, explainability, and human-AI interaction
- Applies across aerospace & defense, aviation, medical, automotive, industrial & manufacturing, and other safety-critical domains

***Rationale:*** *Traditional software assurance practices assume deterministic, logic-based systems with stable behavior after deployment. AI and ML systems challenge this paradigm through probabilistic outputs, emergent behavior from training data, and potential performance degradation over time. These characteristics require verification approaches designed specifically for data-driven systems.*

## 1.2 SCOPE AND APPLICABILITY

This framework applies to any system, subsystem, or component incorporating artificial intelligence or machine learning capabilities, as determined under Section 1.2.1. The classification assigned under Section 1.2.2 (Safety-Critical, Mission-Critical, or Operational Support) determines which requirements apply and at what

stringency; classification selects requirements, it does not gate entry. The framework is built for safety-critical and mission-critical operations, is domain-agnostic, and is applicable to aerospace & defense, aviation, medical, automotive, industrial & manufacturing, and other domains where AI outputs carry operational consequence.

The framework structures requirements into two tiers: Core Requirements (Section 2) apply to every AI system the framework certifies, regardless of underlying technology; Architecture and Paradigm Requirements (Section 3) apply conditionally based on the system's design choices, including multi-model coordination (AI-11), neural network use (AI-12), and continuous learning (AI-13). Section 4 establishes verification methods and the verification matrix. Section 5 provides informative implementation patterns. Appendices contain the glossary and supporting reference material.

**Table 1-1: Applicability Determination Process**

| Step | Action | Outcome |
|---|---|---|
| 1 | Identify AI/ML presence using Section 1.2.1 criteria | Yes → Step 2; No → standard software assurance practices apply |
| 2 | Classify AI system (Safety-Critical, Mission-Critical, Operational Support) | Determines applicable requirements |
| 3 | Document applicability determination | Retained as verification evidence |
| 4 | Establish tailoring if applicable | Tailoring rationale documented |

Should tailoring of listed requirements be considered, reference Section 1.2.6 on how to assess tailoring approaches.

### 1.2.1 Defining AI

Traditional software behavior is explicitly defined by code. Verification confirms the code implements the required logic, and once verified, behavior remains stable unless the code is changed. AI/ML systems operate differently: behavior emerges from patterns learned during training rather than from explicitly programmed instructions.

This fundamental difference means code inspection alone cannot verify AI behavior, validated performance may not persist over time, and outputs may be non-deterministic. These characteristics do not make AI systems unsuitable for safety-critical operations, but they do require verification approaches designed for data-

driven, emergent behavior. The requirements in this framework provide those approaches.

**Table 1-2: Applicability Criteria**

| Criterion | Indicator | Verification Impact |
|---|---|---|
| Learned behavior (anchor criterion) | System behavior is specified by an artifact induced from data (through training, fine-tuning, continuous learning, or retrieval augmentation, for example) rather than by an explicitly written specification against which correctness can be fully verified by established methods | Behavior cannot be verified by code inspection alone; if met, this framework applies |
| Probabilistic outputs (supporting indicator) | System outputs are non-deterministic or include confidence/uncertainty | Identical inputs may not produce identical outputs; verification emphasizes confidence indication and calibration (AI-8.3) |
| Potential drift (supporting indicator) | System behavior may change or degrade over time without code modifications | Validated performance may not persist without ongoing monitoring (AI-4); triggers the Mission-Critical classification floor per Section 1.2.2 |

The Learned behavior criterion is the anchor criterion and is necessary for applicability. The determination is made by inspection of how the system's correctness can be established. If correctness can be fully verified against explicitly written specification by established methods, with no residual reliance on fidelity to any learned artifact (Appendix A), the system is not AI for the purposes of this framework and standard software assurance practices apply. If correctness can only be established empirically, against data or against fidelity to a learned artifact, the anchor criterion is met and the system is AI for the purposes of this framework; proceed to Section 1.2.2 to determine classification and applicable requirements. The mechanisms named in Table 1-2 (training, fine-tuning, continuous learning, retrieval augmentation) illustrate how learned behavior commonly arises; the determination turns on the availability of an explicitly written specification, not on the presence of any particular mechanism. The supporting indicators do not establish applicability on their own; a system that does not meet the anchor

criterion is outside the scope of this framework even where its outputs are probabilistic or its behavior varies over time. Where the anchor criterion is met, the supporting indicators sharpen verification emphasis as noted in Table 1-2. Decision impact, listed as an applicability criterion in prior versions, is a criticality attribute rather than an AI-presence attribute; it is applied during classification in Section 1.2.2.

The applicability determination is performed on a declared determination boundary: a defined and bounded collection of hardware and software items, together with its declared interfaces, containing the candidate learned component or components. This determination level is adapted from the AI/ML constituent concept of the EASA Artificial Intelligence Concept Paper Issue 2, which introduces an intermediate assurance level between the system level and the item level for learning-based components; the same concept appears in the SAE ARP6983/EUROCAE ED-324 working material (informative cross-reference per Section 1.3.3), and comparable software-item granularity is established practice under NPR 7150.2. Performing the determination at the boundary level prevents over-scoping, in which a predominantly conventional system is swept into AI applicability by the presence of a single learned component, and under-scoping, in which a learned component escapes determination because the surrounding system is conventional. Boundary declarations are subject to a coverage condition: the declaration shall name the encompassing system and enumerate every candidate learned component within it; every candidate learned component shall lie within a declared determination boundary, and a negative determination discharges only the boundary for which it was made.

This framework attaches to each component within the determination boundary whose behavior is induced from data, and to any artifact whose correctness criterion is fidelity to such a component, including distilled, compiled, quantized, or otherwise transformed derivatives (Table 1-2b; AI-4.7). Framework obligations end at the declared interfaces of the determination boundary. Obligations placed on the integrating system, including independent output validation (AI-5.5), fallback response (AI-6.3), and human decision authority (AI-9.1), attach where those requirements specify, in the system that consumes the boundary's outputs. The determination boundary and its interfaces are declared in the Applicability Determination Record (Appendix C.1).

For systems containing data-fitted parameters, the determination applies the deletion test: delete every value derived from data, or designed to be derived from data during operation, and inspect what remains. If a behavioral specification remains, the written structure bounding behavior with values re-derivable by a documented derivation that constitutes a correctness argument, the anchor

criterion is not met; if no behavioral specification remains, the learned artifact (Appendix A) is the specification and the anchor criterion is met. Parameters generated at runtime by an explicitly specified procedure are treated as specified behavior only where established analysis of the procedure bounds the correctness of the resulting behavior. Parameter determination or improvement by machine learning (ISO/IEC 22989:2022 3.3.5), including retraining and continuous learning, does not satisfy this condition regardless of when the optimization runs; such systems meet the anchor criterion, and where learning continues during operation, the AI-13 conditional requirements apply. The deletion test is a synthesis of the structure-versus-parameters distinction from system identification practice and the correctness-criterion logic of Table 1-2b; Appendix B.1 states the test as a determination step.

The following system types are not AI for the purposes of this framework, because their behavior is explicitly specified and verifiable by established methods, notwithstanding that several produce probabilistic outputs or exhibit time-varying behavior. Table 1-2b provides worked examples with the reason for exclusion.

**Table 1-2b: Negative Criteria (Not AI for Framework Purposes)**

| System Type | Why Excluded |
|---|---|
| Deterministic control laws (e.g., PID and gain-scheduled controllers) | Behavior is fully specified by programmed control equations and gains; verifiable by code inspection and classical control analysis |
| Classical state estimators (e.g., Kalman filters, fixed-model particle filters) | Outputs are probabilistic, but dynamics, noise models, and covariances are explicitly specified and analytically verifiable; no behavior is induced from data. For adaptive variants with runtime covariance estimation, see the data-fitted parameters row |
| Fixed rule-based systems (hand-authored expert systems, decision tables) | Logic is explicit and inspectable. Rule sets or rule weights induced from data meet the anchor criterion and are in scope |
| Statistical process control | Fixed statistical tests against programmed control limits; no learned behavior |
| Monte Carlo simulation and dispersion analysis tools | Sampling from explicitly specified models and probability distributions; outputs are statistical ensembles, but every generating model and distribution is written specification, verifiable by established methods; no behavior is induced from |

| System Type | Why Excluded |
|---|---|
| | data |
| Lookup tables and interpolation from engineering data | Static, explicitly specified mappings verifiable by standard methods. Artifacts derived from a learned model (distilled tables, compiled or quantized models) inherit AI applicability, with transformation validation per AI-4.7 |
| Data-fitted parameters within explicitly written model structures (calibrated aerodynamic tables, fitted noise covariances, regression coefficients with documented derivation) | Deleting every data-derived value leaves a behavioral specification intact: the written structure bounds behavior, and the values are re-derivable by a documented derivation that constitutes a correctness argument. Parameters generated at runtime by an explicitly specified procedure whose established analysis bounds the correctness of the resulting behavior (adaptive filter covariance estimation) are excluded on the same basis; parameter updating by machine learning, including continuous learning, is not excluded (AI-13 applies where learning continues during operation). Values produced by machine learning and frozen into tables or parameter sets have no such derivation and remain in scope per the lookup table row above |

Appendix B.1 operationalizes this determination as an evidence-cited decision sequence, and the determination, its declared boundary, and the evidence inspected are recorded in the Applicability Determination Record (Appendix C.1).

### 1.2.2 AI System Classification

Once a system has been identified as meeting the applicability criteria in Section 1.2.1, it shall be classified according to Table 1-3. Classification determines which requirements apply and the level of tailoring authority available.

**Table 1-3: AI System Classification**

| Classification | Description | Applicable Requirements |
|---|---|---|
| Safety-Critical AI | AI outputs directly affect human safety or system survivability | All AI-1 through AI-10 (Core Requirements); plus AI-11 if multi-model, AI-12 if neural network (AI-12.2 and AI-12.4 also apply to other statistical machine learning models), |

| Classification | Description | Applicable Requirements |
|---|---|---|
| | | AI-13 if continuous learning (Architecture and Paradigm Requirements) |
| Mission-Critical AI | AI outputs affect operational success but not human safety | AI-1 through AI-6, AI-8, AI-9, AI-10 (Core Requirements); AI-7 tailored; plus AI-11, AI-12, AI-13 conditionally per system design |
| Operational Support AI | AI supports operations but does not drive critical decisions | AI-1 through AI-6 and AI-10 (Core Requirements) minimum; AI-7, AI-8, AI-9 as tailored; plus AI-11, AI-12, AI-13 conditionally per system design |

Decision impact is applied at this step. AI outputs that directly affect human safety or system survivability classify the system as Safety-Critical; outputs that affect operational success but not human safety classify it as Mission-Critical; all remaining AI systems are Operational Support. Appendix B.2 and B.3 operationalize this determination. Classification shall be informed by the severity of the worst credible consequence of AI system malfunction, drawing on the hazard analysis (AI-1.9); where checklist outcomes and severity assessment conflict, the more severe classification shall apply.

The classification determination shall be documented in the Applicability Determination Record (Appendix C.1) and approved at the authority level specified for the system's classification in Table 1-4 (Project Safety Review Board for Safety-Critical AI, Chief Engineer for Mission-Critical AI, Software Assurance authority for Operational Support AI).

Systems meeting the Potential drift supporting indicator in Table 1-2 shall be classified no lower than Mission-Critical and require tailoring approval per Section 1.2.6.

Classification also triggers the ODD declaration rigor specified in Section 1.2.3 Table 1-4, and AI-10 (Privacy and Data Protection) applies where the system processes personal data under any applicable privacy framework; where personal data is not processed, AI-10 sub-requirements may be documented as Not Applicable with rationale.

### 1.2.3 Operational Design Domain

Following the procedural pattern of Section 1.2.1 (Defining AI) and Section 1.2.2 (AI System Classification), this section establishes ODD declaration as a process step that runs alongside classification. The formal requirement is established in AI-1.0.

The AI system shall have a declared Operational Design Domain (ODD) prior to verification. The ODD defines the operational conditions, inputs, users, subjects, regulatory context, and constraints under which the system is designed and validated to operate and serves as the authoritative scope statement for the certification.

Declaration rigor scales with classification per Table 1-4. The formal requirement and verification for ODD declaration are established in AI-1.0. Tailoring of ODD declaration follows the tailoring authority defined in Section 1.2.6.

**Table 1-4: ODD Declaration Rigor by Classification**

| Classification | Required Declaration Rigor |
|---|---|
| Safety-Critical AI | Operational Claim summary statement plus structured ODD using the full attribute taxonomy (all eight categories) with quantitative bounds for all measurable attributes. Exclusions explicitly enumerated using all five exclusion sub-categories. ODD declaration formally reviewed and approved by the Project Safety Review Board. |
| Mission-Critical AI | Operational Claim summary statement plus structured ODD addressing all eight attribute categories. Quantitative bounds where measurable, qualitative descriptors where appropriate. Exclusions explicitly enumerated using applicable sub-categories. ODD declaration approved by the Chief Engineer. |
| Operational Support AI | Operational Claim summary statement plus documented ODD addressing Operational Environment, Operational Phases, Regulatory and Operational Authorization, Subject Population (where applicable), and Exclusions at minimum. Other categories addressed as applicable to the AI function. ODD declaration approved by the Software Assurance authority. |

### 1.2.4 Relationship to Traditional Software Standards

Traditional software engineering standards were developed for deterministic, logic-based systems. These standards implicitly assume characteristics that do not hold for AI/ML systems. This section documents the gaps that necessitate AI-specific requirements.

#### *1.2.4.1 Gap Analysis*

**Table 1-5: Traditional Software vs. AI/ML Gap Analysis**

| Traditional Software | AI/ML Reality | Risk If Unaddressed |
|---|---|---|
| Fault = bug in code | Fault = bias, drift, hallucination | Wrong failure mode analysis |
| V&V at delivery | Requires continuous validation | Post-deployment failures undetected |
| Requirements to Code traceability | Behavior emerges from training data | Loss of traceability |
| Deterministic outputs | Probabilistic outputs | Unpredictable behavior in edge cases |
| Code coverage metrics | No equivalent for trained models | Inadequate test coverage |
| Explicit decision logic | Black-box decision making | Operators cannot verify AI reasoning |
| Scope implicit from documentation | Validated envelope varies by training data | Unbounded certification claim without explicit operational scope |
| Privacy handled procedurally outside the system | Personal data can be reconstructed from models, membership inferred, decisions evade data subject rights | Privacy regulation violation; model inversion and membership inference risk |

#### *1.2.4.2 Gap-to-Requirement Mapping*

Each AI-specific requirement set in Section 2 addresses one or more gaps identified above. Table 1-6 provides traceability from identified gaps to the corresponding requirements.

**Table 1-6: Gap-to-Requirement Mapping**

| Gap / Risk | Requirement Set |
|---|---|
| Unbounded operational scope | AI-1.0: Operational Design Domain |
| Training data integrity and traceability | AI-1: Operational and Data Foundations |
| Bias and fairness issues | AI-2: Addressing AI Bias |
| No test coverage equivalent | AI-3: ML Test Coverage |

| Gap / Risk | Requirement Set |
|---|---|
| Static after deployment / drift | AI-4: Continuous Validation |
| Unpredictable outputs | AI-5: Hallucination Prevention |
| Vulnerable to OOD inputs | AI-6: Out-of-Distribution Detection |
| Vulnerable to adversarial attacks | AI-7: Adversarial Robustness |
| Black-box decision making | AI-8: Explainability |
| Operators cannot verify reasoning | AI-9: Human-AI Teaming |
| Personal data protection and data subject rights | AI-10: Privacy and Data Protection |
| Multi-model coordination failure modes | AI-11: Multi-Model Systems Requirements |
| Neural network architectural concerns (calibration, adversarial vulnerability, training integrity, generative hallucination) | AI-12: Neural Network Requirements |
| Continuous learning paradigm risks (catastrophic forgetting, learning instability, ODD evolution) | AI-13: Continuous Learning and Adaptation Requirements |

### 1.2.5 Integration with Software Lifecycle

All AI software shall also meet existing software engineering requirements. AI systems are software systems, and the foundational practices of software planning, configuration management, verification, and safety assurance remain applicable. This framework does not create a parallel lifecycle; rather, it identifies AI-specific activities that shall be performed within each phase of the existing software lifecycle. One requirement is deliberately an exception: AI-1.9 addresses the system-level hazard analysis, which is conducted under the project's system safety process (see Appendix A) rather than within the software lifecycle; AI-1.9 is this framework's interface to that upstream process.

**Table 1-7: AI Integration with Software Lifecycle**

| Lifecycle Phase | Title | Key Artifacts |
|---|---|---|
| Requirements | Define AI function criticality; identify protected classes | AI applicability determination |

| Lifecycle Phase | Title | Key Artifacts |
|---|---|---|
| Design | Define model architecture; establish explainability approach | AI design documentation |
| Implementation | Partition datasets; document provenance | Dataset partition records |
| Test | Execute AI-specific test matrix | Test reports, coverage reports |
| Verification | Complete verification matrix; compile evidence | Verification matrix, evidence package |
| Operations | Monitor drift, performance, and bias; revalidate per AI-4 | Drift logs, monitoring reports, revalidation records |
| Retirement | Decommission the AI system; disposition models, datasets, retrieval corpora, and logs per their classification and retention requirements (AI-1.6, AI-1.8) | Disposition records, destruction or archival certification |

### 1.2.6 Tailoring Guidance

Should tailoring of listed requirements be considered, Table 1-8 identifies the tailoring authority for each AI system classification. All tailoring shall be documented with rationale and retained as verification evidence. Changes to ODD declarations (AI-1.0) follow the same tailoring authority as other requirements in the corresponding classification.

**Table 1-8: Tailoring Authority by Classification**

| Classification | Tailoring Authority |
|---|---|
| Safety-Critical AI | No tailoring without Project Safety Review Board approval |
| Mission-Critical AI | Tailoring permitted with Chief Engineer approval and documented rationale |
| Operational Support AI | Tailoring permitted with Software Assurance authority approval |

## 1.3 NORMATIVE REFERENCES

This framework distinguishes normative references from informative ones. Table 1-9 lists the normative references: the documents on which conformity with this

framework depends, each to the stated extent of its invocation and no further. Table 1-10 lists informative references: documents the framework draws on for context, provenance, or crosswalk, which carry no conformity burden. The domain standards listed under each requirement are likewise informative (Section 1.3.1). The conformity basis is stated in Section 1.6.

**Table 1-9: Normative References**

| Document | Title | Extent of Invocation |
|---|---|---|
| ISO/IEC 22989:2022 | AI Concepts and Terminology | Terminology imported in Appendix A at the clauses cited there; the machine learning definition at 3.3.5, invoked by the deletion test in Section 1.2.1 and Appendix B.1 |
| ISO/IEC 5338:2023 | AI System Lifecycle Processes | Terminology imported in Appendix A at the clauses cited there |
| IEEE 1012-2024 | System, Software, and Hardware Verification and Validation | Verification method definitions (Inspection, Analysis, Demonstration, Test) adopted in Section 4.1 |

**Table 1-10: Informative References**

| Document | Title | Relevance |
|---|---|---|
| NIST AI RMF 1.0 | AI Risk Management Framework | Risk taxonomy |
| DO-178C | Software Considerations in Airborne Systems | Aviation software |
| ISO 26262 | Road Vehicles - Functional Safety | Automotive safety |
| ISO/IEC 27701:2019 | Privacy Information Management | Privacy controls |
| ISO/IEC 29100:2011 | Privacy Framework | Privacy principles |
| NIST Privacy Framework v1.0 | Enterprise Privacy Risk Management | Privacy risk taxonomy |
| EU AI Act | Regulation on Artificial Intelligence | AI risk tiering (EU) |
| NPR 7150.2D | NASA Software Engineering Requirements | Software lifecycle reference |

| Document | Title | Relevance |
|---|---|---|
| NASA-STD-8739.8 | Software Assurance and Safety Standard | Safety analysis reference |
| ISO/IEC 23053:2022 | Framework for AI Systems Using ML | ML framework concepts |
| ISO/IEC TR 24027:2021 | Bias in AI Systems and Decision Making | Bias assessment guidance |
| NIST AI 600-1 | AI RMF: Generative AI Profile | Generative AI risks |
| FAA AI Safety Roadmap | Roadmap for AI Safety Assurance | Aviation AI certification concepts |
| GDPR (Regulation 2016/679) | EU General Data Protection Regulation | Personal data processing (EU) |
| HIPAA Privacy and Security Rules | 45 CFR 164 Subparts C and E | Protected health information (US) |
| CCPA/CPRA | Cal. Civ. Code §§ 1798.100 et seq. | Consumer privacy (California) |
| NIST SP 800-122 | Guide to Protecting the Confidentiality of PII | PII handling guidance |
| ISO 34503 | Road Vehicles - Taxonomy for ODD | ODD attribute taxonomy (automotive) |
| UL 4600 | Standard for Safety for the Evaluation of Autonomous Products | Autonomous product safety case reference |
| EASA AI Concept Paper Issue 2 | EASA Artificial Intelligence Concept Paper Issue 2: Guidance for Level 1 and 2 machine learning applications | AI/ML constituent concept; determination boundary source (Section 1.2.1) |

### 1.3.1 Domain Standards Citation Convention

The framework requirements in Sections 2 and 3 are algorithm-agnostic and domain-agnostic. Safety Critical Labs certification assessments are frequently conducted in the context of domain-specific regulatory regimes that impose analogous or complementary requirements. To make that relationship visible at the point of reading, each requirement lists related domain standards inline. The list is a crosswalk: it maps the requirement to obligations the applicant's domain regime may already impose, and it levies nothing.

Domain standards citations are organized in a "Related Domain Standards" subsection at the end of each parent requirement and of selected sub-requirements, structured by domain:

- **General:** Domain-agnostic standards (ISO/IEC, NIST, IEEE)
- **Aviation:** FAA regulations, advisory circulars, and guidance
- **Automotive:** SAE standards, ISO automotive safety standards, NHTSA guidance, UNECE regulations
- **Medical:** FDA guidance documents and regulations, HIPAA where applicable
- **EU:** EU AI Act articles, GDPR articles, and related Union regulations
- **Cross-Domain:** NIST AI RMF function and category references, NIST SP 800 series

A domain block is included only where the requirement has relevant citations in that domain. Absence of a domain block indicates that no specific domain standard currently addresses the requirement at the cited level of detail; this does not imply the requirement is inapplicable in that domain.

Related Domain Standards are informative, not normative. No external document is levied by this framework unless a requirement statement names it and states the extent of the invocation; the normative references are listed in Table 1-9, and the conformity basis is stated in Section 1.6. Certification is conducted against the requirements in Sections 2 and 3, and assessment checklists derive only from the [V.AI-x.y] verification statements, never from the crosswalk. During assessment, Safety Critical Labs performs the crosswalk between those requirements and the applicant's domain regime: the applicant submits the evidence its domain regime already requires, as produced; Safety Critical Labs maps that evidence to the framework requirements it satisfies; and the assessment addresses only what the AI-specific failure modes leave uncovered. The applicant does not demonstrate conformance twice. Safety Critical Labs maps and assesses; it does not produce, design, or remediate the applicant's evidence, and it does not certify conformance to any domain standard, which is not a determination this framework issues.

### 1.3.2 Source Traceability Typology

This framework operates in a domain where no single canonical standard yet covers all relevant requirements. The framework therefore extends, combines, and in some cases originates content. To preserve the rigor expected of certification frameworks, every addition is categorized by its relationship to existing sources, and citations follow the convention below.

**Direct Adoption**: Verbatim use of a concept from an existing standard. Cited as: "Per [Standard] [clause]." Example: the IEEE 1012-2024 verification methods used directly in Section 4.1.

**Adaptation**: Modification of an existing concept for AI/ML context. Cited as: "Adapted from [Standard] [clause]; modified to [reason]." Example: the Operational Design Domain concept adapted from SAE J3016 §3.16 (automotive) to a domain-agnostic AI ODD declaration in AI-1.0.

**Synthesis**: Combination of concepts from multiple sources where no single source covers the whole. Cited as: "Synthesis of [Standard A], [Standard B], [Standard C]." Example: the consequence-based classification tier system in Section 1.2.2 (Safety-Critical AI, Mission-Critical AI, Operational Support AI) draws on NPR 8000.4C (consequence classification), ISO 26262 (ASIL), and DO-178C (DAL). Additional example: AI-13 Continuous Learning and Adaptation Requirements (Section 3.3) is a synthesis of ISO/IEC 22989:2022 5.11.9.2 continuous learning provisions, ISO/IEC 23894:2023 AI risk management for evolving systems, and the framework's own ODD evolvability concept.

**Extension**: Addition of new content to an existing concept where the source recognizes the territory but does not address the AI-specific case. Cited as: "Extends [Standard] [clause] to address [AI-specific concern]." Example: the Operations row in Table 1-7 extends NPR 7150.2D §4 (Software Lifecycle, including Operations, Maintenance, and Retirement) to host AI-specific continuous activities (AI-2.4 bias monitoring, AI-4 family continuous validation, AI-6.2 runtime out-of-distribution detection, AI-9.6 operator role verification).

**Framework-defined**: Novel content where no upstream standard covers the territory. Cited as: "Framework-defined to address [specific gap], drawing analogy from [related concept]." Example: the Subject Population versus User Population distinction in AI-1.0, framework-defined to clarify bias concerns (Subject Population, entities the AI acts upon) versus operator concerns (User Population, humans who direct, monitor, or oversee the AI).

This typology is applied to all framework content and is expected to be maintained in future revisions. Where a concept blends categories (most adaptations also synthesize), the dominant relationship is named.

### 1.3.3 Verification Status of Cited References

Inline domain-standard citations throughout this framework have been verified at clause level against the source documents listed in Tables 1-9 and 1-10 to the extent those documents are available in the certification project library. The following commercially-distributed standards are cited in this framework but were

not available for clause-level verification at the time of this revision; their citations are retained as informative cross-references and shall be re-verified when source documents become accessible:

- RTCA DO-326A (Airworthiness Security Process Specification); cited in R.AI-7 Aviation rows
- RTCA DO-330 (Software Tool Qualification Considerations); cited in R.AI-1 Aviation row
- RTCA DO-355 (Information Security Guidance for Continuing Airworthiness); cited in R.AI-7.6 Aviation row
- RTCA DO-356A (Airworthiness Security Methods and Considerations); cited in R.AI-7 Aviation row
- AAMI HE75 (Human Factors Engineering - Design of Medical Devices); cited in R.AI-9.8 Medical row
- SAE ARP6983/EUROCAE ED-324 (Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI); cited in the Section 1.2.1 determination boundary discussion and in the R.AI-11 and R.AI-12 Aviation rows

Per §1.3.1 and Section 1.6, inline domain citations are informative rather than normative; the framework's normative references are listed in Table 1-9. The unverified citations above are retained because they represent the recognized authoritative standards in their respective domains and support cross-regulatory recognition for clients operating under those regimes. Their inclusion does not constitute a normative dependency of this framework.

## 1.4 VERB APPLICATION AND LEXICON

Requirements are defined in Sections 2 and 3 and are denoted by "shall" in the requirements statement. Associated verifications are co-located with their requirements in Sections 2 and 3 and are summarized in the Section 4.2 Verification Matrix. The certification determination rests on these statements as set out in Section 1.6.

Outcome and commitment statements at the parent requirement level are denoted by "will." "Will" statements describe the aggregate outcome that is expected to be met by the sum of the supporting "shall" sub-requirements. "Will" statements do not carry an independent verification; verification is established at the sub-requirement level.

Best practices are denoted “should” and documented in Section 5. “Should” statements do not carry a formal verification requirement.

AI-specific terminology is defined in Appendix A: Glossary, aligned with ISO/IEC 22989:2022 and ISO/IEC 5338:2023 where applicable.

## 1.5 THRESHOLD CATEGORY DEFINITIONS

The following threshold categories shall be documented for all AI systems regardless of classification. These categories are referenced throughout the requirements in Section 2. Threshold values are project-defined; Section 5.1 provides example ranges and selection guidance. Each threshold value shall be documented with its derivation basis and a justification demonstrating consistency with the declared Operational Design Domain (AI-1.0), the system classification, and the hazard analysis (AI-1.9). The adequacy of project-defined threshold values is adjudicated during certification assessment; the adequacy criteria applied by the assessor are defined in the assessment methodology rather than in this document.

- **Minimum performance threshold:** Minimum acceptable accuracy, precision, recall, or domain-appropriate metric below which system response is required
- **Confidence threshold:** Minimum output confidence level for autonomous action versus human review escalation
- **Data drift threshold:** Maximum deviation from training data distribution (using KL divergence, PSI, or equivalent metric) before revalidation is triggered
- **OOD detection threshold:** Boundary conditions that classify inputs as out-of-distribution requiring fallback behavior
- **Human escalation threshold:** Conditions under which AI outputs require mandatory human review before action

***Rationale:*** *While threshold values are project-defined, these categories are necessary for adequate AI system performance and are required as evidence during certification assessment. Additional threshold categories may be needed based on specific AI function characteristics and operational context. Threshold values shall be derived from and bounded by the declared Operational Design Domain (AI-1.0); thresholds that admit conditions outside the declared ODD are inconsistent with the certification claim.*

## 1.6 CONFORMITY BASIS

The certification determination under this framework rests on the requirement statements of this document and on nothing outside it. For the requirements of Sections 2 and 3, the determination rests on each [R.AI-x.y] requirement statement verified per its paired [V.AI-x.y] verification statement, which comprises the verification method, the confirmations the method shall establish, and the success criteria, applied by the methods of Section 4.1 at the cadence and independence Section 4 specifies. The procedural requirements stated with “shall” in Sections 1 and 4, namely the applicability determination and its record (Section 1.2.1, Appendix B.1, Appendix C.1), classification and its approval (Section 1.2.2), the Operational Design Domain declaration (Section 1.2.3, AI-1.0), threshold documentation (Section 1.5), the declarations required by this section, and verification independence (Section 4.5), form part of the conformity basis and are verified by the evidence those sections name. Rationale paragraphs, the implementation patterns of Section 5, and the worked examples in the appendices are informative. The quantified risk score of Appendix D accompanies the determination and does not decide it.

The Related Domain Standards listed under each requirement, the informative references in Table 1-10, and the citations retained under Section 1.3.3 carry no conformity burden. No external document forms part of the conformity basis unless a requirement statement names it and states the extent of the invocation; the documents so invoked are the normative references of Table 1-9. Assessment checklists derive only from the [V.AI-x.y] verification statements, never from the crosswalk. Where a requirement refers to the regulatory regime applicable to the applicant (for example AI-1.5, AI-8.5, and AI-10), that regime is the one the applicant identifies in its Operational Design Domain declaration under Regulatory and Operational Authorization (AI-1.0); the framework does not select it and does not levy it.

The controls of AI-7 assume a foundational cybersecurity posture for the AI system and for the environment in which it is developed, built, and deployed. That posture is a prerequisite of certification, not assessed scope. The foundational cybersecurity baseline shall be declared in the Applicability Determination Record (Appendix C.1), either by attestation or by reference to existing conformity evidence such as an ISO/IEC 27001 certificate, a SOC 2 report, or a compliance record under DO-326A or ISO/SAE 21434. Safety Critical Labs reviews the declaration for completeness and records it; it does not assess the baseline. The declared baseline is carried on the certificate as an element of the certified configuration, in the same manner as the declared Operational Design Domain (AI-1.0) and the declared threshold values (Section 1.5), so that the certificate states the security baseline on which the certified AI-specific controls rest.

The certified configuration comprises the elements the applicant declares and Safety Critical Labs records on the certificate: the determination boundary and classification (Section 1.2), the Operational Design Domain (AI-1.0), the threshold values (Section 1.5), the attack pattern sets declared under AI-7.1 and AI-7.2, and the foundational cybersecurity baseline declared under this section. Certification attaches to that configuration. A change to any of its elements is a change to the certified configuration and is subject to the change and surveillance provisions of the assessment methodology.

# Section 2: Requirements

The requirements in this section apply to AI systems as defined in Section 1.2.1. Each requirement set (AI-1 through AI-10) addresses a specific AI failure mode identified in Section 1.2.4. Threshold categories referenced throughout these requirements are defined in Section 1.5. Section 2 requirements are algorithm-agnostic and do not prescribe specific implementation approaches; architecture-specific and paradigm-specific requirements are established in Section 3 (Architecture and Paradigm Requirements).

Verb application in this section follows Section 1.4.

*Note: Verifications are co-located with requirements in Section 2 for ease of review. They are summarized in the verification matrix in Section 4.2.*

## 2.1 AI-1: OPERATIONAL AND DATA FOUNDATIONS

### [R.AI-1] OPERATIONAL AND DATA FOUNDATIONS

The AI system will establish operational scope and data integrity sufficient to support reliable model behavior and a defensible certification claim throughout the model lifecycle.

Traditional software has direct traceability from requirements to code; an engineer can inspect the code to verify it implements the requirement. AI systems break this traceability because behavior emerges from training data within a declared operational envelope. AI-1 establishes the foundations on which all downstream verification depends: the operational envelope within which the system is designed to function (AI-1.0), the integrity of the data partitions that produced the model (AI-1.1 through AI-1.4), the classification handling controls that govern the data and outputs throughout the lifecycle (AI-1.5 through AI-1.8), the integration of AI failure modes into the project's hazard analysis (AI-1.9), and the controls governing retrieval corpora where retrieval augmentation is used (AI-1.10). AI-1.0 is numbered as a foundational scoping requirement that precedes the data and classification

controls; the Operational Design Domain declared under AI-1.0 establishes the bounds within which all subsequent AI-1.x sub-requirements (and the broader AI-2 through AI-10 requirement sets) apply.

**Applicability:** AI-1.5 through AI-1.8 apply where the AI system or its data pipeline is subject to an information classification or export control regime (CUI, ITAR, EAR, or organization-specific markings). Where no such regime applies, AI-1.5 through AI-1.8 may be documented as Not Applicable with rationale, following the pattern established for AI-10. AI-1.10 applies where the system employs retrieval-augmented generation; where it does not, AI-1.10 may be documented as Not Applicable with rationale.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 5; ISO/IEC 5338:2023 §6.4
- Aviation: DO-178C §5 (Software Development); DO-330 §5 (Tool Qualification); FAA AI Safety Roadmap
- Automotive: ISO 26262-6:2018 §5 (Software development); ISO 21448:2022 §6; ISO 34503 (ODD attributes)
- Medical: FDA AI/ML SaMD Action Plan (2021); FDA PCCP Guidance (2023); IEC 62304:2006+A1:2015 §5.1
- EU: EU AI Act Article 10 (Data and data governance); EU AI Act Article 17 (Quality management)
- Government/Defense: 32 CFR Part 2002 (CUI); 22 CFR 120.17 (ITAR deemed exports); 15 CFR 762 (EAR recordkeeping)
- Cross-Domain: NIST AI RMF MAP-1 through MAP-5; NIST SP 800-53 Rev. 5

### 2.1.0 Operational Design Domain

**[R.AI-1.0] OPERATIONAL DESIGN DOMAIN**

The AI system shall declare an Operational Design Domain (ODD) that defines the operational conditions, inputs, users, subjects, regulatory context, and constraints under which the system is designed and validated to operate. The ODD declaration shall begin with an Operational Claim summary statement and shall address the eight attribute categories defined in Table 2.1.0-A at the declaration rigor established for the system's classification in Table 1-4; Operational Support systems shall address, at minimum, the reduced category set specified there.

***Rationale:*** *AI systems trained against representative data perform reliably within the conditions represented in that data and become unreliable outside those conditions. Without an explicit declaration of operational scope, the certification claim is unbounded; the certificate appears to attest to system behavior under all conditions when verification evidence supports only a subset. ODD declaration establishes the operational envelope as a contractual element of the certification, scoping the certification claim to declared conditions and providing the basis from which downstream requirements are derived: dataset partitioning (AI-1.1), bias evaluation across subject populations (AI-2), operational input distribution testing (AI-3.3), output validity boundaries (AI-4.4), training distribution characterization (AI-6.1), and operational role verification (AI-9.6). ODD is the operational complement to system classification: classification establishes what the system is; ODD establishes where, for whom, and how it is intended to operate.*

**Operational Claim:** The ODD declaration shall begin with a single-sentence Operational Claim summarizing the scope of the certification: *"This AI system is designed and validated to [function] for [subject population] under [operational conditions], operated by [user role] within [regulatory regime]."* The Operational Claim is the elevator-summary of the ODD; the attribute categories below specify it in detail.

The ODD declaration shall address the attribute categories defined in Table 2.1.0-A, at the coverage and rigor established for the system's classification in Table 1-4. Exclusions shall be declared using the five exclusion sub-categories defined in Table 2.1.0-B. Threshold categories (Section 1.5) and operational input distributions (AI-3.3) shall be derived from and consistent with the declared ODD.

**Table 2.1.0-A: Required ODD Attribute Categories**

| Category | Description | Examples | Downstream Requirements |
|---|---|---|---|
| Operational Environment | Physical, spatial, atmospheric, and temporal conditions of operation | Geographic scope, weather and atmospheric conditions, illumination and time-of-day ranges, terrain or surface types, ambient temperature ranges, care setting (medical), facility class | AI-3.3, AI-6.1 |
| System Inputs | Sensor and data source characteristics, supported health | Sensor types and configurations, sensor health states under which operation is supported, data source | AI-3.3, AI-6.1, AI-9.4 |

| Category | Description | Examples | Downstream Requirements |
|---|---|---|---|
| | states, input quality envelopes, and defined system behavior at degradation thresholds | provenance, input quality bounds, defined behavior (degraded mode, disengage, fallback, non-diagnostic message) when degradation thresholds are reached | |
| User Population | Operator qualifications, training, currency, authorized roles, and role within operational workflow | Required operator certifications, system-specific training, currency requirements, role-based authorization levels, supported team configurations, user role in the decision workflow (autonomous, decision support, advisory) | AI-9.1, AI-9.3 |
| Subject Population | Populations or entities the AI's decisions act upon, with validated subject characteristics and excluded subject conditions | Validated patient demographics (medical), vehicle/pedestrian/cyclist populations the perception system handles (automotive), aircraft and obstacle classes (aviation), entity classes for surveillance or targeting systems | AI-2, AI-3.3, AI-4.4 |
| Operational Phases | Intended use cases, operational phase transitions, and phase-level constraints | Supported workflow stages or mission segments, mode transition conditions, supported task categories, maximum phase or session duration, operational tempo (frequency, duty cycle, throughput per operator) | AI-9.4, AI-9.6 |
| Performance Envelope | Per-decision performance limits and timing | Detection ranges, sensitivity and specificity, accuracy thresholds, decision latency, | AI-3.3, AI-4.2, AI-4.4, AI-8.3 |

| Category | Description | Examples | Downstream Requirements |
|---|---|---|---|
| | requirements | throughput bounds, confidence calibration error, computational resource availability | |
| Regulatory and Operational Authorization | Applicable regulatory regime, required certifications, waivers, authorization conditions, and ODD evolvability under change control | Governing regulatory framework (FAA Part 107, FDA SaMD class, ISO 26262 ASIL, EU AI Act risk class), required waivers or special authorizations, jurisdictional limitations, applicable industry standards, ODD evolvability declaration (locked at certification, or evolvable under defined change control mechanism such as FDA PCCP) | All of Section 2 (foundational scoping) |
| Exclusions | Explicit out-of-scope conditions under which the system shall not engage | See Table 2.1.0-B for required exclusion sub-categories | AI-4.4, AI-6.3, AI-9.1 |

*Note on Operational Phases vs. Performance Envelope: Operational Phases captures phase-level or mission-level constraints (duration, sequence, transitions, tempo). Performance Envelope captures per-decision performance (latency, throughput, accuracy at the inference level). Time-bounded attributes that describe the workflow or mission belong in Operational Phases; time-bounded attributes that describe individual inferences belong in Performance Envelope.*

*Note on Subject Population: Not all systems have a distinct subject population from their user population. Where the AI acts on the user themselves (e.g., a personal recommendation system with no third-party subject), the Subject Population entry may state "Subject = User; see User Population." Where subjects are abstract objects rather than people (e.g., obstacle detection in drones), Subject Population still applies and should declare the validated object classes.*

**Table 2.1.0-B: Required Exclusion Sub-Categories**

| Sub-Category | Description | Example |
|---|---|---|
| Excluded Environments | Environmental conditions outside the operational envelope | Night operations, heavy precipitation, visibility below threshold, temperature extremes, terrain types not supported, deployment outside cleared jurisdiction |
| Excluded Use Cases | Mission types or operational scenarios outside intended use | Aerobatic maneuvers, formation operations, payload drop, multi-system coordination, mission durations beyond validated bounds, screening use of a diagnostic-only system |
| Excluded Operator Conditions | Operator states, qualifications, or staffing configurations not supported | Operators without current certification, operators below minimum currency, staffing below required minimum, non-radiologist users of a radiology decision support system |
| Excluded Subject Conditions | Subject populations or conditions outside validated scope | Pediatric patients in an adult-validated medical system, pedestrians in wheelchairs or carrying large objects in a perception system not validated for those cases, subject conditions explicitly excluded from training data |
| Excluded System States | System or sensor degradation conditions that trigger automatic disengage or fallback | Sensor degradation beyond defined thresholds, communication loss conditions, computational resource exhaustion, image quality below threshold, GPS-denied conditions |

**[V.AI-1.0] OPERATIONAL DESIGN DOMAIN**

The AI system shall verify ODD declaration by inspection.

The inspection shall confirm that the ODD declaration includes a single-sentence Operational Claim summary, that the declaration addresses all eight attribute categories required for the system's classification per Section 1.2.3 Table 1-4, that quantitative bounds are documented where the attribute is measurable, that exclusions are explicitly enumerated using the five sub-categories defined in Table 2.1.0-B, that the ODD declaration is approved at the authority level specified for the system's classification, and that downstream artifacts (operational input distributions per AI-3.3, training distribution characterization per AI-6.1, output

validity boundaries per AI-4.4, bias evaluation populations per AI-2) are derived from and consistent with the declared ODD.

**Success Criteria:** The verification shall be considered successful when the inspection shows that the Operational Claim is present and consistent with the detailed ODD, ODD declaration addresses all required attribute categories, quantitative bounds are documented where measurable, exclusions are explicitly enumerated using the required sub-categories, declaration is approved at the required authority level, and downstream artifacts are traceable to the declared ODD.

**ODD Maintenance:** The ODD declaration shall be reviewed at the periodic intervals defined for operational role verification (AI-9.6). Identified divergence between declared ODD and actual operational use shall trigger formal ODD revision under the same authority structure as initial declaration. For systems declared as evolvable under a change control mechanism (e.g., FDA PCCP), ODD evolution shall follow the approved change control plan; ODD changes outside the approved plan shall require recertification.

### 2.1.1 Dataset Separation

**[R.AI-1.1] DATASET SEPARATION**

The AI system shall maintain separation between training, validation, and test datasets.

***Rationale:*** *Data leakage, which occurs when information from test or validation sets contaminates training data, produces artificially inflated performance metrics. A model that has* ***"seen"*** *test data during training will perform well on that data but may fail on novel operational inputs. Strict separation with technical enforcement ensures performance estimates reflect true generalization capability.*

**[V.AI-1.1] DATASET SEPARATION**

The AI system shall verify dataset separation by inspection.

The inspection shall confirm dataset partitions are documented with unique identifiers, partition boundaries are enforced with technical controls (e.g., separate storage locations, access controls, automated pipeline checks), no data leakage exists between partitions, and partition integrity is verified prior to model training and evaluation.

**Success Criteria:** The verification shall be considered successful when the inspection shows that dataset partitions are documented with unique identifiers,

partition boundaries are enforced with technical controls, no data leakage exists between partitions, and partition integrity is verified prior to training and evaluation.

### 2.1.2 Partition Use Restriction

**[R.AI-1.2] PARTITION USE RESTRICTION**

The AI system shall ensure each dataset partition is used only for its intended purpose.

***Rationale:*** *Even with proper separation, using partitions incorrectly (e.g., using validation data for final testing, or repeatedly evaluating against test data during development) compromises the integrity of performance estimates. Restricting partition use to intended purposes and logging all access ensures data integrity is maintained throughout the development lifecycle.*

**[V.AI-1.2] PARTITION USE RESTRICTION**

The AI system shall verify partition use restriction by inspection and analysis.

The inspection shall confirm intended use for each partition is documented (training for model fitting, validation for hyperparameter tuning and model selection, test for final performance evaluation).

The analysis shall confirm access controls restrict partition use to authorized processes, audit logs capture dataset access and usage, and violation of intended use triggers alert and investigation.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that intended use for each partition is documented, access controls restrict partition use to authorized processes, audit logs capture access and usage, and violations trigger alerts and investigation.

### 2.1.3 Data Provenance

**[R.AI-1.3] DATA PROVENANCE**

The AI system shall document data provenance for all datasets used in model development.

***Rationale:*** *AI model behavior is a direct function of training data. Without documented provenance, the source, quality, and potential biases in training data cannot be assessed. Provenance documentation enables root cause analysis when model failures occur, supports retraining decisions, and provides traceability required for safety-critical systems.*

**[V.AI-1.3] DATA PROVENANCE**

The AI system shall verify data provenance by inspection.

The inspection shall confirm data sources are identified and documented for each dataset, data collection methods are documented (including sensors, sampling rates, time periods, and environmental conditions), data transformations and preprocessing steps are recorded (including normalization, augmentation, filtering, and feature extraction), and provenance documentation is traceable to source records.

**Success Criteria:** The verification shall be considered successful when the inspection shows that data sources are identified and documented, collection methods are documented, transformations and preprocessing are recorded, and provenance is traceable to source records.

### 2.1.4 Ground Truth Labeling Quality

**[R.AI-1.4] GROUND TRUTH LABELING QUALITY**

The AI system shall validate ground truth labeling quality for all supervised learning datasets.

***Rationale:*** *Supervised learning models learn to predict labels. If labels are incorrect, the model learns incorrect behavior. Label errors are particularly dangerous because they directly encode wrong answers into the model. For safety-critical applications, label quality must be validated through independent verification, as label errors could cause the model to make dangerous predictions with high confidence.*

**[V.AI-1.4] GROUND TRUTH LABELING QUALITY**

The AI system shall verify ground truth labeling quality by inspection and analysis.

The inspection shall confirm labeling methodology and labeler qualifications are documented.

The analysis shall confirm inter-rater reliability metrics meet defined thresholds for multi-labeler datasets (e.g., Cohen's Kappa $\geq 0.8$ for safety-critical applications), label accuracy is validated through spot-checking or independent verification for safety-critical applications, ambiguous or disputed labels are identified and resolved with documented rationale, and label error rate is estimated and its impact on model performance is assessed.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that labeling methodology and qualifications are documented, inter-rater reliability meets defined thresholds, label accuracy is

validated for safety-critical applications, ambiguous labels are resolved with rationale, and label error rate impact is assessed.

### 2.1.5 Pre-Ingestion Classification Screening

**[R.AI-1.5] PRE-INGESTION CLASSIFICATION SCREENING**

The AI system shall screen all data sources for information classification markings prior to ingestion into model training, fine-tuning, or retrieval-augmented generation (RAG) pipelines.

***Rationale:*** *AI training pipelines ingest data from diverse sources including government databases, technical repositories, contractor deliverables, sensor archives, and publicly available datasets. Any of these sources may contain data subject to information handling controls (CUI, ITAR, EAR, or other classification designations). Once controlled data is ingested into model training, the classification marking is permanently separated from the content. Pre-ingestion screening is the only reliable point at which to prevent classification loss, as no post-training mechanism exists to identify or remove controlled content from model parameters. This requirement establishes screening as a mandatory gate in the data pipeline, analogous to the partition integrity verification required by AI-1.1 prior to every training run.*

**[V.AI-1.5] PRE-INGESTION CLASSIFICATION SCREENING**

The AI system shall verify pre-ingestion classification screening by inspection and analysis.

The inspection shall confirm that a documented screening process exists for all data sources entering AI training, fine-tuning, or RAG pipelines; that screening criteria include CUI markings (per 32 CFR Part 2002 marking requirements), ITAR controlled technical data identifiers, EAR Export Control Classification Numbers (ECCNs), and any organization-specific classification markings; and that screening results are documented for each data source with disposition (approved, excluded, or approved with handling controls).

The analysis shall confirm that automated screening tools are validated against known-positive controlled content at a project-defined detection rate threshold (see Section 5.1); that manual review procedures are defined for data sources where automated screening is insufficient; and that screening coverage is complete, and no data enters the training pipeline without a documented screening disposition.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that a documented screening process exists,

screening criteria cover all applicable classification regimes, screening results are documented with disposition for each data source, automated tools are validated against known-positive content, manual review procedures exist for edge cases, and no data enters the pipeline without screening disposition.

### 2.1.6 Information Classification Inheritance

**[R.AI-1.6] INFORMATION CLASSIFICATION INHERITANCE**

The AI system shall document the information classification level of all training data and shall apply the principle that AI model artifacts and outputs inherit the highest classification level of the training data unless the training pipeline demonstrates, with verifiable evidence, that controlled content was excluded.

***Rationale:*** *In traditional information systems, data classification follows the data: a CUI document copied to a new system requires the new system to handle the copy as CUI. AI systems break this principle because the data transformation from documents to model weights destroys the classification metadata while retaining the controlled content. Without an explicit inheritance rule, the default assumption becomes that AI outputs are uncontrolled, which is incorrect when the training data included controlled content. This requirement establishes classification inheritance as the default, placing the burden of proof on the pipeline to demonstrate that controlled content was excluded rather than on the consumer to determine whether generated outputs contain controlled information. This is consistent with the derivative classification principle in information security: products derived from classified sources inherit the classification of the source material.*

**[V.AI-1.6] INFORMATION CLASSIFICATION INHERITANCE**

The AI system shall verify information classification inheritance by inspection and analysis.

The inspection shall confirm that the classification level of all training datasets is documented in the data provenance record required by AI-1.3; that model artifacts (trained weights, serialized models, deployment packages) are marked with the inherited classification level; and that the classification inheritance determination is documented with justification, either inheriting the highest training data classification or certifying exclusion of controlled content with evidence.

The analysis shall confirm that the classification inheritance chain is complete from data sources through model artifacts to deployment endpoint; that any claim of controlled content exclusion is supported by the screening evidence required by

AI-1.5; and that downstream systems receiving model outputs are authorized to handle information at the inherited classification level.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that training data classification is documented in provenance records, model artifacts carry inherited classification markings, inheritance determination is documented with justification, the inheritance chain is complete, exclusion claims are supported by screening evidence, and downstream systems are authorized for the inherited classification level.

### 2.1.7 Output Classification Awareness

**[R.AI-1.7] OUTPUT CLASSIFICATION AWARENESS**

The AI system shall implement mechanisms to ensure that consumers of AI-generated outputs are aware of the information classification handling requirements applicable to those outputs.

***Rationale:*** *When an AI model trained on or exposed to controlled content generates outputs, those outputs may contain reconstructed controlled information. If the model operates at an inherited classification level per AI-1.6, all outputs must be handled at that level. However, the primary risk is that end users receive AI-generated content without any indication of applicable handling requirements. The user treats the output as uncontrolled general knowledge, copies it to uncontrolled systems, shares it with unauthorized recipients, or disseminates it to foreign persons, each constituting a potential violation of applicable regulations. This requirement ensures that classification awareness propagates to the point of consumption, not just the point of model storage. For ITAR technical data, failure to maintain this awareness constitutes a potential deemed export violation (22 CFR 120.17). For CUI, it violates dissemination controls established by the CUI Registry.*

**[V.AI-1.7] OUTPUT CLASSIFICATION AWARENESS**

The AI system shall verify output classification awareness by test and analysis.

The test shall confirm that AI-generated outputs include or are accompanied by applicable classification handling indicators appropriate to the output channel (e.g., CUI markings on documents, banner markings on display interfaces, metadata tags on programmatic outputs); that output classification indicators are consistent with the inherited classification level determined under AI-1.6; that classification indicators cannot be trivially stripped or bypassed by end users during normal operational use; and that when AI outputs are consumed by downstream systems or pipelines, classification metadata is propagated programmatically.

The analysis shall confirm that output classification awareness mechanisms are appropriate for all operational output channels (display, document generation, API response, data export); that operator training materials address the classification handling requirements for AI-generated outputs; and that the output classification mechanism addresses the case where a model is queried by users at different authorization levels, ensuring outputs are appropriate to the requesting user's authorization.

**Success Criteria:** The verification shall be considered successful when the test and analysis show that outputs include classification handling indicators, indicators are consistent with the inherited classification level, indicators are not trivially bypassed, downstream propagation is implemented, all output channels are covered, operator training addresses output handling, and user authorization levels are respected in output generation.

### 2.1.8 Classification Compliance Audit Trail

**[R.AI-1.8] CLASSIFICATION COMPLIANCE AUDIT TRAIL**

The AI system shall maintain an audit trail documenting classification screening decisions, inheritance determinations, output classification events, and any identified or suspected classification incidents throughout the AI system lifecycle.

***Rationale:*** *Regulatory compliance for CUI, ITAR, and EAR requires demonstrable evidence of proper handling. Traditional information systems maintain access logs and handling records. AI systems require analogous records covering the AI-specific aspects of information classification: what data was screened and what was the disposition, what classification level was inherited and on what basis, what outputs were generated at what classification level, and whether any incidents of suspected classification loss or improper dissemination occurred. Without this audit trail, an organization cannot demonstrate compliance during regulatory review, cannot scope the impact of a discovered classification incident, and cannot support the continuous improvement of classification controls. This requirement is the classification-specific complement to the human-AI interaction logging required by AI-9.5 and the OOD event logging required by AI-6.4.*

**[V.AI-1.8] CLASSIFICATION COMPLIANCE AUDIT TRAIL**

The AI system shall verify classification compliance audit trail by inspection.

The inspection shall confirm that pre-ingestion screening decisions are logged with timestamps, data source identifiers, screening method, and disposition; that classification inheritance determinations are logged with justification and authorizing authority; that output classification events are logged at a level

sufficient to support incident investigation (including query context, output classification level applied, and output channel); that suspected classification incidents are logged with timestamps, discovery method, affected data scope estimate, and remediation actions; and that audit trail records are retained for a period consistent with applicable regulatory requirements (minimum retention periods per 32 CFR Part 2002 for CUI, ITAR record retention requirements per 22 CFR 122.5, and EAR record retention requirements per 15 CFR 762).

**Success Criteria:** The verification shall be considered successful when the inspection shows that screening decisions are logged, inheritance determinations are logged with justification, output classification events are logged, suspected incidents are logged with remediation actions, and retention periods meet applicable regulatory requirements.

### 2.1.9 AI Hazard Analysis Integration

**[R.AI-1.9] AI HAZARD ANALYSIS INTEGRATION**

The AI system shall be included in the system-level hazard analysis conducted under the project's applicable system safety process, and that hazard analysis shall address the AI-specific failure mode categories identified in Section 5.2.2: data-related, model-related, operational, adversarial, privacy, and human factors. Identified AI hazards shall trace to the framework requirements and project-defined thresholds that mitigate them, and hazard analysis outputs shall inform system classification (Section 1.2.2) and threshold selection (Section 1.5).

***Rationale:*** *The foundational principle of this framework is that requirements are defined by the failures they prevent. The hazard analysis is where a project enumerates those failures for its specific system, yet prior versions treated AI hazard analysis as informative guidance only. This requirement establishes the normative hook: it does not create a parallel hazard analysis process, but requires that the hazard analysis the project's system safety process already mandates (FMEA, FMECA, FTA, STPA, or PRA per applicable domain practice) explicitly address AI-specific failure modes and connect its outputs to classification, threshold selection, and requirement tailoring. Section 5.2 provides implementation guidance for this integration.*

**[V.AI-1.9] AI HAZARD ANALYSIS INTEGRATION**

The AI system shall verify AI hazard analysis integration by inspection and analysis.

The inspection shall confirm that a system-level hazard analysis conducted under the project's applicable system safety process exists and includes the AI system; that the analysis addresses each AI-specific failure mode category identified in

Section 5.2.2 or documents the category as not credible for the system with rationale; and that the analysis is maintained current with the system configuration under change control.

The analysis shall confirm that each identified AI hazard traces to the requirement or requirements and the threshold or thresholds that mitigate it, or to a documented risk acceptance, and that hazard analysis outputs are reflected in the system classification and in threshold justifications.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that the hazard analysis includes the AI system, addresses all AI-specific failure mode categories or documents non-credibility with rationale, remains current under change control, and traces identified hazards to mitigating requirements, thresholds, or documented risk acceptances.

**Related Domain Standards**

- General: ISO/IEC 23894:2023 (AI risk management); ISO 31000:2018 (Risk management principles)
- Aviation: SAE ARP4761 (Guidelines for conducting the safety assessment process on civil airborne systems)
- Automotive: ISO 26262-3:2018 Clause 6 (Hazard analysis and risk assessment); ISO 21448:2022 (Safety of the intended functionality)
- Medical: ISO 14971:2019 (Application of risk management to medical devices)
- Aerospace/Defense: NASA-STD-8739.8 (Software safety analysis); MIL-STD-882E (System safety)
- Cross-Domain: NIST AI RMF MAP function; STPA Handbook (Leveson and Thomas, 2018)

### 2.1.10 Retrieval Corpus Controls

**[R.AI-1.10] RETRIEVAL CORPUS CONTROLS**

Where the AI system employs retrieval-augmented generation or otherwise incorporates retrieved content into model inputs or outputs at runtime, the retrieval corpus shall be subject to documented controls including: source provenance meeting the criteria of AI-1.3; content screening prior to corpus admission per AI-1.5; integrity protection and change control for corpus additions, modifications, and removals; and protection against corpus poisoning and indirect prompt injection per AI-7.1 and AI-7.2. Where retrieval augmentation is not used, this sub-requirement may be documented as Not Applicable with rationale.

***Rationale:*** *Retrieval augmentation moves behavior-determining content out of trained parameters and into a corpus that can change at runtime. A corpus change is functionally a model behavior change, but it bypasses the training pipeline controls of AI-1.1 through AI-1.4 unless the corpus is explicitly controlled. An uncontrolled corpus is also the entry point for indirect prompt injection, in which adversarial instructions embedded in retrieved content manipulate system outputs. This requirement extends the data foundation controls to the retrieval pathway so that corpus content receives the same provenance, screening, and integrity discipline as training data. Corpus content enters the system through loaders; the loading of corpus artifacts is subject to AI-7.7 (Artifact Load Safety).*

**[V.AI-1.10] RETRIEVAL CORPUS CONTROLS**

The AI system shall verify retrieval corpus controls by inspection and test.

The inspection shall confirm that corpus sources are documented with provenance per AI-1.3; that each corpus source has a recorded screening disposition per AI-1.5; that corpus additions, modifications, and removals are executed under documented change control with an audit trail retained per AI-1.8; and that corpus access controls restrict modification to authorized processes and personnel.

The test shall confirm that corpus modifications outside the change control process are detected and rejected or alerted, and that content containing the indirect prompt injection patterns in the attack pattern set declared under AI-7.2 is detected during screening or its influence on system outputs is bounded per the protections verified under AI-7.2.

**Success Criteria:** The verification shall be considered successful when the inspection and test show that corpus provenance, screening dispositions, change control, and access controls are documented and operating, unauthorized corpus modifications are detected, and indirect prompt injection protections function as designed. Where retrieval augmentation is not used, AI-1.10 may be documented as Not Applicable with rationale.

**Related Domain Standards**

- General: NIST AI 600-1 (Generative AI Profile); NIST SP 800-53 Rev. 5 CM-3 (Configuration change control) and SI-7 (Software, firmware, and information integrity)
- AI-Specific: OWASP Top 10 for LLM Applications LLM01 (Prompt injection); MITRE ATLAS AML.T0051 (LLM prompt injection); MITRE ATLAS AML.T0020 (Poison training data)

- EU: EU AI Act Article 10 (Data and data governance); EU AI Act Article 15(5) (Resilience against adversarial inputs)
- Cross-Domain: NIST AI RMF MAP-2.3 and MANAGE-3

## 2.2 AI-2: ADDRESSING AI BIAS

### [R.AI-2] ADDRESSING AI BIAS

The AI system will implement bias detection, prevention, and mitigation mechanisms to provide equitable and unbiased decision-making across all applicable operational scenarios and applicable Subject Populations and User Populations declared under AI-1.0.

AI systems trained on historical data can encode and amplify biases present in that data, leading to systematic performance disparities across operational conditions or user populations. Bias in this context refers to any systematic deviation in AI performance, whether across environmental conditions, sensor degradation states, or user characteristics. This requirement addresses the gap between traditional software verification (which assumes uniform behavior) and AI systems whose behavior may vary systematically based on input characteristics. Bias evaluation populations shall be derived from the Subject Population declared under AI-1.0.

**Related Domain Standards**

- General: ISO/IEC TR 24027:2021; ISO/IEC 22989:2022 3.5.4
- Aviation: FAA AI Safety Roadmap (equitable operation across crew demographics) `p.8 ("Focus on Safety Assurance and Safety Enhancements" principle)`
- Medical: FDA Draft Guidance on AI-Enabled Device Software Functions (2025) `Section III (TPLC Approach: General Principles) and Section VIII (Data Management)`; FDA Digital Health Policy Navigator
- EU: EU AI Act Article 10(2)(f) (bias examination and correction); EU AI Act Article 15(3) (accuracy across groups)
- Consumer/Employment: EEOC AI Technical Assistance (2023) (where employment-adjacent)
- Cross-Domain: NIST AI RMF MANAGE-2.3; NIST SP 1270 (Taxonomy of AI Bias, 2022)

### 2.2.1 Bias-Bounded Baseline Performance

**[R.AI-2.1] BIAS-BOUNDED BASELINE PERFORMANCE**

The AI system shall be baselined with performance disparities across user classes and operational contexts bounded within approved bias thresholds.

***Rationale:*** *Establishing baseline performance prior to deployment provides a reference point for continuous monitoring and enables detection of bias drift during operations. Without documented baselines, degradation in performance cannot be objectively measured or addressed.*

**[V.AI-2.1] BIAS-BOUNDED BASELINE PERFORMANCE**

The AI system shall verify bias-bounded baseline performance by analysis and test.

The analysis shall confirm bias performance criteria, including approved disparity thresholds, are defined for all applicable user attributes and operational contexts relevant to the AI function.

The test shall provide results congruent with defined bias performance criteria using validation datasets representative of operational conditions.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that bias-bounded baseline criteria are defined, test results meet defined criteria prior to operational deployment, and a validation report captures baseline values, threshold criteria, and threshold justification.

**Related Domain Standards**

- General: ISO/IEC TR 24027:2021 §6 (Bias assessment approaches); ISO/IEC 22989:2022 3.5.4 (Bias definition)
- Aviation: FAA AI Safety Roadmap (equitable performance across operational demographics) `p.8 ("Focus on Safety Assurance and Safety Enhancements" principle)`
- Automotive: UNECE WP.29 R157 (ALKS) performance consistency requirements `§5.2 (Dynamic Driving Task)`
- Medical: FDA Draft Guidance on AI-Enabled Device Software Functions (2025) `Section VIII (Data Management), Section X.A (Performance Validation), and Appendix C (Performance Validation Considerations)`; FDA Digital Health Policy Navigator

- EU: EU AI Act Article 10(2)(f) (bias examination and correction); EU AI Act Article 10(3) (training data representativeness); EU AI Act Article 15(3) (performance across demographic groups)
- Consumer/Employment (where applicable): EEOC AI Technical Assistance (2023)
- Cross-Domain: NIST AI RMF MANAGE-2.3; NIST SP 1270:2022 (Taxonomy of AI Bias)

### 2.2.2 Bias-Managed Training Data

**[R.AI-2.2] BIAS-MANAGED TRAINING DATA**

The AI system shall use training datasets that account for underrepresented groups, historical biases, and labeling errors.

***Rationale:*** *Training data quality directly determines AI system behavior. Datasets with underrepresented groups, historical biases, or labeling errors will produce models that perpetuate those biases.*

**[V.AI-2.2] BIAS-MANAGED TRAINING DATA**

The AI system shall verify bias-managed training data by inspection.

The inspection shall confirm training datasets document representation across applicable user classes and operational contexts, known historical biases are identified and addressed, and labeling methodology includes quality controls.

**Success Criteria:** The verification shall be considered successful when the inspection shows that training data representation is documented, historical biases are identified and addressed, and labeling quality controls are documented.

### 2.2.3 Training Dataset Validation

**[R.AI-2.3] TRAINING DATASET VALIDATION**

The AI system shall validate training datasets that could impact Safety-Critical or Mission-Critical decisions.

***Rationale:*** *Validation of training datasets prior to model development confirms that the training data bias management requirements (AI-2.2) have been met and provides documented evidence for verification closure.*

**[V.AI-2.3] TRAINING DATASET VALIDATION**

The AI system shall verify training dataset validation by inspection and analysis.

The inspection shall confirm training datasets do not contain data that contributes to biases such as underrepresented groups, historical biases, or labeling errors.

The analysis shall quantify representation across applicable user classes and operational contexts and confirm absence of known bias patterns. Where absence cannot be confirmed, the analysis shall include a risk assessment due to included bias and provide a mitigation strategy for reducing the risk.

**Success Criteria:** The verification shall be considered successful when the inspection confirms training datasets do not contain bias-contributing data, and the analysis confirms either absence of known bias patterns or documented risk assessment with mitigation strategy.

### 2.2.4 Continuous Bias Monitoring

**[R.AI-2.4] CONTINUOUS BIAS MONITORING**

The AI system shall implement continuous bias monitoring appropriate to the operational context.

***Rationale:*** *AI system bias can emerge or amplify during operations due to data drift, changing operational contexts, or feedback loops. Continuous monitoring enables early detection of bias before it impacts safety or operational success. Monitoring frequency and methods should scale with system classification and decision criticality.*

**[V.AI-2.4] CONTINUOUS BIAS MONITORING**

The AI system shall verify continuous bias monitoring by test.

The test shall confirm continuous bias monitoring is operational and metrics are maintained within approved thresholds.

**Success Criteria:** The verification shall be considered successful when the test shows that continuous bias monitoring is operational, and metrics are maintained within approved thresholds.

### 2.2.5 Bias Impact Assessment

**[R.AI-2.5] BIAS IMPACT ASSESSMENT**

The AI system shall conduct bias impact assessments for all AI decisions affecting human safety, resource allocation, mission planning, and operational procedures.

***Rationale:*** *Not all AI decisions carry equal risk of harm from bias. Impact assessments ensure that bias mitigation efforts are prioritized for decisions with the greatest potential consequences for safety and operational success.*

**[V.AI-2.5] BIAS IMPACT ASSESSMENT**

The AI system shall verify bias impact assessments by analysis and test.

The analysis shall document bias impact assessments for each decision category and ensure assessment methodology is traceable to defined criteria.

The test shall confirm assessment results inform mitigation strategies prior to deployment.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that bias impact assessments are documented for each decision category, assessment methodology is traceable to defined criteria, and results inform mitigation strategies prior to deployment.

### 2.2.6 Bias Indicator Definition

**[R.AI-2.6] BIAS INDICATOR DEFINITION**

The AI system shall define bias indicators and thresholds for each AI function.

***Rationale:*** *Bias detection requires defined metrics (indicators) and acceptance boundaries (thresholds). Without explicit definition, bias detection is subjective and unverifiable.*

**[V.AI-2.6] BIAS INDICATOR DEFINITION**

The AI system shall verify bias indicator definition by inspection.

The inspection shall confirm bias indicators are defined for each AI function, threshold values are defined and justified for each indicator, and threshold derivation rationale is documented.

**Success Criteria:** The verification shall be considered successful when the inspection shows bias indicators are defined, thresholds are defined and justified, and derivation rationale is documented.

### 2.2.7 Bias Threshold Alerting

**[R.AI-2.7] BIAS THRESHOLD ALERTING**

The AI system shall generate alerts when bias indicators exceed predefined thresholds.

***Rationale:*** *Detection without notification provides no opportunity for response. Alerts enable operators and systems to respond to detected bias conditions.*

**[V.AI-2.7] BIAS THRESHOLD ALERTING**

The AI system shall verify bias threshold alerting by test.

The test shall confirm alerts are generated when bias indicators exceed thresholds and alerts are delivered to designated recipients within specified latency limits.

**Success Criteria:** The verification shall be considered successful when the test shows alerts are generated when thresholds are exceeded and alerts are delivered within latency limits.

### 2.2.8 Bias Threshold Response

**[R.AI-2.8] BIAS THRESHOLD RESPONSE**

The AI system shall provide bias-corrected decision pathways and alternative recommendations when bias indicators exceed predefined thresholds.

***Rationale:*** *Detection and alerting without corrective action provides no protection. This requirement ensures the system has predefined corrective actions when bias is detected.*

**[V.AI-2.8] BIAS THRESHOLD RESPONSE**

The AI system shall verify bias threshold response by test.

The test shall confirm bias-corrected pathways are invoked when thresholds are exceeded and alternative recommendations are presented.

**Success Criteria:** The verification shall be considered successful when the test shows bias-corrected pathways are invoked and alternative recommendations are presented when thresholds are exceeded.

### 2.2.9 Bias Event Logging

**[R.AI-2.9] BIAS EVENT LOGGING**

The AI system shall log bias detection events, corrective actions taken, and decision rationale to support accountability and post-mission analysis.

***Rationale:*** *Logs provide accountability and enable post-mission analysis to improve future AI systems. Without comprehensive logging, bias events cannot be reconstructed for root cause analysis or used to improve subsequent model versions.*

**[V.AI-2.9] BIAS EVENT LOGGING**

The AI system shall verify bias event logging by inspection.

The inspection shall confirm logs record bias detection events with timestamps, logs capture metrics, decision context, mitigation actions, and operator inputs, logs are immutable and retrievable for post-mission analysis, and log schema documentation is complete.

**Success Criteria:** The verification shall be considered successful when the inspection shows that logs record bias detection events with timestamps, logs capture all required information, logs are immutable and retrievable, and log schema documentation is complete.

### 2.2.10 Safety-Critical Bias Prohibition

**[R.AI-2.10] SAFETY-CRITICAL BIAS PROHIBITION**

The AI system shall implement controls to prevent detected bias from adversely impacting safety-critical operations or human safety.

***Rationale:*** *Bias affecting safety-critical functions poses unacceptable risk. This requirement establishes the fundamental constraint that detected bias must be prevented from affecting safety-critical decisions, with escalation and override as the enforcement mechanism.*

**[V.AI-2.10] SAFETY-CRITICAL BIAS PROHIBITION**

The AI system shall verify safety-critical bias prohibition by analysis and test.

The analysis shall identify all AI decision points that affect safety-critical operations.

The test shall confirm that detected bias at safety-critical decision points triggers protective action preventing biased outputs from affecting safety-critical operations.

**Success Criteria:** The verification shall be considered successful when the analysis identifies safety-critical decision points and the test confirms bias detection triggers protective action at those points.

### 2.2.11 Safety-Critical Bias Escalation

**[R.AI-2.11] SAFETY-CRITICAL BIAS ESCALATION**

The AI system shall escalate any bias detection that would adversely impact safety-critical operations or human safety for immediate human review and potential system override with documented rationale, or, where the approved autonomy level (AI-9.1) or disconnected operations (AI-4.6) preclude timely human

review, shall execute the defined fallback response with the escalation logged for human review at the next available opportunity.

***Rationale:*** *Bias affecting safety-critical functions requires immediate human intervention. Autonomous continuation in the presence of detected safety-critical bias could result in harm. This requirement ensures human authority is maintained over safety-critical AI decisions per AI-9.1 (Human Decision Authority).*

**[V.AI-2.11] SAFETY-CRITICAL BIAS ESCALATION**

The AI system shall verify safety-critical bias escalation by test.

The test shall simulate safety-critical bias scenarios and confirm mandatory human review states are triggered, or the defined fallback response executes where approved autonomy or disconnected operations preclude timely human review, autonomous execution is blocked until review or fallback completes, system override capability is functional, and escalation logs capture review decisions and rationale.

**Success Criteria:** The verification shall be considered successful when the test shows that safety-critical bias scenarios trigger mandatory human review, or the defined fallback response where human review is precluded, autonomous execution is blocked until review or fallback completes, system override capability is functional and documented, and escalation logs capture review decisions and rationale.

### 2.2.12 Bias Edge Case Definition

**[R.AI-2.12] BIAS EDGE CASE DEFINITION**

The AI system shall define edge-case datasets and stress scenarios for bias validation.

***Rationale:*** *Edge cases and stress conditions represent operational boundaries where bias is most likely to emerge but least likely to be represented in training data. Explicit definition of edge cases ensures systematic coverage during validation rather than ad-hoc testing.*

**[V.AI-2.12] BIAS EDGE CASE DEFINITION**

The AI system shall verify bias edge case definition by inspection.

The inspection shall confirm edge-case datasets are defined and documented, stress scenarios are defined and documented, and edge cases address operational boundaries and conditions underrepresented in training data.

**Success Criteria:** The verification shall be considered successful when the inspection shows that edge-case datasets and stress scenarios are defined and documented.

### 2.2.13 Bias Edge Case Validation

**[R.AI-2.13] BIAS EDGE CASE VALIDATION**

The AI system shall validate bias performance under defined edge cases and stress scenarios.

***Rationale:*** *Bias may not manifest under nominal conditions but emerge at operational boundaries or under stress. Validation across edge cases ensures the system maintains performance within approved bias thresholds even in conditions that may not be well-represented in training data.*

**[V.AI-2.13] BIAS EDGE CASE VALIDATION**

The AI system shall verify bias edge case validation by test.

The test shall confirm bias metrics remain within acceptable bounds under defined edge cases and stress scenarios, or threshold exceedances trigger defined mitigation.

**Success Criteria:** The verification shall be considered successful when the test shows bias metrics remain within acceptable bounds under edge cases and stress scenarios, or exceedances trigger defined mitigation.

## 2.3 AI-3: ML TEST COVERAGE

**[R.AI-3] ML TEST COVERAGE**

The AI system will demonstrate adequate test coverage for machine learning models using methods appropriate to data-driven systems where traditional code coverage metrics do not apply.

Traditional software testing relies on code coverage metrics (statement coverage, branch coverage, MC/DC) to demonstrate that the software has been adequately exercised. These metrics are meaningless for ML models: there is no “code” to cover in a neural network’s weights. AI systems require alternative coverage concepts based on input space coverage, operational scenario coverage, and behavioral testing. This requirement addresses the “no test coverage equivalent” gap identified in Section 1.2.4.2.

**Related Domain Standards**

- General: ISO/IEC/IEEE 29119-11:2020 (Testing of AI-based systems); ISO/IEC 5338:2023 §6.4
- Aviation: DO-178C §6 (Verification Process); EUROCAE ED-324/SAE AS6983 (AI/ML aviation)
- Automotive: ISO 26262-6:2018 §9 (Software unit verification); ISO 21448:2022 §7 (Evaluation)
- Medical: IEC 62304:2006+A1:2015 §5.7 (Software system testing)
- EU: EU AI Act Article 15 (Accuracy, robustness and cybersecurity)
- Cross-Domain: NIST AI RMF MEASURE-2; IEEE 7010-2020

### 2.3.1 Test Coverage Metric Definition

**[R.AI-3.1] TEST COVERAGE METRIC DEFINITION**

The AI system shall define test coverage metrics appropriate to each ML model type.

***Rationale:*** *Different ML model types require different coverage approaches. For image classifiers, coverage might focus on variations in lighting, angle, and occlusion. For time-series predictors, coverage might address temporal patterns, seasonality, and anomalies. Metrics must be justified based on model architecture and mission criticality to ensure testing is meaningful rather than merely comprehensive in volume.*

**[V.AI-3.1] TEST COVERAGE METRIC DEFINITION**

The AI system shall verify test coverage metric definition by inspection.

The inspection shall confirm test coverage metrics are defined for each ML model, metrics are justified based on model type and mission criticality, coverage metrics address input space, edge cases, and operational scenarios, and rationale for metric selection is documented.

**Success Criteria:** The verification shall be considered successful when the inspection shows that test coverage metrics are defined for each ML model, metrics are justified based on model type and criticality, coverage addresses input space, edge cases, and operational scenarios, and rationale is documented.

### 2.3.2 Test Coverage Threshold Achievement

**[R.AI-3.2] TEST COVERAGE THRESHOLD ACHIEVEMENT**

The AI system shall achieve defined test coverage thresholds prior to operational deployment.

***Rationale:*** *Defining metrics without achieving them provides no assurance. Coverage thresholds should be established based on system classification: Safety-Critical AI requires higher coverage thresholds than Operational Support AI. Coverage gaps must be identified and dispositioned with documented rationale, either by additional testing or by risk acceptance with mitigation.*

**[V.AI-3.2] TEST COVERAGE THRESHOLD ACHIEVEMENT**

The AI system shall verify test coverage threshold achievement by analysis and test.

The analysis shall confirm coverage thresholds are defined and documented with justification based on system classification and mission risk tolerance.

The test shall execute test cases to achieve required coverage levels, identify and document any coverage gaps, and disposition gaps through additional testing or documented risk acceptance.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that coverage thresholds are defined and documented, test execution achieves required coverage levels, coverage gaps are identified and dispositioned, and coverage reports are retained as verification evidence.

### 2.3.3 Operational Input Distribution Testing

**[R.AI-3.3] OPERATIONAL INPUT DISTRIBUTION TESTING**

The AI system shall test ML models across representative operational input distributions.

***Rationale:*** *ML models perform well on data similar to their training data but may fail on data that differs from training distributions. Testing must cover the full range of inputs expected during actual operations, including nominal conditions, off-nominal conditions, and mission phase transitions. Gaps between test and operational distributions represent untested regions where model behavior is uncertain. Operational input distributions shall be derived from the declared ODD (AI-1.0).*

**[V.AI-3.3] OPERATIONAL INPUT DISTRIBUTION TESTING**

The AI system shall verify operational input distribution testing by test.

The test shall confirm operational input distributions are characterized and documented based on mission analysis and operational scenarios, test datasets represent the full range of expected operational inputs, model performance is validated across the operational input space, and gaps between test and operational distributions are identified and mitigated through additional test data collection, synthetic data generation, or documented risk acceptance.

**Success Criteria:** The verification shall be considered successful when the test shows that operational input distributions are characterized, test datasets represent the full operational range, model performance is validated across the input space, and distribution gaps are identified and mitigated.

### 2.3.4 Boundary and Edge Case Testing

**[R.AI-3.4] BOUNDARY AND EDGE CASE TESTING**

The AI system shall perform boundary and edge case testing for all ML models.

***Rationale:*** *ML models often fail at input boundaries: values at the edges of the training distribution or combinations of inputs not well-represented in training data. Edge cases are particularly important for safety-critical functions because they represent conditions where model behavior is least certain. Systematic boundary testing ensures the model behaves acceptably at operational limits.*

**[V.AI-3.4] BOUNDARY AND EDGE CASE TESTING**

The AI system shall verify boundary and edge case testing by test.

The test shall confirm input boundaries are defined for each ML model based on operational constraints and physical limits, edge cases are identified and documented through systematic analysis (e.g., equivalence partitioning, boundary value analysis adapted for ML), model behavior at boundaries is tested and acceptable, and edge case test results are recorded with pass/fail criteria.

**Success Criteria:** The verification shall be considered successful when the test shows that input boundaries are defined for each ML model, edge cases are identified and documented, model behavior at boundaries is tested and acceptable, and test results are recorded with pass/fail criteria.

## 2.4 AI-4: CONTINUOUS VALIDATION

**[R.AI-4] CONTINUOUS VALIDATION**

The AI system will implement continuous validation processes to ensure model performance remains within acceptable bounds throughout operational deployment.

AI model performance can degrade during operations due to data drift, concept drift, or environmental changes. Continuous validation detects performance degradation before it impacts safety or operational success, enabling timely corrective action such as model retraining, threshold adjustment, or fallback to manual operations.

**Related Domain Standards**

- General: ISO/IEC 5338:2023 §6.4.14; ISO/IEC 22989:2022 5.11.9
- Aviation: FAA AI Safety Roadmap (operational monitoring)
- Automotive: ISO 21448:2022 §13 (Field monitoring); UNECE WP.29 R157 (ALKS) §5.2 `(Dynamic Driving Task)`
- Medical: FDA PCCP Guidance (2023); FDA AI/ML-Based SaMD Action Plan §5 (Real-World Performance)
- EU: EU AI Act Article 17 (Quality management system); EU AI Act Article 72 (Post-market monitoring)
- Cross-Domain: NIST AI RMF MEASURE-3; NIST AI RMF MEASURE-4

### 2.4.1 Data Drift Monitoring

**[R.AI-4.1] DATA DRIFT MONITORING**

The AI system shall monitor for data drift that could cause inaccurate decisions or predictions.

***Rationale:*** *Data drift occurs when the statistical characteristics of production data diverge from the training data distribution. Sensor degradation, environmental changes, or novel operational scenarios can cause inputs to shift from what the model was trained on, leading to degraded prediction accuracy. Without drift detection, performance degradation may go unnoticed until a critical failure occurs.*

**[V.AI-4.1] DATA DRIFT MONITORING**

The AI system shall verify data drift monitoring by test.

The test shall confirm data drift detection mechanisms are implemented and operational, drift is measured against baseline training data distributions using defined statistical methods (e.g., KL divergence, population stability index), drift metrics are logged with timestamps, and alerts are generated when drift exceeds defined thresholds.

**Success Criteria:** The verification shall be considered successful when the test shows that data drift detection mechanisms are operational, drift is measured against baseline distributions, drift metrics are logged with timestamps, and alerts are generated when drift exceeds defined thresholds.

### 2.4.2 Performance Threshold Maintenance

**[R.AI-4.2] PERFORMANCE THRESHOLD MAINTENANCE**

The AI system shall maintain model performance above defined thresholds throughout operation.

***Rationale:*** *AI model performance can degrade over time even without data drift through concept drift, model staleness, or accumulated edge cases. Continuous performance monitoring against defined thresholds provides early warning of degradation before it impacts operational success or safety. Threshold values should be derived from mission risk tolerance per Section 5.1 (Threshold Selection Guidance).*

**[V.AI-4.2] PERFORMANCE THRESHOLD MAINTENANCE**

The AI system shall verify performance threshold maintenance by analysis and test.

The analysis shall confirm performance thresholds are defined and justified based on mission risk tolerance, and that performance metrics are appropriate to the AI function (e.g., accuracy, precision, recall, F1 score, domain-specific metrics).

The test shall confirm performance metrics are continuously computed during operation, performance degradation below threshold triggers defined mitigation response, and performance logs are maintained for trend analysis.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that performance thresholds are defined and justified, performance metrics are continuously computed, degradation below threshold triggers mitigation response, and performance logs support trend analysis.

### 2.4.3 Model Maintenance Criteria

**[R.AI-4.3] MODEL MAINTENANCE CRITERIA**

The AI system shall define criteria for model maintenance, retraining, or version updates when performance deviates beyond acceptable bounds.

***Rationale:*** *When continuous monitoring detects performance degradation, the system must have predefined response procedures. Without documented criteria for when and how to update models, operators may allow degraded performance*

*to continue or may apply ad-hoc fixes that introduce new risks. This requirement ensures maintenance decisions are systematic and traceable. Model maintenance changes also invoke the model integrity controls established under AI-7.4 and the deployment format validation established in AI-4.7; these three control points apply in concert whenever a deployed model is updated.*

**[V.AI-4.3] MODEL MAINTENANCE CRITERIA**

The AI system shall verify model maintenance criteria by inspection.

The inspection shall confirm retraining and update trigger criteria are documented, version control procedures exist for model updates, regression testing requirements are defined for updated models, and change history is maintained with rationale for each update.

**Success Criteria:** The verification shall be considered successful when the inspection shows that retraining/update trigger criteria are documented, version control procedures exist, regression testing requirements are defined, and change history with rationale is maintained.

### 2.4.4 Output Validity Boundaries

**[R.AI-4.4] OUTPUT VALIDITY BOUNDARIES**

The AI system shall define and enforce boundaries for output data validity.

***Rationale:*** *AI systems can produce outputs that are technically valid (within data type constraints) but operationally invalid (physically impossible, outside safe operating ranges, or inconsistent with known constraints). Output boundary enforcement provides a safety envelope that catches erroneous outputs regardless of the internal cause, whether from drift, adversarial inputs, or model failure. Operational limits that bound acceptable outputs shall be derived from the Performance Envelope and Exclusions declared under AI-1.0.*

**[V.AI-4.4] OUTPUT VALIDITY BOUNDARIES**

The AI system shall verify output validity boundaries by test.

The test shall confirm output boundaries are defined for each AI function based on physical constraints, operational limits, and safety margins. The test shall inject inputs designed to produce boundary-case outputs and confirm outputs exceeding boundaries are flagged or rejected, boundary violations are logged with context, and out-of-bounds outputs do not propagate to downstream systems without human review or invocation of defined fallback behavior.

**Success Criteria:** The verification shall be considered successful when the test shows that output boundaries are defined for each AI function, outputs exceeding boundaries are flagged or rejected, boundary violations are logged, and out-of-bounds outputs do not propagate without human review or fallback behavior.

### 2.4.5 Periodic Model Validation

**[R.AI-4.5] PERIODIC MODEL VALIDATION**

The AI system shall perform model validation at defined intervals and maintain baseline validation scores.

***Rationale:*** *Continuous monitoring of individual metrics may not capture gradual, systemic degradation. Periodic comprehensive validation against baseline scores provides a systemic check that the model continues to perform as originally validated. Validation frequency should scale with system classification and operational tempo.*

**[V.AI-4.5] PERIODIC MODEL VALIDATION**

The AI system shall verify periodic model validation by test.

The test shall confirm validation frequency is defined and justified based on system classification and mission profile, validation executes automatically at specified intervals, validation scores are compared against baseline thresholds, and validation failures trigger defined response procedures.

**Success Criteria:** The verification shall be considered successful when the test shows that validation frequency is defined and justified, validation executes automatically at specified intervals, validation scores are compared against baselines, and validation failures trigger defined response procedures.

### 2.4.6 Operational Validation Without Ground Truth

**[R.AI-4.6] OPERATIONAL VALIDATION WITHOUT GROUND TRUTH**

The AI system shall define and justify proxy validation mechanisms for operational environments where ground truth data is unavailable or inaccessible.

***Rationale:*** *AI-4 requirements assume continuous validation against known correct outputs. This assumption holds during development and in connected operational environments where labeled data or authoritative reference sources remain accessible. However, many safety-critical operational contexts lack ground truth entirely: disconnected systems, autonomous platforms, deep space missions, submarine operations, or any environment where the "correct answer" cannot be independently determined in real time. In these environments, validation must rely*

*on proxy mechanisms that provide evidence of continued reliable performance without direct comparison to ground truth. These proxies include output boundary enforcement (per AI-4.4), statistical drift detection against baseline distributions (per AI-4.1), independent cross-validation using dissimilar methods (per AI-5.5), and temporal consistency checks against the system's own recent output history. The selection and sufficiency of proxy mechanisms are context dependent. A system operating autonomously for hours faces different validation challenges than one operating autonomously for months. Projects shall document which proxy mechanisms apply, justify why those proxies are sufficient for the specific operational context and duration, and define the conditions under which proxy validation is no longer adequate and operations should be suspended or degraded.*

**[V.AI-4.6] OPERATIONAL VALIDATION WITHOUT GROUND TRUTH**

The AI system shall verify operational validation without ground truth by inspection and analysis.

The inspection shall confirm the operational environment is characterized with respect to ground truth availability, proxy validation mechanisms are identified and documented for each AI function operating without ground truth, and the justification for proxy sufficiency is documented with traceability to operational context, mission duration, and system classification.

The analysis shall confirm selected proxy mechanisms collectively address performance degradation detection, output validity assurance, and behavioral consistency monitoring; that the analysis identifies conditions under which proxy validation is insufficient and documents the required system response (degraded mode, human escalation, or operational suspension); and that proxy validation coverage is assessed against the operational timeline to confirm adequacy for the expected duration of ground truth unavailability.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that the operational environment is characterized for ground truth availability, proxy mechanisms are documented and justified, proxy coverage addresses degradation detection, output validity, and behavioral consistency, insufficiency conditions are defined with required system response, and proxy adequacy is confirmed for the expected operational duration.

### 2.4.7 Deployment Format Validation

**[R.AI-4.7] DEPLOYMENT FORMAT VALIDATION**

The AI system shall revalidate deployment-format models against the original trained model before operational use, where deployment-format models are

produced through transformations such as quantization, pruning, knowledge distillation, or framework conversion.

***Rationale:*** *Model transformations applied for deployment can introduce numerical, accuracy, or behavioral drift relative to the originally validated model. Without explicit revalidation, an artifact that passed all training-time verification may exhibit degraded or unsafe behavior at runtime, a gap not closed by retraining triggers (AI-4.3) or load-time integrity checks (AI-7.4). Side-by-side comparison on representative data establishes that the deployed artifact retains its validated properties.*

Common transformations requiring revalidation include:

- **Quantization:** reducing model precision (e.g., FP32 to INT8) to enable execution on resource-constrained hardware.
- **Pruning:** removing model weights, neurons, or layers to reduce computational cost.
- **Knowledge distillation:** training a smaller model to approximate the behavior of a larger original.
- **Format conversion:** converting between frameworks (e.g., PyTorch to ONNX to TensorRT) for runtime compatibility.

**[V.AI-4.7] DEPLOYMENT FORMAT VALIDATION**

The AI system shall verify deployment format validation by test.

The test shall demonstrate that the deployment-format model meets all performance thresholds established for the original trained model under representative test conditions. Performance degradation attributable to the transformation shall be documented and accepted within margins established per Section 5.1 (Threshold Selection Guidance), or the transformation shall be revised. Verification evidence shall include side-by-side performance comparisons on representative test datasets covering the declared Operational Design Domain. Deployment Format Validation operates as a control point complementary to Model Maintenance Criteria (AI-4.3) and Model Integrity (AI-7.4); a deployed transformation shall satisfy all three before operational use.

**Success Criteria:** The verification shall be considered successful when the test shows that the deployment-format model meets or exceeds all performance thresholds established for the original trained model under representative test conditions, performance deltas are documented and accepted per Section 5.1, and

the verification evidence is retained per AI-1.8 (Classification Compliance Audit Trail).

**Related Domain Standards**

- Aerospace: NPR 7150.2D §4 (Software Lifecycle Processes including Verification and Validation); NASA-STD-8739.8B (Software Assurance)
- Aviation: DO-178C §6.4 (Software Testing) and §11.5 (Verification of Verification Process Results)
- Automotive: ISO 26262-6:2018 §11 (Testing of the embedded software); ISO 21448:2022 §12 (Evaluation of the achievement of the SOTIF)
- Medical: IEC 62304:2006+A1:2015 §5.7 (Software system testing) and §5.8 (Software release)
- Industrial: ISO 13849-1 (Functional safety of control systems)

## 2.5 AI-5: HALLUCINATION PREVENTION

### [R.AI-5] HALLUCINATION PREVENTION

The AI system will implement mechanisms to detect, prevent, and mitigate AI outputs that are fabricated, unsupported by input data, or inconsistent with operational reality.

***Rationale:*** *AI systems, particularly neural networks and generative models, can produce outputs that appear valid but are completely fabricated or contradicted by physical reality. NIST AI 600-1 defines hallucination as* ***"the production of confidently stated but erroneous or false content."*** Unlike traditional software bugs that produce obviously wrong outputs (e.g., null values, exceptions), hallucinations appear plausible and may not trigger conventional error handling. This requirement addresses the "unpredictable outputs" gap identified in Section 1.2.4.2.

**Related Domain Standards**

- General: NIST AI 600-1 (Generative AI Profile, 2024) §2.2 (Confabulation); ISO/IEC 22989:2022 5.11
- Automotive: ISO 21448:2022 §7 (Functional insufficiencies and triggering conditions)

- Medical: FDA Clinical Decision Support Software Guidance (January 2026) §IV(3)-(4) (Interpretation of Criteria; supports independent review of basis for recommendations)
- EU: EU AI Act Article 15 (Robustness)
- Cross-Domain: NIST AI RMF MEASURE-2.7

### 2.5.1 Hallucination Criteria Definition

**[R.AI-5.1] HALLUCINATION CRITERIA DEFINITION**

The AI system shall define what constitutes a hallucination within the operational context.

***Rationale:*** *Hallucination is context-dependent. An output that would be a hallucination in one domain may be valid in another. For example, a predicted temperature of -270°C would be valid near absolute zero but a hallucination for cabin temperature. Defining mission-specific hallucination criteria enables automated detection and ensures operators understand when an output should be considered unreliable. Ground truth baselines provide authoritative reference points against which outputs can be validated.*

**[V.AI-5.1] HALLUCINATION CRITERIA DEFINITION**

The AI system shall verify hallucination criteria definition by inspection.

The inspection shall confirm hallucination criteria are defined for each AI function, criteria include outputs outside expected bounds, contradictory outputs, and outputs unsupported by inputs, definition is traceable to mission-specific operational constraints, examples of hallucination versus valid edge-case outputs are documented, and ground truth baselines are established for each AI function including authoritative data sources, physical constraints and conservation laws, cross-validation with independent sensors or systems, and operator-verified reference datasets.

**Success Criteria:** The verification shall be considered successful when the inspection shows that hallucination criteria are defined for each AI function, criteria address bounds violations, contradictions, and unsupported outputs, definitions are traceable to operational constraints, examples distinguish hallucinations from valid edge cases, and ground truth baselines are established.

### 2.5.2 Hallucination Detection

**[R.AI-5.2] HALLUCINATION DETECTION**

The AI system shall implement detection mechanisms for hallucinated outputs.

***Rationale:*** *Once hallucination criteria are defined, automated detection mechanisms can identify suspect outputs before they propagate to downstream systems or influence decisions. Detection should occur at inference time, not just during post-mission analysis, to prevent real-time operational impact. Detection mechanisms may include consistency checks, physical constraint validation, cross-sensor comparison, or confidence thresholding.*

**[V.AI-5.2] HALLUCINATION DETECTION**

The AI system shall verify hallucination detection by test.

The test shall confirm detection algorithms are implemented for each AI function, detection mechanisms identify outputs meeting hallucination criteria, detection is performed prior to output delivery to downstream systems, and detection performance is validated against known hallucination test cases including injected hallucinations that violate physical constraints, contradict sensor inputs, or fall outside valid operating ranges.

**Success Criteria:** The verification shall be considered successful when the test shows that detection algorithms are implemented for each AI function, detection identifies outputs meeting hallucination criteria, detection occurs prior to downstream delivery, and detection performance is validated against known test cases.

### 2.5.3 Hallucination Response

**[R.AI-5.3] HALLUCINATION RESPONSE**

The AI system shall prevent hallucinated outputs from being acted upon without human review or defined fallback behavior appropriate to the AI function's criticality.

***Rationale:*** *Detection without response provides no protection. When a potential hallucination is detected, the system must prevent autonomous action based on that output, either escalating to human review or invoking defined fallback behavior, consistent with the response pattern established for out-of-distribution inputs (AI-6.3). This ensures either human judgment or a validated safe response is applied before potentially erroneous AI outputs influence safety-critical decisions. The response should be proportionate to the criticality of the AI function and the confidence of hallucination detection.*

**[V.AI-5.3] HALLUCINATION RESPONSE**

The AI system shall verify hallucination response by test.

The test shall inject hallucinated outputs and confirm detected hallucinations are flagged and quarantined, autonomous execution is blocked for flagged outputs, human review workflow is triggered or defined fallback behavior is invoked for flagged outputs as required by system classification, and override requires documented rationale.

**Success Criteria:** The verification shall be considered successful when the test shows that detected hallucinations are flagged and quarantined, autonomous execution is blocked, human review workflow is triggered or defined fallback behavior is invoked as required by system classification, and override requires documented rationale.

### 2.5.4 Hallucination Logging

**[R.AI-5.4] HALLUCINATION LOGGING**

The AI system shall log all detected hallucinations for post-mission analysis.

***Rationale:*** *Comprehensive logging of hallucination events enables post-mission analysis to identify patterns, root causes, and potential model improvements. Logs support trend analysis to determine if hallucination frequency is increasing (indicating model degradation) and provide evidence for retraining decisions. Logging also supports accountability and incident investigation.*

**[V.AI-5.4] HALLUCINATION LOGGING**

The AI system shall verify hallucination logging by inspection.

The inspection shall confirm hallucination events are logged with timestamps, inputs, and outputs, logs capture detection method and confidence, logs capture human review decision and rationale, and logs are retrievable for trend analysis and model improvement.

**Success Criteria:** The verification shall be considered successful when the inspection shows that hallucination events are logged with timestamps, inputs, and outputs, logs capture detection method, confidence, and human review decisions, and logs are retrievable for trend analysis.

### 2.5.5 Independent Output Validation

**[R.AI-5.5] INDEPENDENT OUTPUT VALIDATION**

The AI system shall validate AI outputs affecting safety-critical or mission-critical functions against an independent source prior to execution.

***Rationale:*** *For safety-critical functions, relying solely on the AI system's internal consistency checks is insufficient; an AI can produce confidently wrong outputs that pass its own validation. Independent validation using dissimilar methods, sensors, or processing provides a defense-in-depth layer that can catch errors the primary AI cannot detect. The independent source must not share common failure modes with the primary AI to ensure that a single fault (e.g., corrupted sensor data, adversarial input, systematic model error) cannot defeat both the primary output and its validation.*

### [V.AI-5.5] INDEPENDENT OUTPUT VALIDATION

The AI system shall verify independent output validation by analysis and test.

The analysis shall identify independent validation sources for each safety-critical AI output and confirm validation sources use dissimilar methods, sensors, or processing from the primary AI. The analysis shall demonstrate that independent validation does not share common failure modes with the primary AI, including common training data sources, common sensor inputs, common processing hardware, or common software dependencies.

The test shall inject discrepancies between AI output and independent validation sources and confirm discrepancies trigger human review or defined fallback behavior, the system does not execute safety-critical actions when validation fails, and validation latency meets operational timing requirements.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that independent validation sources are identified for each safety-critical AI output, validation sources use dissimilar methods from the primary AI, discrepancies trigger human review or fallback behavior, and independent validation does not share common failure modes with the primary AI.

## 2.6 AI-6: OUT-OF-DISTRIBUTION DETECTION

### [R.AI-6] OUT-OF-DISTRIBUTION DETECTION

The AI system will detect and appropriately handle inputs that fall outside the distribution of training data.

ML models are only reliable within the distribution of data they were trained on. When inputs differ significantly from training data due to sensor anomalies, novel operational scenarios, or environmental conditions not anticipated during development, model outputs become unreliable and potentially dangerous. Traditional software has defined input/output bounds that can be checked explicitly; AI systems are vulnerable to OOD inputs that may appear valid but produce unreliable predictions. This requirement addresses the "vulnerable to OOD

inputs" gap identified in Section 1.2.4.2, ensuring the system recognizes when it is operating outside its validated envelope.

**Related Domain Standards**

- General: ISO/IEC 23053:2022; ISO/IEC 5338:2023 §6.4.14
- Aviation: FAA AI Safety Roadmap (operational envelope)
- Automotive: ISO 21448:2022 §6–§8; ISO 34503 (ODD boundary definitions)
- Medical: FDA AI/ML-Based SaMD Action Plan (real-world performance monitoring)
- EU: EU AI Act Article 15 (Robustness to errors)
- Cross-Domain: NIST AI RMF MEASURE-2.6

### 2.6.1 Training Distribution Characterization

**[R.AI-6.1] TRAINING DISTRIBUTION CHARACTERIZATION**

The AI system shall characterize the training data distribution for each ML model.

***Rationale:*** *OOD detection requires knowing what* ***"in-distribution"*** *looks like. Characterizing the training data distribution provides the baseline against which operational inputs are compared. Documentation should include statistical measures (mean, variance, range, density estimates) that enable automated OOD detection at runtime. Without explicit distribution characterization, OOD detection cannot be implemented systematically. The distribution shall be characterized within the scope declared under AI-1.0; characterization need not extend beyond the declared Operational Environment and Subject Population.*

**[V.AI-6.1] TRAINING DISTRIBUTION CHARACTERIZATION**

The AI system shall verify training distribution characterization by inspection.

The inspection shall confirm training data distribution is documented for each model, distribution boundaries and characteristics are defined including statistical measures (e.g., mean, variance, covariance, range, density estimates), documentation includes statistical measures of distribution properties sufficient to support runtime OOD detection, and distribution characterization is traceable to training datasets.

**Success Criteria:** The verification shall be considered successful when the inspection shows that training data distribution is documented for each model,

distribution boundaries and characteristics are defined, statistical measures are documented, and characterization is traceable to training datasets.

### 2.6.2 Runtime OOD Detection

**[R.AI-6.2] RUNTIME OOD DETECTION**

The AI system shall implement mechanisms to detect out-of-distribution inputs at runtime.

***Rationale:*** *OOD detection must occur during operation, not just during development testing. Runtime detection mechanisms compare incoming inputs against the characterized training distribution and flag inputs that fall outside defined boundaries. Detection methods may include statistical distance measures (Mahalanobis distance), density estimation, ensemble disagreement, or dedicated OOD detection models. Detection must meet operational timing requirements to enable timely response.*

**[V.AI-6.2] RUNTIME OOD DETECTION**

The AI system shall verify runtime OOD detection by test.

The test shall confirm OOD detection mechanisms are implemented for each ML model, detection correctly identifies inputs outside training distribution using defined statistical methods, detection performance is validated against known OOD test cases including both synthetic OOD inputs and realistic operational scenarios, and detection latency meets operational timing requirements for the AI function.

**Success Criteria:** The verification shall be considered successful when the test shows that OOD detection mechanisms are implemented for each model, detection correctly identifies OOD inputs, detection is validated against known test cases, and latency meets operational timing requirements.

### 2.6.3 OOD Response

**[R.AI-6.3] OOD RESPONSE**

The AI system shall prevent autonomous action on out-of-distribution inputs without human review or defined fallback behavior.

***Rationale:*** *Detecting OOD inputs without a defined response provides no protection. When an input is flagged as OOD, the system must either escalate to human review or invoke predefined fallback behavior appropriate to the AI function's criticality. Autonomous execution based on potentially unreliable OOD predictions could lead to hazardous outcomes. The response should be*

*proportionate to system classification: Safety-Critical AI may require mandatory human review, while Operational Support AI may use automated fallback.*

**[V.AI-6.3] OOD RESPONSE**

The AI system shall verify OOD response by test.

The test shall inject OOD inputs and confirm OOD inputs trigger defined response procedures, autonomous execution is blocked or fallback behavior is invoked, human review workflow is available for OOD conditions when required by system classification, and system behavior for OOD inputs is documented and tested across all defined OOD scenarios.

**Success Criteria:** The verification shall be considered successful when the test shows that OOD inputs trigger defined response procedures, autonomous execution is blocked or fallback is invoked, human review workflow is available when required, and OOD behavior is documented and tested.

### 2.6.4 OOD Event Logging

**[R.AI-6.4] OOD EVENT LOGGING**

The AI system shall log all out-of-distribution detection events.

***Rationale:*** *OOD events provide valuable information about operational conditions that differ from training assumptions. Logging enables post-mission analysis to identify patterns, assess whether the operational environment has shifted from training assumptions, and inform decisions about model retraining or operational envelope expansion. OOD logs also support trend analysis to detect gradual distribution shift that might indicate data drift.*

**[V.AI-6.4] OOD EVENT LOGGING**

The AI system shall verify OOD event logging by inspection.

The inspection shall confirm OOD events are logged with timestamps and input data (or input characterization sufficient for analysis), logs capture detection confidence and method, logs capture system response and operator actions, and logs are retrievable for trend analysis and model improvement.

**Success Criteria:** The verification shall be considered successful when the inspection shows that OOD events are logged with timestamps and input data, logs capture detection confidence, method, and system response, and logs are retrievable for analysis.

## 2.7 AI-7: ADVERSARIAL ROBUSTNESS

### [R.AI-7] ADVERSARIAL ROBUSTNESS

The AI system will implement mechanisms to detect, prevent, and mitigate adversarial attacks that could compromise AI integrity, manipulate outputs, or degrade system performance.

AI systems are vulnerable to attack vectors that do not exist for traditional software. Adversarial attacks can manipulate training data (poisoning), craft inputs that cause misclassification (adversarial examples), extract model parameters (model theft), or reconstruct sensitive training data (model inversion). Adversarial compromise could affect safety or operational success. This requirement addresses the "vulnerable to adversarial attacks" gap identified in Section 1.2.4.

AI-specific adversarial robustness does not replace foundational cybersecurity engineering; it extends it. The controls in AI-7.1 through AI-7.7 address attack vectors unique to AI systems, and they assume an underlying cybersecurity posture for the system and for the environment in which it is developed, built, and deployed. That posture is a prerequisite of certification rather than assessed scope: the applicant declares its foundational cybersecurity baseline by attestation or by reference to existing conformity evidence, Safety Critical Labs records the declaration, and the declared baseline is carried on the certificate as an element of the certified configuration (Section 1.6). NIST SP 800-53 Rev. 5 (Security and Privacy Controls for Information Systems and Organizations), in particular the System and Information Integrity (SI) and System and Communications Protection (SC) control families, and NIST SP 800-218 (Secure Software Development Framework, SSDF) for the ML pipeline code, training infrastructure, and deployment tooling, are representative baselines and are listed with the related domain standards below; this framework does not levy them. AI-7 sub-requirements complement rather than duplicate the declared baseline: data poisoning protection (AI-7.1) extends integrity controls to training data; model integrity (AI-7.4) extends them to model artifacts; supply chain security (AI-7.6) extends secure development practices to AI-specific dependencies including pre-trained models and ML frameworks; and artifact load safety (AI-7.7) extends them to the loading of every artifact the system admits.

**Related Domain Standards**

- General: NIST AI 100-2e2023 (Adversarial ML Taxonomy); ISO/IEC 5338:2023 §6.4.3
- Aviation: DO-326A (Airworthiness Security Process); DO-356A (Airworthiness Security Methods); FAA AI Safety Roadmap

- Automotive: ISO/SAE 21434:2021 (Road vehicles cybersecurity); UNECE WP.29 R155 (Cyber security management)
- Medical: FDA Premarket Cybersecurity Guidance (2023); FDA Postmarket Cybersecurity Guidance (2016)
- EU: EU AI Act Article 15 (Cybersecurity); EU Cyber Resilience Act (EU Cyber Resilience Act article reference held)
- Cross-Domain: NIST AI RMF MANAGE-2.2; NIST SP 800-218 (Secure Software Development Framework); NIST SP 800-53 Rev. 5 SI family

### 2.7.1 Data Poisoning Protection

**[R.AI-7.1] DATA POISONING PROTECTION**

The AI system shall implement protections against training data poisoning attacks.

***Rationale:*** *Data poisoning injects malicious samples into training datasets to influence model behavior in attacker-controlled ways. Poisoned models may appear to function normally on most inputs but produce incorrect outputs on specific attacker-chosen inputs. Poisoning could create backdoors that trigger dangerous behavior under specific conditions. Protection requires authenticating data sources, validating data integrity, and detecting statistically anomalous samples.*

**[V.AI-7.1] DATA POISONING PROTECTION**

The AI system shall verify data poisoning protection by analysis and test.

The analysis shall confirm training data sources are authenticated and validated through documented provenance, and data integrity checks detect unauthorized modifications.

The test shall confirm anomaly detection identifies statistically suspicious training samples, and poisoning attack scenarios are tested against the poisoning attack pattern set declared by the project in its verification evidence (Appendix C.2), for example backdoor triggers, label flipping, and gradient-based poisoning. The declared set shall be documented with its derivation basis (the threat model, the system classification, and the declared Operational Design Domain); its adequacy is adjudicated during certification assessment in the same manner as threshold values (Section 1.5), and the declared set is recorded as an element of the certified configuration (Section 1.6).

**Success Criteria:** The verification shall be considered successful when the analysis and test show that training data sources are authenticated, data integrity checks

detect modifications, anomaly detection identifies suspicious samples, and poisoning attack scenarios are tested against the declared attack pattern set.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SI-7 (Software, firmware, and information integrity); ISO/IEC 27001:2022 A.8.28 (Secure coding)
- AI-Specific: NIST AI 100-2e2023 §2.3 (Poisoning Attacks and Mitigations); MITRE ATLAS AML.T0020 (Poison training data)
- Data Integrity: ISO/IEC 5259-3:2024 (Data quality management, verification); ISO 8000-61:2016 (Data quality management process reference model)
- Aviation: RTCA DO-326A §3.5 (Security risk assessment); SAE ARP 4754A §4 (Development assurance)
- Automotive: ISO/SAE 21434:2021 §15 (Threat analysis and risk assessment methods)
- Cross-Domain: NIST AI RMF MANAGE-2.1 (Risk response to identified AI risks); EU AI Act Article 15(4) (Resilience against attempts to manipulate training data)

### 2.7.2 Adversarial Input Protection

**[R.AI-7.2] ADVERSARIAL INPUT PROTECTION**

The AI system shall implement protections against adversarial input manipulation at runtime.

***Rationale:*** *Adversarial inputs are carefully crafted to cause model misclassification while appearing valid to human observers. Small perturbations imperceptible to humans can cause dramatic changes in model outputs. Attack patterns in current use include FGSM (Fast Gradient Sign Method), PGD (Projected Gradient Descent), and C&W (Carlini & Wagner) attacks; the set a system is tested against is declared by the project and reviewed during assessment. Runtime protection requires input validation, perturbation detection, and verification that outputs are stable under small input changes.*

**[V.AI-7.2] ADVERSARIAL INPUT PROTECTION**

The AI system shall verify adversarial input protection by test.

The test shall confirm input validation detects malformed or anomalous inputs, adversarial perturbation detection mechanisms are implemented, model outputs are stable under small input perturbations within operational bounds, and the

adversarial attack pattern set declared by the project in its verification evidence (Appendix C.2), for example FGSM, PGD, and C&W, is tested and model response is documented; for systems incorporating generative models or retrieval augmentation, the declared set shall include prompt injection attack patterns, both direct and indirect through retrieved or ingested content, and the test shall additionally confirm that those patterns are tested and the system response is documented. The declared set shall be documented with its derivation basis (the threat model, the system classification, and the declared Operational Design Domain); its adequacy is adjudicated during certification assessment in the same manner as threshold values (Section 1.5), and the declared set is recorded as an element of the certified configuration (Section 1.6).

**Success Criteria:** The verification shall be considered successful when the test shows that input validation detects anomalous inputs, perturbation detection is implemented, outputs are stable under small perturbations, and the declared attack pattern set, including prompt injection for generative and retrieval-augmented systems, is tested.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SI-10 (Information input validation); NIST AI 100-2e2023 §2 (Evasion attacks)
- AI-Specific: OWASP ML Top 10 (2023) ML01 (Input manipulation attack); OWASP Top 10 for LLM Applications LLM01 (Prompt injection); MITRE ATLAS AML.T0043 (Craft adversarial data); MITRE ATLAS AML.T0051 (LLM prompt injection)
- Robustness: ISO/IEC TR 24029-1:2021 (AI robustness assessment overview); IEEE 7009-2024 (Fail-safe design of autonomous systems)
- Aviation: RTCA DO-326A §3.7 (Security architecture); SAE ARP 6983 (AI/ML in airborne systems, draft)
- Automotive: ISO/SAE 21434:2021 §9.5 (Cybersecurity concept)
- Cross-Domain: NIST AI RMF MEASURE-2.7 (AI system security); EU AI Act Article 15(5) (Resilience against adversarial inputs)

### 2.7.3 Model Inversion and Extraction Protection

**[R.AI-7.3] MODEL INVERSION AND EXTRACTION PROTECTION**

The AI system shall implement protections against model inversion and extraction attacks.

***Rationale:*** *Model inversion attacks reconstruct sensitive training data by analyzing model outputs, potentially exposing sensitive information. Model extraction attacks systematically query the model to replicate its behavior, enabling attackers to study vulnerabilities offline. Protection requires access controls, query rate limiting, and output perturbation techniques that prevent systematic probing while maintaining operational utility. Where differential privacy or analogous techniques are applied for data subject privacy purposes, see also AI-10.4 (Privacy-Enhancing Techniques); the same mechanism may satisfy both requirements under distinct threat models.*

**[V.AI-7.3] MODEL INVERSION AND EXTRACTION PROTECTION**

The AI system shall verify model inversion and extraction protection by inspection and analysis.

The inspection shall confirm model access is restricted to authorized queries only, and training data sensitivity classification is documented.

The analysis shall confirm query rate limiting prevents systematic model probing, and output perturbation or differential privacy techniques are applied where applicable based on training data sensitivity.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that model access is restricted, training data sensitivity is classified, query rate limiting is implemented, and appropriate privacy techniques are applied.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SC-8 (Transmission confidentiality) and SC-28 (Protection of information at rest)
- AI-Specific: NIST AI 100-2e2023 §2.4 (Privacy Attacks: membership inference, model inversion, extraction); MITRE ATLAS AML.T0024 (Exfiltration via ML inference API)
- Privacy-Preserving ML: NIST SP 800-226 (Guidelines for evaluating differential privacy guarantees)
- Data Protection: ISO/IEC 27701:2019 (Privacy information management); GDPR Article 32 (Security of processing)
- Cross-Domain: NIST AI RMF MAP-4.1 (Approaches for mapping AI technology and legal risks); EU AI Act Article 15(4) (Resilience against confidentiality attacks)

### 2.7.4 Model Integrity

**[R.AI-7.4] MODEL INTEGRITY**

The AI system shall maintain model integrity throughout the operational lifecycle.

***Rationale:*** *Model integrity controls address two threat cases. In the first, the artifact was sound when validated and is modified, corrupted, or replaced afterward: cryptographic signing provides tamper-evident verification, and integrity checks at load time and at defined intervals during operation detect the change. The controls of this requirement are necessary for that case, in which unauthorized modification could otherwise introduce backdoors, degrade performance, or alter behavior in ways that circumvent all prior verification. In the second, the artifact is hostile at its origin: a backdoored or code-bearing artifact published by an adversary carries a valid signature and complete provenance, so the same controls verify the authenticity of a compromised artifact and pass. Signing establishes origin, not soundness. The controls of this requirement are therefore not sufficient for the adversarial-origin case, which is addressed by AI-7.6 (provenance and risk assessment of pre-trained components) and AI-7.7 (Artifact Load Safety). Model integrity controls apply in concert with the maintenance criteria established under AI-4.3 and the deployment format validation established in AI-4.7; legitimate model updates shall pass through all three control points before operational deployment.*

**[V.AI-7.4] MODEL INTEGRITY**

The AI system shall verify model integrity by inspection.

The inspection shall confirm model artifacts are cryptographically signed, integrity verification occurs at model load and at defined intervals during operation, unauthorized model modifications trigger alerts and fallback procedures, and model version history is maintained with tamper-evident logging.

**Success Criteria:** The verification shall be considered successful when the inspection shows that model artifacts are signed, integrity verification occurs at load and defined intervals, unauthorized modifications trigger alerts, and version history is maintained with tamper-evident logging.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SI-7 (Integrity) and CM-5 (Access restrictions for change)

- Secure Development: NIST SP 800-218 (SSDF, Secure Software Development Framework); SLSA Framework v1.0 (Supply-chain Levels for Software Artifacts)
- Model Signing: Sigstore / Linux Foundation Model Transparency project (cryptographic signing of model weights)
- Aviation: RTCA DO-326A §3.9 (Security assurance); RTCA DO-178C §11.5 (Configuration management)
- Automotive: ISO/SAE 21434:2021 §11 (Cybersecurity validation)
- Cross-Domain: NIST AI RMF MANAGE-2.1 (Risk response); EU AI Act Article 15(1) (Accuracy, robustness, and cybersecurity throughout lifecycle)

### 2.7.5 Adversarial Event Logging

**[R.AI-7.5] ADVERSARIAL EVENT LOGGING**

The AI system shall log all adversarial detection events and security incidents for post-mission analysis.

***Rationale:*** *Comprehensive logging of adversarial events supports post-mission forensic analysis, enables detection of attack patterns, and informs security improvements for future missions. Logs must be immutable to prevent attacker manipulation and must be retained for sufficient duration to support post-mission analysis.*

**[V.AI-7.5] ADVERSARIAL EVENT LOGGING**

The AI system shall verify adversarial event logging by inspection.

The inspection shall confirm security events are logged with timestamps, event type, and severity, logs capture detection confidence and system response, logs are immutable and retrievable for forensic analysis, and log retention meets mission duration plus post-mission analysis requirements.

**Success Criteria:** The verification shall be considered successful when the inspection shows that security events are logged with timestamps, type, and severity, logs capture detection confidence and response, logs are immutable and retrievable, and retention meets requirements.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 AU-2 (Event logging), AU-3 (Content of audit records), AU-12 (Audit record generation)

- Security Monitoring: ISO/IEC 27001:2022 A.8.15 (Logging) and A.8.16 (Monitoring activities)
- Incident Response: NIST SP 800-61 Rev. 2 (Computer security incident handling guide)
- Aviation: RTCA DO-326A §3.10 (Security monitoring)
- Automotive: ISO/SAE 21434:2021 §8.3 (Cybersecurity monitoring)
- Cross-Domain: NIST AI RMF MEASURE-4.3 (Ongoing monitoring); EU AI Act Article 12 (Record-keeping of events relevant to risk identification)

### 2.7.6 AI Supply Chain Security

**[R.AI-7.6] AI SUPPLY CHAIN SECURITY**

The AI system shall assess and document supply chain risks for all AI/ML frameworks, libraries, and pre-trained model components.

***Rationale:*** *AI systems depend on complex software supply chains including ML frameworks (e.g., TensorFlow, PyTorch), libraries (e.g., NumPy, ONNX), and potentially pre-trained models or transfer learning components. These dependencies may contain vulnerabilities, backdoors, or undocumented behaviors introduced through foreign contribution, compromised maintainers, or malicious updates. Pre-trained models may encode biases, backdoors, or behaviors from unknown training data. Supply chain assessment ensures that AI system integrity extends beyond the project's own development to encompass all external dependencies.*

**[V.AI-7.6] AI SUPPLY CHAIN SECURITY**

The AI system shall verify AI supply chain security by inspection.

The inspection shall confirm AI/ML frameworks and libraries are identified and documented with version numbers and sources, supply chain risk assessment addresses foreign contribution, maintenance status, and known vulnerabilities, pre-trained models (if used) have documented provenance and integrity verification, and mitigation strategies are documented for identified supply chain risks.

**Success Criteria:** The verification shall be considered successful when the inspection shows that AI/ML frameworks and libraries are documented, supply chain risk assessment addresses contribution sources and maintenance, pre-trained model provenance is documented, and mitigation strategies are documented for identified risks.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SR-3 (Supply chain controls) and SR-5 (Acquisition strategies)
- Supply Chain: NIST SP 800-161 Rev. 1 (C-SCRM for systems and organizations); NIST SP 800-218 (SSDF)
- Model Bill-of-Materials: NTIA Minimum Elements for SBOM (2021); CISA SBOM guidance; extensions for ML (ML-BOM)
- Aviation: RTCA DO-326A §3.8 (Supplier risk); RTCA DO-355 (Information security guidance for airworthiness)
- Automotive: ISO/SAE 21434:2021 §7 (Distributed cybersecurity activities: supplier relationships)
- Cross-Domain: NIST AI RMF GOVERN-6.1 (Third-party risk management); EU AI Act Article 25 (Obligations along the AI value chain)

### 2.7.7 Artifact Load Safety

**[R.AI-7.7] ARTIFACT LOAD SAFETY**

The AI system shall load model artifacts, and every other artifact admitted into its execution environment through a loader, including tokenizers, processors, datasets, and retrieval corpus content (AI-1.10), using serialization formats and loader configurations that do not execute code during deserialization or load. Any deviation shall be justified and risk assessed under AI-7.6. Loader configurations that execute remote or repository-supplied code at load shall be disabled, or the code so executed shall be brought under the same review and configuration control as project code.

***Rationale:*** *Integrity verification (AI-7.4) establishes that an artifact is the one its publisher produced, and provenance documentation (AI-7.6) records where it came from; neither prevents the artifact from executing code when it is loaded. Executable serialization formats run arbitrary code at deserialization, and loader configurations that execute repository-supplied code at load run whatever the repository contains. A hostile artifact that passes signature verification and carries complete provenance documentation therefore controls the host from the moment it loads. The exposure is the loader, not the artifact type: model weights, tokenizers, processors, datasets, and retrieval corpus content pass through loaders with the same behavior, which is why this requirement binds every artifact load path. Freedom from code execution is a property of the format together with its loader configuration, and several formats execute no code only under*

*configuration (custom operators, embedded graph code), so the requirement binds the loader configuration and not the format name alone. This requirement covers the adversarial-origin case named in the AI-7.4 rationale.*

**[V.AI-7.7] ARTIFACT LOAD SAFETY**

The AI system shall verify artifact load safety by inspection and test.

The inspection shall confirm that every artifact load path is documented with its serialization format and loader configuration; that the format and configuration on each path do not execute code during deserialization or load, or that each deviation is justified and risk assessed under AI-7.6; and that loader options which execute remote or repository-supplied code are disabled, or the code so executed is under project review and configuration control.

The test shall confirm that a crafted artifact carrying an executable payload, presented on each artifact load path, is rejected without execution.

**Success Criteria:** The verification shall be considered successful when the inspection and test show that every artifact load path is documented with a non-executing format and loader configuration or a justified and risk-assessed deviation, that remote code execution at load is disabled or controlled, and that crafted executable artifacts are rejected without execution on every load path.

**Related Domain Standards**

- General: NIST SP 800-53 Rev. 5 SI-3 (Malicious code protection), SI-7 (Software, firmware, and information integrity), and CM-7 (Least functionality)
- Secure Development: NIST SP 800-218 (SSDF, Secure Software Development Framework)
- AI-Specific: MITRE ATLAS AML.T0011.000 (User Execution: Unsafe ML Artifacts); MITRE ATLAS AML.T0010.003 (ML Supply Chain Compromise: Model); OWASP ML Top 10 (2023) ML06 (AI supply chain attacks); OWASP Top 10 for LLM Applications LLM03 (Supply chain)
- Cross-Domain: NIST AI RMF GOVERN-6.1 (Third-party risk management); EU AI Act Article 15(1) (Accuracy, robustness, and cybersecurity throughout lifecycle)

## 2.8 AI-8: EXPLAINABILITY

**[R.AI-8] EXPLAINABILITY**

The AI system will provide explanations of AI decisions and recommendations sufficient for operators to verify AI reasoning and establish appropriate trust.

Traditional software has explicit decision logic that can be inspected through code review: an engineer can trace exactly why a particular output was produced. AI systems, particularly neural networks, make decisions through learned patterns in weights that are not directly interpretable. ISO 22989 Section 3.5.7 defines explainability as the property of an AI system to express important factors influencing results in a way humans can understand. Without explainability, operators cannot verify AI reasoning, leading to either blind over-trust or unwarranted under-trust. This requirement addresses the "black-box decision making" gap identified in Section 1.2.4.2. Where AI-10.3 (Data Subject Rights) requires explanation of automated decisions, AI-8 explainability outputs serve as supporting evidence.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 3.5.7; ISO/IEC TS 6254:2024 (AI system explainability)
- Aviation: FAA AI Safety Roadmap (interpretability requirements)
- Medical: FDA Transparency Principles for Medical Devices with AI/ML (2021); FDA CDS Guidance (January 2026) §IV(4)
- EU: EU AI Act Article 13 (Transparency and provision of information)
- Cross-Domain: NIST AI RMF MANAGE-4.3; NIST SP 1270 (explainability and bias intersection)

### 2.8.1 Explainability Requirements Definition

**[R.AI-8.1] EXPLAINABILITY REQUIREMENTS DEFINITION**

The AI system shall define explainability requirements appropriate to each AI function's criticality and user needs.

***Rationale:*** *Explainability requirements should scale with decision criticality; Safety-Critical AI functions require more rigorous explanation capabilities than Operational Support functions. Requirements must also consider the target audience; operators need operationally relevant explanations, while engineers may need deeper technical detail. Defining requirements upfront ensures explainability is designed in, not retrofitted.*

**[V.AI-8.1] EXPLAINABILITY REQUIREMENTS DEFINITION**

The AI system shall verify explainability requirements definition by inspection.

The inspection shall confirm explainability requirements are defined for each AI function, requirements are scaled to function criticality and decision impact, target audience and their explanation needs are documented (e.g., operators, engineers, analysts), and explainability approach is justified for each AI function (e.g., feature importance, attention visualization, rule extraction, inherently interpretable models).

**Success Criteria:** The verification shall be considered successful when the inspection shows that explainability requirements are defined for each AI function, requirements scale with criticality, target audiences are documented, and explainability approaches are justified.

### 2.8.2 Decision Explanation Generation

**[R.AI-8.2] DECISION EXPLANATION GENERATION**

The AI system shall provide decision explanations in a format understandable to operators.

***Rationale:*** *Explanations must be actionable and comprehensible to the intended audience within operational time constraints. Technical explanations (e.g., feature weights, activation patterns) may be appropriate for post-mission analysis but are inadequate for real-time operational decisions. Explanation format should be validated through human factors assessment to ensure operators actually understand and can act on the information provided.*

**[V.AI-8.2] DECISION EXPLANATION GENERATION**

The AI system shall verify decision explanation generation by demonstration.

The demonstration shall confirm explanations are generated for AI decisions and recommendations, explanation format is appropriate to operational context and timing constraints, explanations identify key factors influencing the decision, and operator comprehension is validated through usability assessment with representative operators performing representative tasks under realistic conditions.

**Success Criteria:** The verification shall be considered successful when the demonstration shows that explanations are generated for AI decisions, format is appropriate to context, key influencing factors are identified, and operator comprehension is validated through usability assessment.

### 2.8.3 Confidence Indication

**[R.AI-8.3] CONFIDENCE INDICATION**

The AI system shall indicate confidence levels for AI outputs.

***Rationale:*** *AI outputs vary in reliability: some predictions are made with high confidence, others are uncertain. Operators need to know when to trust AI outputs and when to apply additional scrutiny or seek alternative information. Confidence must be calibrated; a stated 90% confidence should be correct approximately 90% of the time, to enable appropriate operator trust calibration. ISO 22989 Section 3.5.8 defines predictability as enabling reliable assumptions about outputs; calibrated confidence supports this property.*

**[V.AI-8.3] CONFIDENCE INDICATION**

The AI system shall verify confidence indication by test.

The test shall confirm confidence metrics are computed for AI outputs, confidence is presented to operators with outputs in a format that supports decision-making, confidence calibration is validated against actual accuracy (e.g., Expected Calibration Error ≤ 0.05), and low-confidence outputs are clearly distinguished from high-confidence outputs through visual or auditory differentiation.

**Success Criteria:** The verification shall be considered successful when the test shows that confidence metrics are computed, confidence is presented with outputs, calibration is validated against actual accuracy, and low-confidence outputs are clearly distinguished.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 3.5.8 (Predictability); ISO/IEC TS 6254:2024 §6 (Confidence communication)
- Aviation: FAA AI Safety Roadmap (Interpretability and uncertainty quantification)
- Automotive: SAE J3016:2021 §5.3 (Automation reliability feedback to user)
- Medical: FDA AI/ML Transparency Principles (2021) Principle 3 (Labeling for transparency); FDA CDS Guidance (January 2026) §IV(4) (Criterion 4: basis and evidence supporting recommendations)
- EU: EU AI Act Article 13(3)(b)(iv) (Information on level of accuracy and reliability)
- Cross-Domain: NIST AI RMF MEASURE-2.5

### 2.8.4 Reasoning Inspection Capability

**[R.AI-8.4] REASONING INSPECTION CAPABILITY**

The AI system shall enable operators to query or inspect AI reasoning for safety-critical decisions.

***Rationale:*** *For safety-critical decisions, operators may need to understand AI reasoning in greater depth than routine explanations provide. On-demand inspection capability allows operators to drill down into AI reasoning when uncertainty exists or when AI recommendations conflict with operator expectations. This capability supports the human decision authority required by AI-9.1 (Human Decision Authority) and aligns with ISO 22989 Section 3.5.6 controllability requirements.*

**[V.AI-8.4] REASONING INSPECTION CAPABILITY**

The AI system shall verify reasoning inspection capability by test.

The test shall confirm query or inspection capability is available for safety-critical AI functions, operators can access additional detail on demand beyond routine explanations, inspection does not unacceptably delay operational timelines, and inspection capability is documented in operator procedures.

**Success Criteria:** The verification shall be considered successful when the test shows that inspection capability is available for safety-critical functions, operators can access additional detail on demand, inspection meets timing requirements, and capability is documented in procedures.

### 2.8.5 Public Disclosure Support

**[R.AI-8.5] PUBLIC DISCLOSURE SUPPORT**

The AI system shall provide the artifacts and information necessary to support public disclosure obligations applicable to the deployment context, in a form suitable for stakeholder consumption outside the development organization.

***Rationale:*** *AI systems may be subject to public disclosure requirements including external model cards or system factsheets, consumer-facing information under EU AI Act Article 13, federal AI use case inventory reporting under Executive Order 14110 and OMB M-24-10, and public registration for high-risk systems under EU AI Act Article 71. The specific disclosure obligations depend on deployment context and generally bind the deploying organization rather than the developer; however, the AI system design must support those obligations by producing standardized disclosure artifacts consumable by regulators, users, data subjects, and the public. This requirement complements the internal explainability requirements (AI-8.1 through AI-8.4) by establishing the external-facing disclosure surface. Disclosure artifacts should be derived from existing framework evidence (ODD declaration per AI-1.0, bias evaluation per AI-2, performance characterization per AI-4,*

*explainability outputs per AI-8.1-8.4) rather than generated independently, to ensure external disclosures remain consistent with internal verification evidence. This sub-requirement is optional by classification and may be documented as Not Applicable with rationale where no public disclosure obligation applies.*

**[V.AI-8.5] PUBLIC DISCLOSURE SUPPORT**

The AI system shall verify public disclosure support by inspection.

The inspection shall confirm that applicable public disclosure obligations are identified in the ODD declaration under Regulatory and Operational Authorization (AI-1.0); that disclosure artifact formats are defined (e.g., external model card, AI factsheet, EU AI Act Article 13 user information, federal AI inventory entry); that each required artifact is populated from existing framework evidence rather than generated in parallel; and that artifacts are reviewed prior to release for consistency with ODD declaration, documented limitations, and verification evidence.

The inspection shall further confirm that artifact update procedures are defined so that disclosure artifacts remain consistent with the current system state following updates, drift remediation (AI-4.3), reclassification (AI-9.6), or ODD revision; and that artifacts are retained for the retention period required by the applicable regulatory regime.

**Success Criteria:** The verification shall be considered successful when the inspection shows that applicable disclosure obligations are identified, disclosure artifacts are defined and populated from framework evidence, artifacts are reviewed for consistency before release, update procedures are defined, and retention requirements are satisfied. Where no public disclosure obligation applies, AI-8.5 may be documented as Not Applicable with rationale.

**Related Domain Standards**

- General: ISO/IEC 23894:2023 Annex A.12 (Transparency and explainability to stakeholders); ISO/IEC 5259-4 §7 (Data quality process for ML); ISO/IEC 42001:2023 §8.3 (Communication)
- Aviation: FAA AI Safety Roadmap (stakeholder transparency)
- Automotive: NHTSA AV 4.0 element 2 (Operational Design Domain description); NHTSA AV Transparency and Engagement Principles
- Medical: FDA Predetermined Change Control Plan Guidance (2024) §V (Policy for Predetermined Change Control Plans, including labeling requirements at

lines 844-887 of the guidance); FDA Transparency for Machine Learning-Enabled Device Guidance (2021)

- EU: EU AI Act Article 13 (Transparency and provision of information to deployers); EU AI Act Article 50 (Transparency obligations for general-purpose AI); EU AI Act Article 71 (Public EU database registration for high-risk AI)
- US Federal: Executive Order 14110 Section 10.1(b) (Federal AI inventory); OMB M-24-10 (Federal AI use case inventory requirements); NIST AI 100-1 (AI Risk Management Framework)
- Cross-Domain: NIST AI RMF GOVERN-1.4 (Risk documentation); NIST AI RMF MANAGE-4.1 (Post-deployment monitoring disclosure)

## 2.9 AI-9: HUMAN-AI TEAMING

### [R.AI-9] HUMAN-AI TEAMING

The AI system will support effective human-AI collaboration, enabling operators to maintain situational awareness, appropriate trust calibration, and final decision authority over safety-critical actions.

AI systems in safety-critical operations operate as part of a human-machine team, not as fully autonomous agents. ISO 22989 Section 5.13 defines automation levels ranging from full human control to full autonomy; for safety-critical systems, human oversight remains essential. Operators must understand what the AI is doing, trust it appropriately (neither too much nor too little), and retain authority to override AI actions when necessary. This requirement addresses the “operators cannot verify reasoning” gap and ensures AI augments rather than replaces human judgment in safety-critical decisions.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 3.5.6 (Controllability); §5.13 (Automation levels)
- Aviation: 14 CFR 25.1302 (Installed systems and equipment for use by the flightcrew); FAA AC 25.1322 (Flight Crew Alerting); FAA AI Safety Roadmap (human oversight)
- Automotive: SAE J3016:2021 §5 (Automation levels); NHTSA Automated Vehicles Comprehensive Plan (2021)
- Medical: FDA CDS Guidance (January 2026) §IV(4) (Criterion 4: independent review); HHS/ONC Health IT Certification

- EU: EU AI Act Article 14 (Human oversight); EU AI Act Article 26 (Deployer obligations)
- Cross-Domain: NIST AI RMF GOVERN-1.5; NIST AI RMF MANAGE-2.1

### 2.9.1 Human Decision Authority

**[R.AI-9.1] HUMAN DECISION AUTHORITY**

The AI system shall maintain human decision authority over safety-critical actions.

***Rationale:*** *For safety-critical actions, those that could result in harm, equipment damage, or mission loss, humans must retain final decision authority. AI may recommend, advise, or prepare actions, but execution should require human authorization except where explicitly designed and approved for autonomous operation (e.g., time-critical emergency responses). ISO 22989 Section 3.5.6 defines controllability as the property allowing human intervention; this requirement ensures controllability is maintained for safety-critical functions.*

**[V.AI-9.1] HUMAN DECISION AUTHORITY**

The AI system shall verify human decision authority by test.

The test shall confirm safety-critical actions require human authorization before execution, AI cannot execute safety-critical actions autonomously unless explicitly designed, approved, and documented for autonomous operation, human override capability is always available and functional, and authority boundaries are documented and enforced through technical controls.

**Success Criteria:** The verification shall be considered successful when the test shows that safety-critical actions require human authorization, autonomous execution is limited to explicitly approved cases, human override is always available, and authority boundaries are documented and enforced.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 3.5.6 (Controllability)
- Aviation: 14 CFR 25.1302 (Flight crew installed systems); FAA AC 25.1322 §4 (Alert prioritization and integration)
- Automotive: SAE J3016:2021 §5.2 (Dynamic Driving Task fallback); NHTSA AV 4.0 Voluntary Safety Self-Assessment element 3 (Human Machine Interface)
- Medical: FDA Clinical Decision Support Software Guidance (January 2026) §IV(4) (Criterion 4: basis for practitioner independent review)

- EU: EU AI Act Article 14(4) (Human oversight measures for high-risk AI); EU AI Act Article 14(5) (oversight of certain biometric categorization)
- Cross-Domain: NIST AI RMF GOVERN-1.5; NIST AI RMF MANAGE-2.1

### 2.9.2 Operational Situational Awareness

**[R.AI-9.2] OPERATIONAL SITUATIONAL AWARENESS**

The AI system shall support operator situational awareness of AI state and behavior.

***Rationale:*** *Operators cannot effectively supervise AI systems they do not understand. Situational awareness requires visibility into what the AI is doing, what it is recommending, and why. Changes in AI state (e.g., mode transitions, confidence drops, detected anomalies) must be communicated to operators to prevent surprise and enable proactive response.*

**[V.AI-9.2] OPERATIONAL SITUATIONAL AWARENESS**

The AI system shall verify operator situational awareness support by test.

The test shall confirm AI operational state is visible to operators through appropriate displays, AI actions and recommendations are communicated clearly within operational timing constraints, changes in AI state or confidence are annunciated (e.g., state change notification ≤500ms, confidence drop >15% triggers alert), and operators can determine what the AI is doing and why through available displays and explanations.

**Success Criteria:** The verification shall be considered successful when the test shows that AI state is visible to operators, actions and recommendations are clearly communicated, state changes are annunciated within timing requirements, and operators can determine AI status and reasoning.

### 2.9.3 Trust Calibration

**[R.AI-9.3] TRUST CALIBRATION**

The AI system shall support appropriate trust calibration between operators and AI.

***Rationale:*** *Miscalibrated trust either over-trust or under-trust degrades human-AI team performance. Over-trust leads to automation complacency where operators fail to catch AI errors; under-trust leads to operators ignoring valid AI recommendations. Trust calibration requires that operators understand AI reliability and limitations, that training addresses appropriate trust levels, and that system design promotes calibrated trust through transparency and appropriate confidence communication.*

**[V.AI-9.3] TRUST CALIBRATION**

The AI system shall verify trust calibration support by analysis and demonstration.

The analysis shall confirm AI reliability and limitations are documented and communicated to operators, and training materials address appropriate AI trust levels including scenarios where AI should and should not be trusted.

The demonstration shall confirm system design discourages both over-trust (e.g., through uncertainty indication, limitation warnings) and under-trust (e.g., through reliability demonstration, confidence calibration), and trust calibration is assessed through human factors evaluation with representative operators under realistic operational conditions.

**Success Criteria:** The verification shall be considered successful when the analysis and demonstration show that AI reliability and limitations are communicated, training addresses trust calibration, system design discourages miscalibrated trust, and human factors evaluation confirms appropriate trust calibration.

### 2.9.4 Graceful Degradation

**[R.AI-9.4] GRACEFUL DEGRADATION**

The AI system shall degrade gracefully when AI capabilities are reduced or unavailable.

***Rationale:*** *AI systems may experience capability reduction due to sensor failures, computational resource constraints, detected anomalies, or explicit mode transitions. When AI capabilities are degraded, operators must be notified and manual or backup procedures must be available to continue operations. Abrupt AI failure without graceful degradation could leave operators without situational awareness or control capability. Degraded mode definitions shall be consistent with the System Inputs degradation behaviors declared under AI-1.0.*

**[V.AI-9.4] GRACEFUL DEGRADATION**

The AI system shall verify graceful degradation by test.

The test shall confirm degraded AI modes are defined and documented for each AI function, operators are notified of AI capability reduction through appropriate annunciation, manual or backup procedures are available and documented for operation without full AI capability, and transition to degraded modes is tested and acceptable (e.g., ≤2 operator actions, ≤5 seconds to transition to manual mode).

**Success Criteria:** The verification shall be considered successful when the test shows that degraded modes are defined, operators are notified of capability

reduction, manual/backup procedures are available, and mode transitions are tested and acceptable.

### 2.9.5 Human-AI Interaction Logging

**[R.AI-9.5] HUMAN-AI INTERACTION LOGGING**

The AI system shall log human-AI interactions for post-mission analysis.

***Rationale:*** *Comprehensive logging of human-AI interactions supports post-mission analysis to identify human factors issues, assess trust calibration, evaluate override patterns, and improve future AI systems. Logs should capture what the AI recommended, what the human decided, and the rationale for any overrides. This data is essential for continuous improvement of human-AI teaming effectiveness.*

**[V.AI-9.5] HUMAN-AI INTERACTION LOGGING**

The AI system shall verify human-AI interaction logging by inspection.

The inspection shall confirm human-AI interactions are logged with timestamps, logs capture AI recommendations and human decisions, logs capture overrides with rationale, and logs support post-mission human factors analysis including trust calibration assessment and override pattern analysis.

**Success Criteria:** The verification shall be considered successful when the inspection shows that interactions are logged with timestamps, logs capture recommendations, decisions, and override rationale, and logs support post-mission human factors analysis.

### 2.9.6 Operational Role Verification

**[R.AI-9.6] OPERATIONAL ROLE VERIFICATION**

The AI system shall implement periodic verification that the system's operational role remains consistent with its classification.

***Rationale:*** *AI system classification (Section 1.2.2) determines which requirements apply and at what stringency. Classification is assigned based on the intended operational role of the AI system: whether its outputs are advisory, operational, or safety critical. However, operational usage can drift over time without any change to the system itself. A system classified as Operational Support may produce outputs that operators gradually elevate from advisory to authoritative, relying on AI recommendations as though they were directives. This usage drift is particularly insidious because no technical change triggers a review. The model performs identically; only the human reliance on it has changed. When this occurs, the system is effectively operating at a higher classification tier without the*

*corresponding requirements, verifications, or safeguards. Periodic verification of operational role ensures that classification remains aligned with actual usage and that any drift toward higher reliance triggers formal reclassification and the application of additional requirements. This is the human factors complement to the technical drift monitoring required by AI-4.*

**[V.AI-9.6] OPERATIONAL ROLE VERIFICATION**

The AI system shall verify operational role verification by inspection and analysis.

The inspection shall confirm a periodic review schedule is defined and justified based on system classification and operational tempo, the review process includes assessment of how operators actually use AI outputs (not just how the system is designed to be used), review criteria explicitly address whether operators treat AI outputs as advisory, operational, or authoritative, review criteria include adherence of actual operational use to the declared ODD (AI-1.0), and review findings are documented and retained as verification evidence.

The analysis shall confirm the review process includes mechanisms to detect usage drift, such as operator workflow analysis, override rate trends, or decision dependency assessment; that identified usage drift triggers formal reclassification evaluation per Section 1.2.2; and that when reclassification is warranted, the additional requirements corresponding to the new classification are applied and verified before continued operation at the elevated role.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that a periodic review schedule is defined and justified, the review assesses actual operator usage against classification assumptions, usage drift detection mechanisms are defined, identified drift triggers reclassification evaluation, and reclassification results in application of corresponding requirements.

### 2.9.7 Operator Qualification

**[R.AI-9.7] OPERATOR QUALIFICATION**

The AI system shall require documented operator qualifications appropriate to system classification and operational role, including system-specific training, demonstrated proficiency in normal and abnormal operations, and currency requirements that remain valid during continued operation.

***Rationale:*** *AI system safety and performance depend on operator understanding of system capabilities, limitations, and failure modes. Unqualified operators may over-trust AI outputs, fail to recognize degraded performance, or misinterpret advisory outputs as directives. Qualification rigor should scale with classification:*

*Safety-Critical AI requires formal operator certification, Mission-Critical AI requires role-based qualification, and Operational Support AI requires documented familiarization. Currency requirements prevent qualification decay during extended operation and ensure operators remain proficient following system updates that materially change AI behavior.*

**[V.AI-9.7] OPERATOR QUALIFICATION**

The AI system shall verify operator qualification by inspection and test.

The inspection shall confirm that operator qualification requirements are documented for each operational role; that qualification criteria address AI-specific topics including system capabilities, documented limitations (Section 1.5 thresholds, AI-1.0 ODD bounds), known failure modes (AI-6 OOD, AI-7 adversarial), and appropriate trust calibration (AI-9.3); that currency requirements are defined with intervals justified by system classification and operational tempo; and that qualification records are retained as verification evidence.

The test shall confirm that operators assigned to the AI system hold current qualification for their role, that operators without current qualification are prevented from assuming roles that require it through administrative or technical controls, and that qualification re-validation is triggered when material system updates affect operator-relevant behavior.

**Success Criteria:** The verification shall be considered successful when the inspection and test show that operator qualification requirements are documented by role and classification, AI-specific qualification content is addressed, currency requirements are defined and enforced, and qualification records are retained as verification evidence.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 3.5.5 (Human-AI collaboration); ISO/IEC 25059:2023 §5 (Quality characteristics for AI systems)
- Aviation: 14 CFR 121 Subpart N (Training Program); 14 CFR 121 Subpart Y (Advanced Qualification Program); FAA AI Safety Roadmap (operator training)
- Automotive: SAE J3016:2021 §8 (User considerations); NHTSA AV 4.0 element 12 (Consumer Education and Training)
- Medical: FDA 21 CFR 820.25 (Personnel - Training); FDA Clinical Decision Support Software Guidance (January 2026); note: 'Practitioner learning' is not addressed as a distinct section in the current 2026 version; concept is implicit in §IV(4) Criterion 4 discussion of HCP independent review

- EU: EU AI Act Article 4 (AI literacy); EU AI Act Article 14(4)(a) (Oversight competence); EU AI Act Article 26(2) (Deployer training obligations)
- Cross-Domain: NIST AI RMF GOVERN-2.2 (Accountability for staff); NIST SP 800-53 Rev. 5 AT-3 (Role-Based Training); NIST SP 800-53 Rev. 5 PS-3 (Personnel Screening)

### 2.9.8 Workload Management

**[R.AI-9.8] WORKLOAD MANAGEMENT**

The AI system shall define operational workload parameters and implement design measures that maintain operator cognitive capacity within documented limits under nominal, degraded, and high-tempo conditions.

***Rationale:*** *AI deployment can increase or decrease operator cognitive workload depending on interface design, alert logic, and the degree to which AI outputs require operator verification. Excess workload degrades decision quality, delays responses to abnormal conditions, and increases error probability. Insufficient workload induces complacency and reduces attentional resources available for monitoring AI outputs, which is a documented precursor to automation-induced failures. Workload must be validated as part of the operational design rather than discovered in deployment; this requirement complements trust calibration (AI-9.3) by ensuring the operator retains the cognitive margin required to exercise judgment and complements operator qualification (AI-9.7) by establishing the environment in which qualified operators function.*

**[V.AI-9.8] WORKLOAD MANAGEMENT**

The AI system shall verify workload management by analysis and test.

The analysis shall confirm that operational workload parameters are defined for representative nominal, degraded, and high-tempo scenarios; that cognitive workload limits are documented with reference to an accepted workload assessment method (e.g., NASA-TLX, Bedford, SWAT) or domain guidance (e.g., FAA AC 25.1302, AAMI HE75); that automation-bias and complacency risks are explicitly addressed in the workload design; and that workload monitoring mechanisms (manual override rates, alert acknowledgement latency, task completion times) are specified.

The test shall confirm that under representative operational scenarios, measured workload remains within documented limits; that mitigations are invoked when workload indicators exceed defined thresholds (e.g., alert suppression, task offloading, fallback behaviors); and that operators retain the cognitive capacity necessary to exercise decision authority required by AI-9.1.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that workload parameters are defined for nominal, degraded, and high-tempo scenarios, measured workload remains within documented limits, automation-bias and complacency risks are addressed, and workload mitigations are effective.

**Related Domain Standards**

- General: ISO 9241-11 (Usability: Definitions and concepts); ISO 11064 (Ergonomic design of control centres); ISO/IEC 25059:2023 §5 (AI system quality characteristics)
- Aviation: 14 CFR 25.1302 (Flightcrew workload); FAA AC 25.1302-1 (Installed Systems and Equipment for Use by the Flightcrew); FAA AI Safety Roadmap (cognitive workload)
- Automotive: NHTSA Visual-Manual Driver Distraction Guidelines (2013); SAE J3016:2021 §5.3 (Driver readiness); SAE J2944 (Driving performance terms)
- Medical: FDA Human Factors and Usability Engineering Guidance (2016) §6.3.1 (Task Analysis); AAMI HE75:2009/(R)2018 (Human factors design)
- EU: EU AI Act Article 14(4)(d) (Measures to counter automation bias); Council Directive 89/391/EEC Article 6 (General obligations for safe working conditions)
- Cross-Domain: NIST AI RMF MEASURE-2.8 (Human factors); MIL-STD-1472H (DoD Human Engineering Design Criteria); NUREG-0711 Rev. 3 (Human Factors Engineering Program Review)

### 2.9.9 Training Program Requirements

**[R.AI-9.9] TRAINING PROGRAM REQUIREMENTS**

The AI system shall be supported by a training program that provides initial qualification, recurrent training, and scenario-based exercises covering AI system capabilities, documented limitations, failure modes, and appropriate trust calibration.

***Rationale:*** *Training programs for AI systems must address considerations that differ from traditional system training: non-deterministic behavior, performance variability across operational conditions, AI-specific failure modes such as OOD inputs and adversarial attacks, and the calibration between operator reliance and demonstrated AI performance. Initial qualification alone is insufficient because AI system behavior may evolve with updates, data drift, or operational context.*

*Recurrent training maintains proficiency during extended operation, and scenario-based exercises expose operators to rare failure modes before they occur operationally. Training program requirements complement operator qualification (AI-9.7) by defining the content and cadence that produce qualified operators, and complement workload management (AI-9.8) by ensuring operators possess the knowledge required to recognize and respond to high-workload conditions.*

**[V.AI-9.9] TRAINING PROGRAM REQUIREMENTS**

The AI system shall verify training program requirements by inspection.

The inspection shall confirm that an initial qualification curriculum is documented and covers AI system capabilities, documented limitations (Section 1.5 thresholds, AI-1.0 ODD bounds), known failure modes (AI-6 OOD, AI-7 adversarial, AI-4 drift), explainability outputs (AI-8), and trust calibration guidance (AI-9.3); that recurrent training is scheduled at documented intervals justified by classification and operational tempo; that scenario-based exercises cover both nominal operations and rare off-nominal conditions including AI-specific failure modes; and that training effectiveness is measured through assessment of operator performance rather than attendance alone.

The inspection shall further confirm that material system updates trigger targeted training for the affected operator population before the updated system is released to operational use; that training content is updated to reflect current system behavior, thresholds, and ODD; and that training records are retained as verification evidence supporting AI-9.7 qualification currency.

**Success Criteria:** The verification shall be considered successful when the inspection shows that initial qualification, recurrent training, and scenario-based exercises are documented and delivered; that AI-specific content is addressed; that training is updated in response to material system changes; and that training records support AI-9.7 qualification currency.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 5.5 (Operation and monitoring); ISO/IEC 27002:2022 §6.3 (Information security awareness, education and training)
- Aviation: 14 CFR 121 Subpart Y (Advanced Qualification Program); FAA AC 120-35D (Line Operational Simulations); FAA AI Safety Roadmap §4 (Operator proficiency)
- Automotive: NHTSA AV 4.0 element 12 (Consumer education); SAE J3016:2021 §8.3 (User training for higher levels of driving automation)

- Medical: FDA 21 CFR 820.25 (Personnel training); FDA Clinical Decision Support Software Guidance (January 2026); note: dedicated 'Practitioner learning and proficiency' section does not exist in the current 2026 version; concept is implicit in §IV(4)
- EU: EU AI Act Article 4 (AI literacy); EU AI Act Article 26(2) (Deployer obligations, competence of natural persons assigned oversight)
- Cross-Domain: NIST AI RMF GOVERN-2.2 (Accountability for staff); NIST SP 800-53 Rev. 5 AT-2 (Literacy Training), AT-3 (Role-Based Training), AT-4 (Training Records)

## 2.10 AI-10: PRIVACY AND DATA PROTECTION

### [R.AI-10] PRIVACY AND DATA PROTECTION

The AI system will implement personal data protection mechanisms that enable lawful processing, respect data subject rights, and minimize privacy risk throughout the AI lifecycle.

AI systems frequently process personal data: patient records in medical AI, biometric data in identity systems, behavioral data in recommendation systems, sensor-derived personal information in autonomous platforms. The handling of this data is governed by privacy regulations distinct from information classification regimes: GDPR in the EU, HIPAA in US healthcare, CCPA/CPRA in California, PIPEDA in Canada, and analogous frameworks globally. AI systems also introduce privacy risks not present in traditional software: training data may be reconstructable from model parameters (model inversion), trained models may leak membership information about training subjects, and the opacity of AI decisions complicates the exercise of data subject rights. This requirement addresses personal data handling throughout the AI lifecycle, complementing the information classification handling established in AI-1.5 through AI-1.8.

**Applicability:** AI-10 applies to AI systems that process personal data as defined under any applicable privacy framework. Where the system does not process personal data, AI-10 sub-requirements may be documented as Not Applicable with rationale. Determination of personal data processing shall be documented in the ODD declaration (AI-1.0) under Subject Population and Regulatory and Operational Authorization.

**Related Domain Standards**

- General: ISO/IEC 27701:2019 (Privacy Information Management); ISO/IEC 29100:2011 (Privacy framework)

- Medical: HIPAA Privacy Rule (45 CFR 164 Subpart E); HIPAA Security Rule (45 CFR 164 Subpart C)
- EU: GDPR (Regulation 2016/679); EU AI Act Article 10(5) (special categories of personal data)
- US State: CCPA/CPRA (Cal. Civ. Code §§ 1798.100 et seq.)
- Cross-Domain: NIST Privacy Framework v1.0; NIST SP 800-53 Rev. 5 PT family; NIST SP 800-122 (PII)

### 2.10.1 Personal Data Identification and Minimization

**[R.AI-10.1] PERSONAL DATA IDENTIFICATION AND MINIMIZATION**

The AI system shall identify personal data within all training, validation, test, and operational datasets, and shall apply data minimization principles to limit personal data collection, retention, and processing to what is necessary for the declared purpose.

***Rationale:*** *Personal data identification is a prerequisite for any privacy control. Without knowing what personal data is present, no other privacy requirement can be satisfied. Direct identifiers (name, email, government ID, biometric data, medical record number) are typically straightforward to detect; indirect identifiers (combinations of attributes that enable re-identification) require more sophisticated assessment. Data minimization, processing only what is necessary for the declared purpose, is foundational to GDPR Article 5(1)(c), CCPA reasonable necessity provisions, and HIPAA minimum necessary standard. AI systems often violate minimization implicitly by retaining all collected data for potential future training; explicit minimization requires documented justification for each data category retained.*

**[V.AI-10.1] PERSONAL DATA IDENTIFICATION AND MINIMIZATION**

The AI system shall verify personal data identification and minimization by inspection and analysis.

The inspection shall confirm that personal data identification methods are documented and applied to all datasets in scope; that direct and indirect identifiers are documented per dataset; that data minimization analysis documents the necessity of each personal data category retained; and that minimization decisions are reviewed at the cadence of the dataset refresh cycle.

The analysis shall confirm that automated identification tools are validated against known personal data samples; that re-identification risk assessment addresses

indirect identifier combinations; and that retention periods for personal data are bounded and justified relative to the declared processing purpose.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that personal data identification methods are documented and applied, identifiers are documented per dataset, minimization analysis is complete with documented necessity, automated tools are validated, re-identification risk is assessed for indirect identifiers, and retention periods are bounded and justified.

### 2.10.2 Lawful Basis and Consent Management

**[R.AI-10.2] LAWFUL BASIS AND CONSENT MANAGEMENT**

The AI system shall document the lawful basis for personal data processing under each applicable privacy framework, and shall implement consent management mechanisms where consent is the lawful basis.

***Rationale:*** *Personal data processing requires a lawful basis under all major privacy frameworks. GDPR Article 6 enumerates six lawful bases (consent, contract, legal obligation, vital interests, public task, legitimate interests); HIPAA distinguishes covered entity processing, business associate processing, and authorization-based processing; CCPA provides a notice-and-opt-out framework with specific bases for sensitive personal information. The lawful basis selected determines the obligations that follow: consent-based processing requires consent records, withdrawal mechanisms, and re-consent triggers; legitimate-interests-based processing requires a documented balancing test. AI systems trained on data collected under one lawful basis may be unable to be retrained on the same data if the basis changes; documenting basis throughout the lifecycle is necessary for ongoing lawful operation.*

**[V.AI-10.2] LAWFUL BASIS AND CONSENT MANAGEMENT**

The AI system shall verify lawful basis and consent management by inspection.

The inspection shall confirm that the lawful basis for personal data processing is documented under each applicable privacy framework; that the basis is documented per dataset and per processing purpose; that consent records are maintained where consent is the basis, including consent text, capture mechanism, timestamp, and withdrawal status; that consent withdrawal mechanisms are implemented and accessible to data subjects; and that processing ceases or is justified under a different basis when consent is withdrawn.

**Success Criteria:** The verification shall be considered successful when the inspection shows that lawful basis is documented per dataset and processing

purpose under each applicable framework, consent records are complete and accessible, withdrawal mechanisms are functional, and basis changes trigger documented review.

### 2.10.3 Data Subject Rights Implementation

**[R.AI-10.3] DATA SUBJECT RIGHTS IMPLEMENTATION**

The AI system shall implement procedures to support data subject rights as required by applicable privacy frameworks, including rights of access, rectification, erasure, restriction, portability, and objection.

***Rationale:*** *Data subject rights are a defining feature of modern privacy frameworks. GDPR Articles 15 through 22 establish access, rectification, erasure, restriction, portability, objection, and rights regarding automated decision-making. HIPAA establishes patient access rights and amendment rights under 45 CFR 164.524 and 164.526. CCPA establishes access, deletion, and opt-out rights. AI systems present unique implementation challenges: the right to erasure may require model retraining if personal data influenced learned parameters; the right to access may require explanation of automated decisions; the right to rectification of training data does not retroactively correct already-trained model behavior. These challenges require explicit procedural design rather than ad-hoc handling. Where data subject rights require explanation of automated decisions, AI-8 (Explainability) outputs serve as supporting evidence.*

**[V.AI-10.3] DATA SUBJECT RIGHTS IMPLEMENTATION**

The AI system shall verify data subject rights implementation by inspection and test.

The inspection shall confirm that procedures are documented for each data subject right applicable under the relevant privacy frameworks; that response timelines meet regulatory requirements (e.g., 30 days under GDPR Article 12(3), 30 days under HIPAA 164.524, 45 days under CCPA); that procedures address AI-specific implementation challenges including erasure from trained models and access to automated decision explanations; and that data subject request intake mechanisms are accessible and documented.

The test shall confirm that representative data subject rights requests can be processed end-to-end within required timelines; that erasure requests result in documented removal from datasets and assessment of model retraining necessity; and that access requests produce intelligible responses including, where applicable, explanation of automated decision logic per AI-8.

**Success Criteria:** The verification shall be considered successful when the inspection and test show that procedures exist for each applicable right, timelines meet regulatory requirements, AI-specific challenges are addressed procedurally, intake mechanisms are accessible, end-to-end processing meets timelines, erasure assessments include model retraining necessity, and access responses include automated decision explanation where applicable.

### 2.10.4 Privacy-Enhancing Techniques

**[R.AI-10.4] PRIVACY-ENHANCING TECHNIQUES**

The AI system shall apply privacy-enhancing techniques appropriate to the personal data sensitivity, processing purpose, and re-identification risk of the system.

***Rationale:*** *Privacy-enhancing techniques (PETs) reduce privacy risk through technical design rather than procedural controls alone. Anonymization removes the link between data and data subjects; pseudonymization replaces identifiers with reversible tokens; differential privacy adds calibrated noise to prevent membership inference; federated learning trains models without centralizing raw data; secure aggregation enables computation without exposing inputs; encryption protects data at rest, in transit, and increasingly in computation. The selection of PETs depends on the threat model, the utility requirements, and the regulatory regime. PETs are not interchangeable: anonymization that meets HIPAA Safe Harbor may not meet GDPR anonymization standards; differential privacy parameters that satisfy academic definitions may not satisfy regulatory expectations. Where differential privacy is applied as defense against model inversion attacks, see also AI-7.3 (Model Inversion and Extraction Protection); AI-10.4 addresses PETs as design choices for data subject privacy, while AI-7.3 addresses inversion as an adversarial threat. The same technical mechanism may satisfy both requirements.*

**[V.AI-10.4] PRIVACY-ENHANCING TECHNIQUES**

The AI system shall verify privacy-enhancing techniques by analysis and test.

The analysis shall confirm that the privacy-enhancing techniques applied are appropriate to the personal data sensitivity, processing purpose, and re-identification risk; that the selection rationale is documented including consideration of alternatives; that anonymization or pseudonymization claims are validated against the applicable regulatory standard (e.g., GDPR Article 4(5), HIPAA Safe Harbor at 45 CFR 164.514(b)(2), or Expert Determination at 164.514(b)(1)); that differential privacy parameters where applied are documented with privacy budget and composition analysis; and that encryption methods meet the cryptographic

standard declared by the project in its verification evidence (Appendix C.2), with the derivation basis for that choice documented.

The test shall confirm that applied techniques function as designed under representative operational conditions; that anonymization is robust against re-identification using documented adversary capability assumptions; and that privacy guarantees do not degrade unacceptably under operational use patterns.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that PETs are appropriate to risk and purpose, selection is documented with alternatives considered, anonymization meets the applicable regulatory standard, differential privacy parameters are documented with composition analysis, encryption meets the declared cryptographic standard, techniques function under operational conditions, and privacy guarantees are robust against documented adversary assumptions.

### 2.10.5 Cross-Border Data Transfer Controls

**[R.AI-10.5] CROSS-BORDER DATA TRANSFER CONTROLS**

The AI system shall identify and control cross-border transfers of personal data using mechanisms appropriate to the source and destination jurisdictions.

***Rationale:*** *Personal data transferred across jurisdictional boundaries is subject to additional regulatory requirements. GDPR Chapter V restricts transfers outside the European Economic Area to jurisdictions with adequacy decisions, transfers under Standard Contractual Clauses, transfers under Binding Corporate Rules, or transfers under enumerated derogations. Other jurisdictions impose analogous restrictions (e.g., China PIPL, India DPDP, Brazil LGPD). AI systems often involve transfers that are not obvious: cloud-hosted training infrastructure, third-party annotation services, regional inference endpoints, model weights derived from trans-jurisdictional training data. Each transfer point must be identified and brought under an appropriate transfer mechanism, or the system must be architected to avoid the transfer.*

**[V.AI-10.5] CROSS-BORDER DATA TRANSFER CONTROLS**

The AI system shall verify cross-border data transfer controls by inspection and analysis.

The inspection shall confirm that personal data flows are mapped across all AI lifecycle stages including data collection, training infrastructure, third-party services, model storage, and inference endpoints; that cross-border transfers are identified with source and destination jurisdictions; that transfer mechanisms are documented for each transfer (adequacy decision, SCCs, BCRs, derogations, or

other applicable mechanism); and that jurisdictional limitations of system deployment are documented in the ODD under Regulatory and Operational Authorization.

The analysis shall confirm that selected transfer mechanisms are valid under current regulatory guidance; that transfer impact assessments have been performed where required (e.g., post-Schrems II requirements for transfers to certain jurisdictions); and that contingency plans exist for transfer mechanism invalidation.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that personal data flows are mapped, cross-border transfers are identified with jurisdictions, transfer mechanisms are documented and valid, transfer impact assessments are complete where required, and contingency plans exist for mechanism invalidation.

### 2.10.6 Privacy Impact Assessment

**[R.AI-10.6] PRIVACY IMPACT ASSESSMENT**

The AI system shall conduct a Privacy Impact Assessment (PIA) or Data Protection Impact Assessment (DPIA) where required by applicable regulation, addressing privacy risks across the AI lifecycle and documenting risk mitigation measures.

***Rationale:*** *Privacy impact assessment is mandated for high-risk processing under GDPR Article 35 and recommended as a best practice across most modern privacy frameworks. AI systems frequently meet GDPR DPIA triggering criteria: systematic evaluation of personal aspects (Article 35(3)(a)), large-scale processing of special categories (Article 35(3)(b)), and systematic monitoring (Article 35(3)(c)). Beyond regulatory compliance, PIAs surface privacy risks that would otherwise be invisible: re-identification risk in anonymized datasets, membership inference susceptibility, function creep beyond original purpose, and downstream privacy harms from automated decisions. The PIA is the integrating artifact that ties together AI-10's other sub-requirements.*

**[V.AI-10.6] PRIVACY IMPACT ASSESSMENT**

The AI system shall verify privacy impact assessment by inspection.

The inspection shall confirm that a PIA or DPIA has been conducted addressing the AI system; that the assessment covers personal data flows, lawful basis, data subject rights implementation, privacy-enhancing techniques applied, cross-border transfers, and re-identification and inference attack risks; that the assessment is traceable to the declared ODD Subject Population (AI-1.0); that mitigation measures are documented for each identified risk; that the assessment

has been reviewed and approved by an authority with privacy responsibility (e.g., Data Protection Officer where required, Privacy Officer, or designated equivalent); and that the assessment is reviewed at periodic intervals or upon material change to the system.

**Success Criteria:** The verification shall be considered successful when the inspection shows that the PIA is complete and covers all required topics, is traceable to the declared ODD, includes documented mitigation measures, is approved by a privacy authority, and is subject to periodic review and change-triggered review.

### 2.10.7 Privacy Compliance Audit Trail

**[R.AI-10.7] PRIVACY COMPLIANCE AUDIT TRAIL**

The AI system shall maintain an audit trail of privacy-relevant events including data subject rights requests and dispositions, consent grants and withdrawals, cross-border transfer authorizations, identified privacy incidents, and PIA reviews and updates.

***Rationale:*** *Regulatory compliance under GDPR, HIPAA, CCPA, and analogous frameworks requires demonstrable evidence of proper privacy handling. GDPR Article 30 specifies records of processing activities; HIPAA 164.316 requires policies and procedures documentation with retention; CCPA requires verification records for data subject requests. Beyond compliance, the privacy audit trail supports incident scoping, regulatory inquiries, and demonstration of accountability. This requirement is the privacy-domain complement to the classification compliance audit trail established in AI-1.8; the two audit trails serve distinct regulatory regimes and are maintained separately even where the underlying logging infrastructure is shared.*

**[V.AI-10.7] PRIVACY COMPLIANCE AUDIT TRAIL**

The AI system shall verify privacy compliance audit trail by inspection.

The inspection shall confirm that data subject rights requests are logged with timestamps, requestor identification (or appropriate pseudonymous identifier), request type, response timeline, and disposition; that consent grants and withdrawals are logged with timestamps and consent text version; that cross-border transfer authorizations are logged with destination jurisdiction and transfer mechanism; that identified privacy incidents are logged with timestamps, discovery method, scope estimate, notification actions, and remediation; that PIA reviews and updates are logged with timestamps and approver; and that audit trail records are retained for periods consistent with applicable regulatory requirements.

**Success Criteria:** The verification shall be considered successful when the inspection shows that data subject rights, consent events, transfer authorizations, privacy incidents, and PIA reviews are logged with required content, and that retention periods meet applicable regulatory requirements under each privacy framework that applies to the system.

# Section 3: Architecture and Paradigm Requirements

Section 3 establishes requirements that apply when an AI system is built using specific architectures or follows specific learning paradigms. Unlike Section 2 (Core Requirements), which applies to every AI system regardless of underlying technology, the requirements in this section apply conditionally based on the system's design choices.

Architecture refers to structural design choices including multi-model coordination, neural network use, ensemble methods, foundation model use, and other architectural patterns that introduce specific failure modes beyond those addressed by the algorithm-agnostic core.

Paradigm refers to learning or operational pattern choices including continuous learning, federated learning, online learning, and other paradigms that introduce failure modes related to how the system learns, adapts, or operates over time.

The current architecture and paradigm requirement sets are AI-11 (Multi-Model Systems), AI-12 (Neural Networks), and AI-13 (Continuous Learning and Adaptation). Future revisions of this framework will add architecture and paradigm requirement sets as the field matures and as Safety Critical Labs begins certifying additional system classes.

## 3.1 AI-11: MULTI-MODEL SYSTEMS REQUIREMENTS

### [R.AI-11] MULTI-MODEL SYSTEMS REQUIREMENTS

The AI system will, when implemented as a multi-model system, manage inter-model coordination, combined confidence representation, cascading failure mitigation, and operator-facing display of multi-model decisions in a manner that preserves overall system reliability and operator interpretability.

Multi-model systems combine outputs from two or more AI models. Examples include a perception model feeding a planning model, multiple specialized classifiers contributing to a fused decision, ensemble architectures with voting or weighted aggregation, and retrieval-augmented generation pipelines. These systems exhibit failure modes that single-model systems do not, including

coordination drift, conflicting outputs, error propagation, combined-confidence misrepresentation, and correlated failures. The following sub-requirements apply to AI systems that incorporate two or more interacting AI models.

**Related Domain Standards**

- General: ISO/IEC 22989:2022; ISO/IEC 5338:2023
- Aerospace: NPR 7150.2D §4 (Software Lifecycle Processes); NASA-STD-8739.8B (Software Assurance); neither directly addresses multi-model AI architecture; framework-defined
- Aviation: SAE ARP4754B §4 (Development assurance for systems and items); SAE ARP6983/EUROCAE ED-324 (forward reference, Q1 2026)
- Automotive: ISO 26262-6:2018 §7 (Software architectural design); ISO 21448:2022 §5 (Specification and design)
- Medical: IEC 62304:2006+A1:2015 §5.3 (Software architectural design); FDA AI-Enabled Device Software Functions guidance Section IX (Model Description and Development)
- Industrial: ISO 13849-1 (Functional safety of control systems); not multi-model-specific; framework-defined for multi-model AI

### 3.1.1 Multi-Model Architecture Documentation

**[R.AI-11.1] MULTI-MODEL ARCHITECTURE DOCUMENTATION**

The AI system shall document the multi-model architecture including the role of each model, the data flow between models, the interface specifications for model-to-model data exchange, the combined system-level performance requirements, and the failure mode analysis addressing cascading and correlated failure scenarios.

***Rationale:*** *Multi-model architectures introduce coordination concerns that single-model systems do not face. Without explicit architecture documentation, inter-model dependencies are implicit and the system cannot be assessed against the failure modes that multi-model coordination introduces.*

**[V.AI-11.1] MULTI-MODEL ARCHITECTURE DOCUMENTATION**

The AI system shall verify multi-model architecture documentation by inspection.

The inspection shall confirm the architecture documentation identifies the role of each model, specifies the data flow between models, defines the interface specifications for model-to-model data exchange, states the combined system-

level performance requirements, and includes a failure mode analysis addressing cascading and correlated failure scenarios.

**Success Criteria:** The verification shall be considered successful when the inspection shows that the architecture documentation identifies the role of each model, specifies inter-model data flow, defines model-to-model interface specifications, states combined system-level performance requirements, and includes a failure mode analysis addressing cascading and correlated failures, and the documentation is complete, accurate, and available for assessor review.

### 3.1.2 Cascading Failure Mitigation

**[R.AI-11.2] CASCADING FAILURE MITIGATION**

The AI system shall implement and document mitigations for cascading failure scenarios in chained model architectures, including error propagation between sequential models, confidence degradation across model chains, out-of-distribution propagation where upstream OOD outputs appear in-distribution to downstream models, and latency accumulation in time-critical decision paths.

***Rationale:*** *When AI models are chained, errors can amplify or be masked, combined confidence is generally lower than individual model confidence, and OOD detection at the system entry point may not catch outputs that drift OOD between models. Latency accumulation in sequential processing affects time-critical decisions.*

**[V.AI-11.2] CASCADING FAILURE MITIGATION**

The AI system shall verify cascading failure mitigation by analysis and test.

The analysis shall confirm mitigations are documented for error propagation between sequential models, confidence degradation across model chains, out-of-distribution propagation where upstream out-of-distribution outputs appear in-distribution to downstream models, and latency accumulation in time-critical decision paths.

The test shall demonstrate that the documented mitigations operate as designed under representative cascading-failure conditions.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that mitigations are documented for error propagation between sequential models, confidence degradation across model chains, out-of-distribution propagation, and latency accumulation, and that the mitigations are demonstrated to operate as designed under representative test conditions.

### 3.1.3 Ensemble Coordination

**[R.AI-11.3] ENSEMBLE COORDINATION**

The AI system shall implement and document coordination mechanisms for ensemble and voting architectures, including controls against correlated failures (where models trained on similar data fail on the same inputs), documented voting threshold rationale (majority, unanimous, or weighted voting schemes), and disagreement handling when models disagree beyond defined thresholds.

***Rationale:*** *Ensemble systems are intended to provide robustness through model diversity, but correlated failures can defeat this intent. Voting threshold selection materially affects ensemble behavior and requires justification.*

**[V.AI-11.3] ENSEMBLE COORDINATION**

The AI system shall verify ensemble coordination by analysis and test.

The analysis shall confirm the coordination mechanisms include controls against correlated failures, documented rationale for the selected voting threshold (majority, unanimous, or weighted), and defined handling for cases where models disagree beyond established thresholds.

The test shall demonstrate that the coordination mechanisms operate per the documented design under representative scenarios, including disagreement cases.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that the coordination mechanisms include controls against correlated failures, documented voting threshold rationale, and defined disagreement handling, and that the mechanisms operate per the documented design under representative test scenarios including disagreement cases.

### 3.1.4 Combined Confidence Representation

**[R.AI-11.4] COMBINED CONFIDENCE REPRESENTATION**

The AI system shall compute and display combined confidence values that accurately represent the aggregate uncertainty across all participating models. Multi-model confidence displays shall include combined system-level confidence clearly labeled, individual model contributions visible on demand, visual indication when models disagree beyond defined thresholds, and clear mapping between confidence levels and recommended operator action per Section 5.1 (Threshold Selection Guidance).

***Rationale:*** *Operators interacting with multi-model systems need to understand both the aggregate confidence and the contribution of each model to diagnose*

*unusual outputs and exercise appropriate authority. Cascaded model architectures generally compute combined confidence as the product of individual confidences (assuming independence) or via Bayesian propagation; ensemble architectures may use weighted average, minimum (most conservative), or ensemble-specific calibration. The presentation method must support operator interpretation.*

**[V.AI-11.4] COMBINED CONFIDENCE REPRESENTATION**

The AI system shall verify combined confidence representation by test.

The test shall confirm combined confidence is computed by the documented method and accurately represents the aggregate uncertainty across all participating models, the combined system-level confidence is clearly labeled, individual model contributions are visible on demand, model disagreement beyond defined thresholds is visually indicated, and confidence levels map to recommended operator action per Section 5.1 (Threshold Selection Guidance).

**Success Criteria:** The verification shall be considered successful when the test shows that combined confidence is computed by the documented method, the combined confidence and individual model contributions are displayed as required, model disagreement is visually indicated, confidence levels map to recommended operator action, and the operator interface is validated through scenario-based testing including disagreement cases.

### 3.1.5 Multi-Model Audit Trail

**[R.AI-11.5] MULTI-MODEL AUDIT TRAIL**

The AI system shall maintain an audit trail of multi-model decisions including which models contributed to each output, the contribution weights or confidence values, the aggregation method applied, and the resulting decision rationale.

***Rationale:*** *Post-mission analysis, incident investigation, and certification surveillance require the ability to reconstruct multi-model decisions including the contribution of each constituent model.*

**[V.AI-11.5] MULTI-MODEL AUDIT TRAIL**

The AI system shall verify the multi-model audit trail by inspection.

The inspection shall confirm the audit trail records which models contributed to each output, the contribution weights or confidence values, the aggregation method applied, and the resulting decision rationale, and that records are retained per AI-1.8 (Classification Compliance Audit Trail).

**Success Criteria:** The verification shall be considered successful when the inspection shows that the audit trail records the contributing models, their contribution weights or confidence values, the aggregation method applied, and the resulting decision rationale, and that the audit trail is complete, accurate, and retained per AI-1.8 (Classification Compliance Audit Trail).

## 3.2 AI-12: NEURAL NETWORK REQUIREMENTS

### [R.AI-12] NEURAL NETWORK REQUIREMENTS

The AI system will, when implemented using neural network architectures, address the specific failure modes of neural architectures including confidence miscalibration, the absence of inspectable decision logic, training integrity concerns, adversarial vulnerability, generative hallucination, and transformation-induced behavioral drift.

Neural network architectures include deep learning models, convolutional neural networks, recurrent neural networks, transformer-based systems, and generative neural models. While the algorithm-agnostic requirements in Section 2 apply to any AI system, neural networks present specific failure modes, including the absence of inspectable decision logic, prevalence of confidence miscalibration, susceptibility to adversarial inputs, sensitivity to model transformations, and hallucination in generative outputs, that warrant dedicated requirements. The following sub-requirements apply to AI systems that incorporate neural network components. AI-12.2 (Confidence Calibration) and AI-12.4 (Training Integrity) additionally apply to AI systems implemented with other statistical machine learning models, including gradient-boosted trees, random forests, and support vector machines, which exhibit the same miscalibration and overfitting failure modes; for such systems the remaining AI-12 sub-requirements apply only where neural network components are present.

**Related Domain Standards**

- General: ISO/IEC 22989:2022; ISO/IEC 5338:2023
- Aerospace: NPR 7150.2D §4 (Software Lifecycle Processes); NASA-STD-8739.8B (Software Assurance); neither neural-network-specific; framework-defined
- Aviation: DO-178C §5.2 (Software Design Process); SAE ARP6983/EUROCAE ED-324 (forward reference, Q1 2026, AI/ML aviation)
- Automotive: ISO 26262-6:2018 §7 (Software architectural design); emerging ISO/PAS 8800 (Road Vehicles - Safety and AI) when published

- Medical: FDA AI-Enabled Device Software Functions guidance Section IX (Model Description and Development); directly addresses neural network architecture documentation
- Industrial: ISO 13849-1 (Functional safety of control systems); not neural-network-specific; framework-defined

### 3.2.1 Neural Network Architecture Documentation

**[R.AI-12.1] NEURAL NETWORK ARCHITECTURE DOCUMENTATION**

The AI system shall document neural network architecture details in addition to standard software documentation, including layer types, dimensions, and connectivity; activation functions and normalization schemes; total parameter count and memory footprint; computational requirements (FLOPs) for inference; optimizer configuration, learning rate schedule, and regularization; batch size and training duration; data augmentation techniques; hardware and framework versions used; accuracy metrics across data subsets and operational scenarios; calibration curves for confidence outputs; latency and throughput on target hardware; known failure modes and limitations; and the deployment artifact chain with traceability and cryptographic hashes at each transformation step.

***Rationale:*** *Neural network behavior depends on architectural choices, training configuration, and deployment transformations in ways not present in traditional software. Comprehensive architecture documentation is required for assessor evaluation and post-mission analysis.*

**[V.AI-12.1] NEURAL NETWORK ARCHITECTURE DOCUMENTATION**

The AI system shall verify neural network architecture documentation by inspection.

The inspection shall confirm the documentation records the network architecture (layer types, dimensions, connectivity, activation and normalization schemes, parameter count, memory footprint, and inference computational requirements), the training configuration (optimizer, learning rate schedule, regularization, batch size, training duration, data augmentation, and hardware and framework versions), the performance characterization (accuracy across data subsets and operational scenarios, calibration curves, latency and throughput on target hardware, and known failure modes and limitations), and the deployment artifact chain with traceability and cryptographic hashes at each transformation step.

**Success Criteria:** The verification shall be considered successful when the inspection shows that the architecture, training configuration, performance

characterization, and deployment artifact chain are documented as required, and the documentation is complete, accurate, and current with the deployed system.

### 3.2.2 Confidence Calibration

**[R.AI-12.2] CONFIDENCE CALIBRATION**

The AI system shall demonstrate that model confidence outputs are calibrated against actual prediction accuracy. Calibration techniques such as temperature scaling, Platt scaling, or isotonic regression shall be applied where calibration error exceeds thresholds established per Section 5.1 (Threshold Selection Guidance), and calibration shall be validated on representative held-out data.

***Rationale:*** *Neural network confidence scores (typically softmax outputs) are often poorly calibrated. A model reporting 90% confidence may be correct only 70% of the time. Operators relying on AI confidence to make decisions are misled by uncalibrated confidence outputs.*

**[V.AI-12.2] CONFIDENCE CALIBRATION**

The AI system shall verify confidence calibration by analysis and test.

The analysis shall confirm calibration error is measured against actual prediction accuracy, calibration thresholds are established per Section 5.1 (Threshold Selection Guidance), and any calibration techniques applied (such as temperature scaling, Platt scaling, or isotonic regression) are justified.

The test shall demonstrate that model confidence outputs are calibrated against actual accuracy on representative held-out data and that calibration error remains within the established thresholds.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that confidence calibration is validated against actual accuracy on representative held-out data, calibration error is documented, and any calibration techniques applied are justified.

### 3.2.3 Neural Network Explainability Methods

**[R.AI-12.3] NEURAL NETWORK EXPLAINABILITY METHODS**

The AI system shall implement explainability methods appropriate to the neural network architecture and decision criticality, with method selection documented and justified. For Safety-Critical AI functions, the AI system shall document the rationale for selecting neural network architectures over inherently interpretable alternatives (decision trees, rule-based systems, linear models) when interpretable alternatives could achieve required performance.

***Rationale:*** *Neural networks present inherent explainability challenges because decision logic is distributed across learned weights rather than explicit rules. Post-hoc methods such as SHAP, LIME, attention visualization, and layer-wise relevance propagation provide explanations of individual predictions but do not constitute proof of correct decision logic. For high-criticality functions, the choice to use a neural network rather than an interpretable alternative is itself a design decision requiring justification.*

**[V.AI-12.3] NEURAL NETWORK EXPLAINABILITY METHODS**

The AI system shall verify neural network explainability methods by inspection and test.

The inspection shall confirm explainability methods appropriate to the architecture and decision criticality are documented with selection rationale, and that any Safety-Critical use of a neural network over an inherently interpretable alternative is justified per AI-8 (Explainability).

The test shall demonstrate that the implemented explainability methods generate explanations for individual predictions as documented.

**Success Criteria:** The verification shall be considered successful when the inspection and test show that explainability methods are implemented and validated, method selection rationale is documented, and any Safety-Critical use of neural networks over interpretable alternatives is justified per AI-8 (Explainability).

### 3.2.4 Training Integrity

**[R.AI-12.4] TRAINING INTEGRITY**

The AI system shall implement and document training integrity controls including: documented overfitting indicators and the threshold beyond which retraining is required; documented relationship between model architecture complexity and training data volume; documented random seeds and training configurations to enable reproducibility; and validation set contamination controls beyond [R.AI-1.1] (Dataset Separation) and [R.AI-1.2] (Partition Use Restriction), specifically addressing hyperparameter tuning on validation data and architecture decisions informed by validation performance.

***Rationale:*** *Neural network and other statistical machine learning training introduces risks not present in traditional software development. Models with sufficient capacity can memorize training data rather than learning generalizable patterns. Hyperparameter tuning effectively incorporates validation information into the model. Stochastic training processes introduce reproducibility concerns.*

*These risks require specific controls beyond the algorithm-agnostic dataset partitioning requirements.*

**[V.AI-12.4] TRAINING INTEGRITY**

The AI system shall verify training integrity by inspection and analysis.

The inspection shall confirm training integrity controls are documented, including overfitting indicators and the retraining threshold, the relationship between architecture complexity and training data volume, random seeds and training configurations enabling reproducibility, and validation set contamination controls beyond AI-1.1 (Dataset Separation) and AI-1.2 (Partition Use Restriction).

The analysis shall confirm the documented controls were applied during development and are verifiable from training records.

**Success Criteria:** The verification shall be considered successful when the inspection and analysis show that training integrity controls are documented, applied during development, and verifiable from training records.

### 3.2.5 Neural Network Out-of-Distribution Detection Extensions

**[R.AI-12.5] NEURAL NETWORK OUT-OF-DISTRIBUTION DETECTION EXTENSIONS**

The AI system shall implement out-of-distribution detection methods appropriate to neural network architectures, recognizing that softmax confidence alone is insufficient for OOD detection. OOD detection for neural networks shall use methods that account for the tendency of neural networks to produce high-confidence predictions on inputs far from the training distribution, such as ensemble disagreement, energy-based methods, feature-space distance metrics, or deep generative model likelihood.

***Rationale:*** *Neural networks often produce high softmax confidence scores on inputs outside the training distribution. Standard OOD detection per [R.AI-6.2] requires neural-network-specific extensions to be effective.*

**[V.AI-12.5] NEURAL NETWORK OUT-OF-DISTRIBUTION DETECTION EXTENSIONS**

The AI system shall verify neural network out-of-distribution detection extensions by test.

The test shall demonstrate that the out-of-distribution detection methods are effective on representative out-of-distribution test data, account for the tendency of neural networks to produce high-confidence predictions on inputs far from the training distribution, and that the chosen method (such as ensemble disagreement, energy-based methods, feature-space distance metrics, or deep generative model

likelihood) is justified relative to the neural network architecture and extends the standard detection required by AI-6.2.

**Success Criteria:** The verification shall be considered successful when the test shows that out-of-distribution detection methods are validated on representative out-of-distribution test data and the chosen method is justified relative to the neural network architecture.

### 3.2.6 Adversarial Vulnerability Mitigation

**[R.AI-12.6] ADVERSARIAL VULNERABILITY MITIGATION**

The AI system shall implement adversarial robustness measures specific to neural network attack classes, including testing against gradient-based attacks (FGSM, PGD, C&W) and gradient-free attacks. Adversarial defenses shall be evaluated for gradient masking, where defenses appear to reduce gradient-based attack success but do not actually improve robustness.

***Rationale:*** *Neural networks are uniquely susceptible to adversarial examples: inputs with imperceptible perturbations that cause dramatic misclassification. This vulnerability is fundamental to gradient-based learning and cannot be fully eliminated. Some adversarial defenses create a false sense of security by masking gradients rather than improving robustness, requiring evaluation with gradient-free attacks. Neural-network-specific adversarial measures supplement [R.AI-7.2] (Adversarial Input Manipulation Protection).*

**[V.AI-12.6] ADVERSARIAL VULNERABILITY MITIGATION**

The AI system shall verify adversarial vulnerability mitigation by test.

The test shall demonstrate that adversarial robustness is evaluated against both gradient-based attack classes (such as FGSM, PGD, and C&W) and gradient-free attack classes, that apparent robustness is not attributable to gradient masking, and that residual vulnerability is documented, supplementing the protection required by AI-7.2 (Adversarial Input Manipulation Protection).

**Success Criteria:** The verification shall be considered successful when the test shows that adversarial robustness is validated against both gradient-based and gradient-free attack classes, gradient masking is excluded as the source of any apparent robustness, and residual vulnerability is documented.

### 3.2.7 Generative Model Hallucination Constraints

**[R.AI-12.7] GENERATIVE MODEL HALLUCINATION CONSTRAINTS**

For neural networks that generate outputs (text, trajectories, plans, code, images), the AI system shall implement domain-specific hallucination constraints including physical plausibility checks (conservation laws, kinematic limits, domain-specific bounds), cross-validation with independent sensors or models where available, confidence-bounded outputs (refusing to generate beyond confidence thresholds), and output format constraints that limit the space of generable outputs to verifiable forms.

***Rationale:*** *Generative neural models can produce plausible-seeming outputs that are factually incorrect or physically infeasible. Hallucination detection per [R.AI-5] requires generative-model-specific constraints to be effective in safety-critical contexts.*

**[V.AI-12.7] GENERATIVE MODEL HALLUCINATION CONSTRAINTS**

The AI system shall verify generative model hallucination constraints by analysis and test.

The analysis shall confirm the documented constraints include physical plausibility checks (conservation laws, kinematic limits, domain-specific bounds), cross-validation with independent sensors or models where available, confidence-bounded outputs, and output format constraints that limit generable outputs to verifiable forms, supplementing the hallucination detection required by AI-5.

The test shall demonstrate that the implemented constraints bound the system's output space within verifiable limits per the operational context.

**Success Criteria:** The verification shall be considered successful when the analysis and test show that generative output constraints are documented, implemented, and demonstrated to bound the system's output space within verifiable limits per the operational context.

### 3.2.8 Deployment Transformation Validation (Neural Network Considerations)

**[R.AI-12.8] DEPLOYMENT TRANSFORMATION VALIDATION (NEURAL NETWORK CONSIDERATIONS)**

Neural network deployment transformations shall be validated per [R.AI-4.7] (Deployment Format Validation), with neural-network-specific attention to: quantization sensitivity (per-layer impact analysis, post-quantization calibration on representative data); pruning effects (validation that pruned models do not exhibit different failure modes than original models); knowledge distillation (validation that distilled student models preserve teacher model behavior on out-of-distribution inputs and edge cases); and framework conversion (numerical equivalence

validation between source and target frameworks with documented acceptable tolerances).

***Rationale:*** *Neural networks are particularly sensitive to deployment-format transformations. Reducing precision can cause catastrophic accuracy degradation in some layers, pruning can introduce different failure modes, distillation may not preserve out-of-distribution behavior, and framework conversion can introduce numerical differences due to operator implementation variations. AI-4.7 establishes the general normative requirement; this sub-requirement specifies the neural-network-specific concerns.*

#### [V.AI-12.8] DEPLOYMENT TRANSFORMATION VALIDATION (NEURAL NETWORK CONSIDERATIONS)

The AI system shall verify deployment transformation validation by test, per [V.AI-4.7] (Deployment Format Validation).

The test shall demonstrate the validation evidence required by [V.AI-4.7], with neural-network-specific attention to quantization sensitivity (per-layer impact analysis and post-quantization calibration on representative data), pruning effects (confirmation that pruned models do not exhibit different failure modes than the original), knowledge distillation (confirmation that distilled student models preserve teacher behavior on out-of-distribution inputs and edge cases), and framework conversion (numerical equivalence between source and target frameworks within documented tolerances).

**Success Criteria:** The verification shall be considered successful when the test shows that deployment transformations satisfy [V.AI-4.7] with neural-network-specific validation evidence demonstrating that the quantization, pruning, distillation, and framework-conversion concerns are addressed.

## 3.3 AI-13: CONTINUOUS LEARNING AND ADAPTATION REQUIREMENTS

#### [R.AI-13] CONTINUOUS LEARNING AND ADAPTATION REQUIREMENTS

The AI system will, when implementing continuous learning or adaptation during operations, control the conditions under which learning occurs, validate what is being learned, preserve the ability to detect and recover from learning that produces degraded behavior, and maintain transparency over the system's learning history.

Continuously-learning systems update their model parameters during operation based on observed data. ISO/IEC 22989:2022 5.11.9.2 defines continuous learning as incremental training that occurs during operation. While continuous learning enables AI systems to adapt to operational environments and address drift

(improving mission effectiveness when properly controlled), it introduces failure modes that statically-trained systems do not exhibit, including catastrophic forgetting, learning data drift, online learning instability, and the inability to verify the operational model against a fixed validation baseline. The following sub-requirements apply to AI systems implementing continuous learning during operations.

**Related Domain Standards**

- General: ISO/IEC 22989:2022 5.11.9.2; ISO/IEC 5338:2023
- Aerospace: NPR 7150.2D §4 (Software Lifecycle Processes); NASA-STD-8739.8B (Software Assurance); current NASA guidance heavily restricts post-deployment learning; framework-defined for continuous learning
- Aviation: FAA AI Safety Roadmap, p.10 ("Differentiate Between Learned AI and Learning AI" principle); directly addresses learning-AI safety assurance methodology
- Automotive: ISO 21448:2022 §13 (Operation phase activities) for field-monitoring of evolving systems; UNECE WP.29 R157 §5.2 for post-market ALKS changes
- Medical: FDA AI-Enabled Device Software Functions guidance Section III (TPLC Approach: General Principles); FDA Predetermined Change Control Plan (PCCP) Guidance; central FDA framework for post-market AI/ML changes
- Industrial: ISO 13849-1 (Functional safety of control systems); does not address continuous learning; framework-defined

### 3.3.1 Continuous Learning Permission Criteria

**[R.AI-13.1] CONTINUOUS LEARNING PERMISSION CRITERIA**

The AI system shall declare the conditions under which continuous learning is permitted, including the system classification tier, the operational scenarios in which learning may occur, the data sources from which learning may proceed, and the controls required at each tier. Safety-Critical AI continuous learning requires rollback capability, approval for model updates prior to activation, shadow mode validation before operational use, and automated performance bounds monitoring. Mission-Critical AI continuous learning requires performance monitoring per AI-4, automated rollback at defined thresholds, logging of all model updates, and periodic review of learning trajectory. Operational Support AI continuous learning

requires baseline performance tracking and anomaly alerting for significant behavior changes.

***Rationale:*** *Continuous learning has different risk profiles depending on system classification and operational context. Permission criteria scale the required controls to the consequences of degraded learning outcomes.*

**[V.AI-13.1] CONTINUOUS LEARNING PERMISSION CRITERIA**

The AI system shall verify continuous learning permission criteria by inspection.

The inspection shall confirm the permission criteria declare the system classification tier, the operational scenarios in which learning may occur, the permitted data sources, and the controls required at each tier, and that the controls are classification-appropriate (Safety-Critical AI requiring rollback capability, pre-activation approval, shadow mode validation, and automated performance bounds monitoring; Mission-Critical AI requiring performance monitoring per AI-4, automated rollback at defined thresholds, update logging, and periodic learning-trajectory review; and Operational Support AI requiring baseline performance tracking and anomaly alerting).

**Success Criteria:** The verification shall be considered successful when the inspection shows that continuous learning permission criteria are documented, classification-appropriate, and enforced by system controls.

### 3.3.2 Runtime Learning Controls

**[R.AI-13.2] RUNTIME LEARNING CONTROLS**

The AI system shall implement runtime controls that prevent learning when defined preconditions are not met, including controls for safety-critical operational windows, controls when learning data sources are suspect, and controls when system confidence is below validated thresholds.

***Rationale:*** *Continuous learning permitted under all conditions risks incorporating poisoned data, learning during anomalous operational states, or amplifying low-confidence behavior. Runtime controls operationalize the permission criteria declared under AI-13.1.*

**[V.AI-13.2] RUNTIME LEARNING CONTROLS**

The AI system shall verify runtime learning controls by test.

The test shall demonstrate that runtime controls prevent learning when defined preconditions are not met, including during safety-critical operational windows,

when learning data sources are suspect, and when system confidence is below validated thresholds, operationalizing the permission criteria declared under AI-13.1.

**Success Criteria:** The verification shall be considered successful when the test shows that runtime learning controls prevent learning under each documented precondition violation through test scenarios.

### 3.3.3 Learning Data Validation

**[R.AI-13.3] LEARNING DATA VALIDATION**

The AI system shall validate learning data quality before incorporating it into model updates, including data freshness (recency relative to operational distribution), source legitimacy (trusted data origin), and absence of poisoning indicators (anomaly detection on learning data prior to use).

***Rationale:*** *Learning data quality directly determines what the model learns. Without validation, continuous learning is vulnerable to data poisoning, distribution shift in learning data, and incorporation of stale data that misrepresents current operational conditions.*

**[V.AI-13.3] LEARNING DATA VALIDATION**

The AI system shall verify learning data validation by test and analysis.

The analysis shall confirm the validation criteria address data freshness (recency relative to the operational distribution), source legitimacy (trusted data origin), and poisoning indicators (anomaly detection on learning data prior to use).

The test shall demonstrate that learning data failing the freshness, legitimacy, or poisoning checks is rejected before incorporation into model updates.

**Success Criteria:** The verification shall be considered successful when the test and analysis show that learning data validation is implemented and demonstrated to reject learning data that fails freshness, legitimacy, or poisoning checks.

### 3.3.4 Catastrophic Forgetting Mitigation

**[R.AI-13.4] CATASTROPHIC FORGETTING MITIGATION**

The AI system shall implement and verify mitigations against catastrophic forgetting (loss of previously-validated capabilities through new learning), using methods such as elastic weight consolidation, progressive neural networks, replay buffers, or other techniques appropriate to the model architecture and learning paradigm.

***Rationale:*** *Per ISO/IEC 22989:2022, continuous learning systems are susceptible to catastrophic forgetting where new learning overwrites previously acquired knowledge. Without mitigation, continuous learning can degrade validated capabilities even as it improves performance on new tasks.*

**[V.AI-13.4] CATASTROPHIC FORGETTING MITIGATION**

The AI system shall verify catastrophic forgetting mitigation by test.

The test shall demonstrate that the implemented mitigations (such as elastic weight consolidation, progressive neural networks, or replay buffers appropriate to the model architecture and learning paradigm) retain previously-validated capabilities after continuous learning episodes.

**Success Criteria:** The verification shall be considered successful when the test shows that catastrophic forgetting mitigations retain previously-validated capabilities after continuous learning episodes.

### 3.3.5 Learning Validation

**[R.AI-13.5] LEARNING VALIDATION**

The AI system shall validate continuous learning updates against benchmark performance criteria before the updated model becomes the operational model. Updates that fail validation shall be rejected and the prior validated model retained.

***Rationale:*** *Without pre-activation validation, learning updates that produce degraded behavior become operational before the degradation is detected. Pre-activation validation is the gate that prevents this.*

**[V.AI-13.5] LEARNING VALIDATION**

The AI system shall verify learning validation by test.

The test shall demonstrate that continuous learning updates are validated against benchmark performance criteria before becoming the operational model, that updates failing validation are rejected with the prior validated model retained, and that accepted updates meet or exceed prior performance on the benchmark suite.

**Success Criteria:** The verification shall be considered successful when the test shows that learning validation rejects updates that fail benchmark criteria, and accepted updates meet or exceed prior performance on the benchmark suite.

### 3.3.6 Rollback Procedures

**[R.AI-13.6] ROLLBACK PROCEDURES**

The AI system shall implement rollback procedures that restore a prior validated model state when continuous learning updates produce degraded performance after activation. Rollback shall be invokable both automatically (on threshold-triggered detection) and manually (by qualified operator authority).

***Rationale:*** *Pre-activation validation reduces but does not eliminate the risk of degraded learning outcomes reaching operations. Rollback procedures provide the recovery path when post-activation degradation is detected.*

**[V.AI-13.6] ROLLBACK PROCEDURES**

The AI system shall verify rollback procedures by test.

The test shall demonstrate that rollback restores a prior validated model state when continuous learning updates produce degraded performance after activation, invokable both automatically (on threshold-triggered detection) and manually (by qualified operator authority).

**Success Criteria:** The verification shall be considered successful when the test shows that rollback restores the prior validated model state both automatically under threshold scenarios and on manual operator command.

### 3.3.7 ODD Evolvability Declaration

**[R.AI-13.7] ODD EVOLVABILITY DECLARATION**

The AI system shall declare its ODD evolvability per AI-1.0, specifying whether the operational design domain is permitted to evolve through continuous learning, the bounds within which evolution may occur, and the conditions that trigger ODD reassessment.

***Rationale:*** *Continuous learning that changes the model's effective operational envelope changes the system's ODD. Without explicit ODD evolvability declaration, the relationship between continuous learning and the validated operational envelope is implicit and unverifiable.*

**[V.AI-13.7] ODD EVOLVABILITY DECLARATION**

The AI system shall verify the ODD evolvability declaration by inspection.

The inspection shall confirm the declaration specifies, per AI-1.0, whether the operational design domain is permitted to evolve through continuous learning, the bounds within which evolution may occur, and the conditions that trigger ODD reassessment, and that the declaration is consistent with the AI-1.0 ODD declaration.

**Success Criteria:** The verification shall be considered successful when the inspection shows that ODD evolvability is declared and consistent with the AI-1.0 ODD declaration, and the bounds and trigger conditions are documented.

### 3.3.8 Continuous Learning Audit Trail

**[R.AI-13.8] CONTINUOUS LEARNING AUDIT TRAIL**

The AI system shall maintain an audit trail of all continuous learning events, including data sources used for each learning event, parameter updates applied, validation outcomes, and any rollback actions taken.

***Rationale:*** *Post-incident analysis, certification surveillance, and operational accountability require complete reconstruction of the system's learning history*.

**[V.AI-13.8] CONTINUOUS LEARNING AUDIT TRAIL**

The AI system shall verify the continuous learning audit trail by inspection.

The inspection shall confirm the audit trail records the data sources used for each learning event, the parameter updates applied, the validation outcomes, and any rollback actions taken, and that records are retained per AI-1.8 (Classification Compliance Audit Trail).

**Success Criteria:** The verification shall be considered successful when the inspection shows that the continuous learning audit trail is complete, accurate, and retained per AI-1.8 (Classification Compliance Audit Trail).

# Section 4: Verification

## 4.0 OVERVIEW

Section 4 establishes the verification methods, evidence requirements, and verification matrix that implement the [V.AI-x.y] verification statements declared under each Section 2 and Section 3 sub-requirement. The methods defined in 4.1 are applied per the matrix in 4.2; cadence (one-time, periodic, or continuous) is indicated per row to reflect lifecycle phase coverage including post-deployment Operations and Sustainment per Table 1-7. Section 4.4 elaborates the continuous and periodic verification activities required for deployed AI systems. Section 4.5 establishes the verification independence required by system classification.

## 4.1 VERIFICATION METHODS

Verification of AI-specific requirements uses standard methods defined in IEEE 1012-2024, with tailoring for AI/ML system characteristics.

| Method | Definition | AI Application |
|---|---|---|
| Inspection | Examination of documentation, artifacts, or code | Dataset documentation, provenance records, log schemas |
| Analysis | Evaluation using mathematical modeling or analytical techniques | Bias risk assessment, coverage analysis, distribution characterization |
| Demonstration | Qualitative observation of system behavior under realistic conditions | Operator comprehension assessment, human-AI interaction evaluation |
| Test | Execution of the system or component under controlled conditions with defined inputs, producing quantitative measurements evaluated against pass/fail criteria | Drift detection, OOD detection, hallucination detection |

## 4.2 VERIFICATION MATRIX SUMMARY

The verification matrix traces each sub-requirement to its verification method and required evidence artifact.

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| AI-1.0 | Operational Design Domain declaration | Inspection | One-time | ODD Declaration Record |
| AI-1.1 | Dataset separation | Inspection | One-time | Partition documentation |
| AI-1.2 | Partition use restriction | Inspection, Analysis | One-time | Access logs, partition audit |
| AI-1.3 | Data provenance | Inspection | One-time | Provenance records |
| AI-1.4 | Ground truth labeling quality | Inspection, Analysis | One-time | Labeling QA report |
| AI-1.5 | Pre-ingestion classification screening | Inspection, Analysis | Continuous | Screening report, source certification |
| AI- | Information | Inspection, | Continuous | Inheritance |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| 1.6 | Classification inheritance | Analysis | | documentation |
| AI-1.7 | Output classification awareness | Test, Analysis | Continuous | Output marking validation |
| AI-1.8 | Classification compliance audit trail | Inspection | Continuous | Audit trail records |
| AI-1.9 | AI hazard analysis integration | Inspection, Analysis | One-time | AI hazard analysis records |
| AI-1.10 | Retrieval corpus controls | Inspection, Test | Continuous | Corpus control and screening records |
| AI-2.1 | Bias-bounded baseline performance | Analysis, Test | One-time | Validation report |
| AI-2.2 | Bias-managed training data | Inspection | One-time | Dataset representation report |
| AI-2.3 | Training dataset validation | Inspection, Analysis | One-time | Dataset audit report |
| AI-2.4 | Continuous bias monitoring | Test | Continuous | Bias monitoring logs |
| AI-2.5 | Bias impact assessment | Analysis, Test | One-time | Impact assessment records |
| AI-2.6 | Bias indicator definition | Inspection | One-time | Indicator and threshold record |
| AI-2.7 | Bias threshold alerting | Test | One-time | Alert test report |
| AI-2.8 | Bias threshold response | Test | One-time | Response pathway test report |
| AI-2.9 | Bias event logging | Inspection | Continuous | Bias event logs |
| AI-2.10 | Safety-critical bias prohibition | Analysis, Test | One-time | Protective action test report |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| AI-2.11 | Safety-critical bias escalation | Test | One-time | Escalation test report |
| AI-2.12 | Bias edge case definition | Inspection | One-time | Edge case catalog |
| AI-2.13 | Bias edge case validation | Test | One-time | Edge case test report |
| AI-3.1 | Test coverage metric definition | Inspection | One-time | Coverage metric documentation |
| AI-3.2 | Test coverage threshold achievement | Analysis, Test | One-time | Coverage report |
| AI-3.3 | Operational input distribution testing | Test | One-time | Distribution test report |
| AI-3.4 | Boundary and edge case testing | Test | One-time | Boundary test report |
| AI-4.1 | Data drift monitoring | Test | Continuous | Drift detection logs |
| AI-4.2 | Performance threshold maintenance | Analysis, Test | Continuous | Performance monitoring records |
| AI-4.3 | Model maintenance criteria | Inspection | Periodic | Update procedures, version history |
| AI-4.4 | Output validity boundaries | Test | Continuous | Boundary test report |
| AI-4.5 | Periodic model validation | Test | Periodic | Validation records |
| AI-4.6 | Operational validation without ground truth | Inspection, Analysis | Continuous | Proxy validation documentation |
| AI-4.7 | Deployment format validation | Test | One-time per transformation | Side-by-side performance comparison report |
| AI-5.1 | Hallucination criteria definition | Inspection | One-time | Criteria documentation |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| AI-5.2 | Hallucination detection | Test | Continuous | Detection test report |
| AI-5.3 | Hallucination response | Test | Continuous | Response test report |
| AI-5.4 | Hallucination logging | Inspection | Continuous | Hallucination event logs |
| AI-5.5 | Independent output validation | Analysis, Test | One-time | Independent validation report |
| AI-6.1 | Training distribution characterization | Inspection | One-time | Distribution documentation |
| AI-6.2 | Runtime OOD detection | Test | Continuous | OOD test report |
| AI-6.3 | OOD response | Test | One-time | OOD response test report |
| AI-6.4 | OOD event logging | Inspection | Continuous | OOD event logs |
| AI-7.1 | Data poisoning protection | Analysis, Test | One-time | Poisoning test report |
| AI-7.2 | Adversarial input protection | Test | Periodic | Adversarial test report |
| AI-7.3 | Model inversion and extraction protection | Inspection, Analysis | Periodic | Access control and privacy report |
| AI-7.4 | Model integrity | Inspection | One-time | Integrity verification logs |
| AI-7.5 | Adversarial event logging | Inspection | Continuous | Security event logs |
| AI-7.6 | AI supply chain security | Inspection | One-time | Supply chain risk assessment |
| AI-7.7 | Artifact load safety | Inspection, Test | One-time | Load path inspection record; load rejection test report |
| AI- | Explainability | Inspection | One-time | Explainability |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| 8.1 | requirements definition | | | requirements |
| AI-8.2 | Decision explanation generation | Demonstration | One-time | Explanation usability assessment |
| AI-8.3 | Confidence indication | Test | Periodic | Confidence calibration report |
| AI-8.4 | Reasoning inspection capability | Test | One-time | Inspection capability test report |
| AI-8.5 | Public disclosure support | Inspection | One-time | Disclosure artifacts and release review records |
| AI-9.1 | Human decision authority | Test | One-time | Authority test report |
| AI-9.2 | Operational situational awareness | Test | One-time | SA test report |
| AI-9.3 | Trust calibration | Analysis, Demonstration | One-time | Human factors assessment |
| AI-9.4 | Graceful degradation | Test | One-time | Degraded mode test report |
| AI-9.5 | Human-AI interaction logging | Inspection | Continuous | Interaction logs |
| AI-9.6 | Operational role verification | Inspection, Analysis | Continuous | Periodic review records |
| AI-9.7 | Operator qualification | Inspection, Test | Periodic | Qualification records, role-assignment controls |
| AI-9.8 | Workload management | Analysis, Test | Continuous | Workload parameter analysis and scenario test results |
| AI-9.9 | Training program requirements | Inspection | Periodic | Training curricula, records, and update |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| | | | | procedures |
| AI-10.1 | Personal data identification and minimization | Inspection, Analysis | One-time | PII identification and minimization record |
| AI-10.2 | Lawful basis and consent management | Inspection | Continuous | Lawful basis and consent records |
| AI-10.3 | Data subject rights implementation | Inspection, Test | One-time | DSR procedures and response evidence |
| AI-10.4 | Privacy-enhancing techniques | Analysis, Test | Continuous | PET selection and validation report |
| AI-10.5 | Cross-border data transfer controls | Inspection, Analysis | Continuous | Data flow map, transfer mechanism documentation |
| AI-10.6 | Privacy Impact Assessment | Inspection | Continuous | PIA/DPIA record |
| AI-10.7 | Privacy compliance audit trail | Inspection | Continuous | Privacy event audit trail |
| AI-11.1 | Multi-model architecture documentation | Inspection | One-time | Architecture documentation package |
| AI-11.2 | Cascading failure mitigation | Analysis, Test | One-time | Cascading failure analysis and test report |
| AI-11.3 | Ensemble coordination | Analysis, Test | One-time | Ensemble coordination test report |
| AI-11.4 | Combined confidence representation | Test | One-time | Combined confidence calibration report |
| AI-11.5 | Multi-model audit trail | Inspection | Continuous | Multi-model decision logs |
| AI-12.1 | Neural network architecture documentation | Inspection | One-time | Architecture documentation package |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| AI-12.2 | Confidence calibration | Analysis, Test | Periodic | Calibration validation report |
| AI-12.3 | Neural network explainability methods | Inspection, Test | One-time | Explainability method validation report |
| AI-12.4 | Training integrity | Inspection, Analysis | One-time | Training integrity records |
| AI-12.5 | Neural network OOD detection extensions | Test | Continuous | OOD extension test report |
| AI-12.6 | Adversarial vulnerability mitigation | Test | Periodic | Adversarial robustness test report |
| AI-12.7 | Generative model hallucination constraints | Analysis, Test | Continuous | Output constraint test report |
| AI-12.8 | Deployment transformation validation (neural network considerations) | Test | One-time per transformation | Side-by-side performance comparison report |
| AI-13.1 | Continuous learning permission criteria | Inspection | One-time | Permission criteria documentation |
| AI-13.2 | Runtime learning controls | Test | Continuous | Learning control test report |
| AI-13.3 | Learning data validation | Test, Analysis | Continuous | Learning data validation logs |
| AI-13.4 | Catastrophic forgetting mitigation | Test | Periodic | Regression test report |
| AI-13.5 | Learning validation | Test | Periodic | Learning validation records |
| AI-13.6 | Rollback procedures | Test | One-time | Rollback test report |
| AI- | ODD evolvability | Inspection | One-time | ODD Declaration |

| Req ID | Requirement Summary | Method | **Cadence** | Evidence Artifact |
|---|---|---|---|---|
| 13.7 | declaration | | | Record (evolvability section) |
| AI-13.8 | Continuous learning audit trail | Inspection | Continuous | Learning event logs |

## 4.3 DEPLOYMENT FORMAT VALIDATION

Verification of deployment-format model transformations is established normatively under V.AI-4.7. Refer to AI-4.7 for the verification statement, success criteria, and applicable domain standards. The corresponding row in the Section 4.2 Verification Matrix lists the verification method (Test) and required evidence artifact (side-by-side performance comparison report).

## 4.4 CONTINUOUS AND PERIODIC VERIFICATION

A subset of verification activities established under Section 2 cannot be satisfied by a one-time pre-deployment activity. These activities operate continuously or on a defined periodicity throughout the Operations and Sustainment lifecycle phase per Table 1-7. The Cadence column of the Section 4.2 Verification Matrix identifies which sub-requirements carry this obligation.

Continuous verification activities include:

- Bias monitoring per AI-2.4, evaluating performance metrics across declared Subject Population strata as the system processes operational data.
- Performance threshold maintenance per AI-4.2 and drift detection per AI-4.1, comparing live performance to validated baselines.
- Operational validation without ground truth per AI-4.6, applying surrogate validation methods where outcome data is delayed or unavailable.
- Out-of-distribution detection per AI-6.2, flagging inputs outside the declared Operational Design Domain at runtime.
- Operator role and authority verification per AI-9.6, confirming that human decision authority remains correctly assigned during operations.

Periodic verification activities include:

- Model maintenance review per AI-4.3 and periodic model validation per AI-4.5, conducted on the cadence established by mission risk per Section 5.1.

- Operator qualification renewal per AI-9.7 and training program review per AI-9.9, conducted on the schedule established by the operating organization.
- Adversarial robustness re-testing per AI-7.x, conducted when the threat environment evolves or on the schedule established by the security program.

Continuous and periodic verification evidence shall be retained per AI-1.8 (Classification Compliance Audit Trail) and shall be available for review during certification surveillance, incident investigation, or recertification activities.

## 4.5 VERIFICATION INDEPENDENCE

Verification activities shall be performed with independence scaled by system classification, as specified in the table below. Independence applies to the execution and disposition of the [V.AI-x.y] verifications defined in Sections 2 and 3, including the judgment that project-defined thresholds (Section 1.5) have been met.

| Classification | Required Independence |
|---|---|
| Safety-Critical AI | Verification performed by individuals who did not develop the artifact under verification, with the verification function organizationally distinct from the development function. Verification results dispositioned by the independent verification function. |
| Mission-Critical AI | Verification performed by individuals who did not develop the artifact under verification. Organizational independence recommended but not required. |
| Operational Support AI | Self-verification permitted. Verification evidence independently reviewed by the Software Assurance authority prior to closure. |

This independence scaling follows the pattern established for software verification independence in DO-178C and for independent verification and validation in IEEE 1012. Certification assessment under this framework additionally provides third-party examination of the verification evidence itself; that examination does not substitute for the project-level independence required above.

# Section 5: Implementation Patterns

## 5.1 THRESHOLD SELECTION GUIDANCE

This section provides informative guidance for threshold selection. It does not establish normative requirements. Threshold documentation obligations are established normatively in Section 1.5; adequacy of selected values is evaluated during certification assessment. Projects should document threshold selection rationale as part of verification evidence.

Section 2 requires projects to define and document thresholds for bias metrics, drift detection, performance monitoring, and other AI-specific measures. This section provides example threshold ranges and selection considerations to assist projects in establishing appropriate values. These examples represent typical industry practice and should be tailored based on mission risk tolerance, operational context, and system classification.

### 5.1.1 Example Threshold Ranges by Requirement

The following table provides example threshold ranges for selected verification metrics. Values are illustrative starting points; projects should confirm appropriate values through operational testing and document selection rationale per AI-1.8 (Classification Compliance Audit Trail).

| Req | Metric | Example | Selection Considerations |
|---|---|---|---|
| **AI-2** | Demographic parity difference | ≤2% for safety-critical; ≤5% for mission-critical | Tighter thresholds for crew health/safety functions; consider protected class sample sizes |
| **AI-2** | Equalized odds difference | ≤3% TPR/FPR difference across groups | Critical for detection/diagnostic systems where false negatives have safety implications |
| **AI-4** | Data drift (PSI) | <0.1 negligible; 0.1-0.25 moderate; >0.25 significant | Alert at moderate; halt autonomous operation at significant for safety-critical |
| **AI-4** | Data drift (KL divergence) | >0.1 warrants investigation; >0.2 requires action | Monitor per feature; aggregate thresholds may mask localized drift |
| **AI-4** | Performance degradation | ≤5% accuracy drop from baseline | Baseline established during V&V; consider mission phase |

| Req | Metric | Example | Selection Considerations |
|---|---|---|---|
| | | | sensitivity |
| **AI-5** | Hallucination detection rate | ≥95% detection with ≤1% false positive rate | Balance detection sensitivity against operational false alarm burden |
| **AI-3** | Input space coverage | ≥95% of ODD-derived input equivalence classes (safety-critical); ≥90% (mission-critical) | Derive equivalence classes from declared ODD attributes (AI-1.0); complement with boundary and edge case testing per AI-3.4; applicable regardless of model architecture |
| **AI-6** | OOD detection rate | ≥97% detection with ≤2% false positive rate | Validate against realistic OOD scenarios from operational environment |
| **AI-8** | Confidence calibration error | ECE ≤0.05; reliability slope 0.9-1.1 | Critical for operator trust calibration; validate across operational scenarios |
| **AI-9** | State change notification | ≤500ms; confidence drop >15% triggers alert | Align with human factors reaction time requirements |
| **AI-9** | Manual mode transition | ≤2 operator actions; ≤5 seconds | Validate under time pressure; include in crew training scenarios |
| **AI-1.5** | Pre-ingestion classification screening detection rate | ≥95% Safety-Critical; ≥90% Mission-Critical; ≥85% Operational Support | Detection rate measured against known-positive controlled content. Values proposed; confirm with operational testing. |
| **AI-4.7** | Deployment transformation performance margin | ≤5% performance delta vs. original model (Safety-Critical); ≤10% (Mission-Critical); ≤15% (Operational Support); document and accept with justification | Margins measured per relevant performance metric (accuracy, F1, calibration, etc.) on representative test data. Values proposed; confirm with operational testing. |

### 5.1.2 Scaling Thresholds by Classification

Threshold stringency should scale with AI system classification:

**Safety-Critical AI:** Most stringent thresholds. Conservative margins. Multiple independent detection mechanisms recommended. Thresholds derived from hazard analysis and safety margins.

**Mission-Critical AI:** Thresholds balanced against mission objectives. May accept higher false positive rates to ensure detection. Thresholds derived from mission success criteria.

**Operational Support AI:** Thresholds may be relaxed where operational impact is limited. Focus on trend detection over absolute thresholds. Thresholds derived from operational efficiency considerations.

### 5.1.3 Threshold Documentation

When documenting threshold selection rationale, projects should capture:

- Selected threshold value and units
- Derivation basis (analysis, heritage, industry standard, simulation)
- Relationship to mission risk tolerance and safety margins
- System response when threshold is exceeded
- Validation evidence demonstrating threshold appropriateness
- Conditions under which threshold should be revisited

## 5.2 AI-SPECIFIC HAZARD ANALYSIS INTEGRATION

This section provides informative guidance for integrating AI-specific failure modes into hazard analysis. The normative requirement to perform and integrate this analysis is established in AI-1.9 (AI Hazard Analysis Integration); this section guides its implementation and does not itself establish normative requirements.

### 5.2.1 Purpose

Traditional hazard analysis methods (FMEA, FTA, HAZOP) assume deterministic failure modes. AI systems introduce probabilistic, emergent, and adversarial failure modes that require augmented analysis approaches.

### 5.2.2 AI Failure Mode Categories

When conducting hazard analysis for AI systems, include the following failure mode categories:

- **Data-Related:** bias, drift, poisoning, leakage
- **Model-Related:** hallucination, overconfidence, underfitting, overfitting
- **Operational:** OOD inputs, concept drift, degradation
- **Adversarial:** input manipulation, model theft, poisoning attacks
- **Privacy:** re-identification of training data subjects, membership inference attacks revealing training set composition, training-data leakage through model outputs (inversion attacks), inferred attribute disclosure beyond intended outputs.
- **Human Factors:** automation bias, over-trust, under-trust

## 5.3 HUMAN-AI TEAMING IMPLEMENTATION GUIDANCE

This section provides implementation guidance for the human-AI teaming requirements established under AI-9. Effective human-AI teaming requires more than well-designed interfaces; it requires deliberate attention to operator workload, trust calibration, role clarity, and the conditions under which authority transfers between human and AI.

### 5.3.1 Operator Workload and Attention Management

Implementation of [R.AI-9.8] (Workload Management) should consider both cognitive load (the mental demand of monitoring AI behavior) and attention demand (the frequency and duration of operator engagement). Workload assessment methods include NASA-TLX subjective ratings, secondary-task performance measures, and physiological indicators where available. AI systems intended for sustained operation should demonstrate sustainable workload across the longest contiguous operating period in the declared ODD.

### 5.3.2 Trust Calibration

Calibrated trust, where operator reliance on AI matches the AI's actual reliability, is the goal. Both over-trust (automation bias) and under-trust (rejection of valid AI recommendations) are failure modes. Implementation of [R.AI-9.3] (Trust Calibration) should include confidence indicators that are themselves calibrated (per [R.AI-8.3]), consistent presentation of AI uncertainty, and operator training that explicitly addresses both failure modes. Verification activities should include

scenario-based exercises measuring operator response to both AI failures and AI successes outside the operator's expectation.

### 5.3.3 Operator Role Verification

[R.AI-9.6] requires periodic verification that operator role assignments remain consistent with the original safety analysis. Drift can occur when operating organizations consolidate roles, when shift schedules change the personnel filling roles, or when the operational tempo deviates from the design assumption. Verification should be cadenced per the operating organization's change-control process and at minimum on each system upgrade.

### 5.3.4 Operator Qualification and Training

[R.AI-9.7] (Operator Qualification) and [R.AI-9.9] (Training Program Requirements) together establish the qualification and training infrastructure. Training programs should include both initial qualification and recurrent qualification on a defined cadence, with explicit coverage of: AI system capability limits, the declared ODD and its implications, escalation procedures when AI behavior departs from expectation, and the human authority preserved under [R.AI-9.1].

### 5.3.5 Authority Transfer

Where AI systems and operators share decision authority, the conditions under which authority transfers between them should be unambiguous and trained. Implementation should specify: which actions remain under human authority unconditionally per [R.AI-9.1], which decisions the AI may execute autonomously, the conditions that trigger AI escalation to human authority, and the conditions that allow operators to override AI recommendations or revoke AI authority. Ambiguous authority allocation is a known root cause of automation surprises and should be eliminated by design.

## 5.4 PRIVACY AND DATA PROTECTION IMPLEMENTATION GUIDANCE

This section provides implementation guidance for the privacy and data protection requirements established under AI-10. AI systems often process personal data at scale and may infer attributes operators did not explicitly intend to expose. Implementation should anticipate both the data the system processes directly and the inferences it can derive.

### 5.4.1 Data Minimization

Implementation of [R.AI-10] data minimization principles should begin at the data ingestion boundary. Where the operational task can be accomplished without personal data, personal data should not enter the AI processing path. Where

personal data is required, the minimum sufficient subset should be retained and the remainder discarded or replaced with derived features. Data minimization should be a design-time decision, not a post-hoc filter.

### 5.4.2 De-Identification and Re-Identification Risk

De-identification techniques (k-anonymity, l-diversity, differential privacy, synthetic data generation) should be selected based on the threat model and the system's classification tier. Safety-Critical AI processing personal data should assume re-identification attempts will occur and select techniques whose privacy guarantees survive the assumed adversary capabilities. Re-identification risk should be re-evaluated whenever the input data distribution or the auxiliary data available to potential adversaries materially changes.

### 5.4.3 Access Controls and Audit

Access to personal data within the AI processing path should be controlled, logged, and auditable. Access logs should be retained per the AI-1.8 Classification Compliance Audit Trail and should include both human access (developers, operators, auditors) and automated access (training pipelines, model serving infrastructure, monitoring systems). Anomalous access patterns should trigger investigation.

### 5.4.4 Data Retention

Personal data retention should be limited to the minimum period required for the operational task and any legally mandated retention. Retention schedules should be documented per data category and enforced through automated deletion where possible. Aggregated or de-identified outputs derived from personal data may have different retention requirements than the source data; both schedules should be explicit.

### 5.4.5 Inference and Membership Concerns

AI systems can leak training data through inversion attacks, membership inference attacks, and confidence-based extraction techniques. Implementation should consider: differential privacy or related techniques during training where the threat model warrants, output filtering or noise injection at inference time for sensitive applications, and adversarial testing for membership inference vulnerabilities for systems trained on sensitive data. The risk surface depends on model architecture, training data sensitivity, and adversary access; all three should inform the mitigation strategy.

### 5.4.6 Cross-Border Data Transfer

Where personal data crosses jurisdictional boundaries, applicable data protection regulations (GDPR, regional equivalents) apply in addition to the AI-10 requirements. Implementation should document the legal basis for transfer, the safeguards applied, and the data subject rights preserved. Cross-border considerations should be addressed at design time, not as a post-hoc compliance overlay.

## Appendix A: GLOSSARY

The glossary contains 90 terms organized into 10 topical categories: Core AI/ML, System Classification, Operational Design Domain, Data, Risk and Failure Mode, Verification and Validation, Operational, Security, Explainability and Trust, and Privacy and Data Protection. Definitions are framework-defined or mapped to international standards where applicable; each entry indicates its source. Cross-references appear at the end of related entries.

### CORE AI/ML TERMS

AI System: Engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives. (ISO/IEC 22989:2022 3.1.4)

Artificial Intelligence (AI): A system that, for a given set of human-defined objectives, generates outputs such as predictions, recommendations, classifications, or decisions, and that exhibits the capacity to perform tasks commonly associated with human cognition (e.g., perception, reasoning, learning, decision-making) under varying conditions. (Adapted from ISO/IEC 22989:2022 3.1.4; modified to add capacities associated with human cognition. ISO/IEC 22989 3.1.3 defines artificial intelligence as the discipline.)

Continuous Learning: Incremental training of an AI system that takes place on an ongoing basis during the operation phase of the AI system life cycle. (ISO/IEC 22989:2022 3.1.9)

Determination Boundary: A defined and bounded collection of hardware and software items, together with its declared interfaces, on which the AI applicability determination of Section 1.2.1 is performed. Framework requirements attach to learned components within the boundary and to artifacts whose correctness criterion is fidelity to a learned component; obligations on the integrating system attach where the respective requirements specify. Adapted from the AI/ML constituent concept of the EASA Artificial Intelligence Concept Paper Issue 2; modified to a domain-agnostic determination unit for framework applicability.

Format Conversion: A model transformation that translates a model between framework representations (e.g., PyTorch to ONNX to TensorRT) to enable execution on target runtimes. Subject to revalidation per AI-4.7.

Human-Machine Teaming: Integration of human interaction with machine intelligence capabilities. (ISO/IEC 22989:2022 3.3.3)

Hyperparameter: Characteristic of a machine learning algorithm that affects its learning process and is selected prior to training. (ISO/IEC 22989:2022 3.3.4)

Inference: Reasoning by which conclusions are derived from known premises. (ISO/IEC 22989:2022 3.1.17)

Knowledge Distillation: A model transformation that trains a smaller “student” model to approximate the behavior of a larger “teacher” model, producing a compact model with comparable functional behavior. Subject to revalidation per AI-4.7.

Learned Artifact: An artifact whose values are determined or improved from data by machine learning (ISO/IEC 22989:2022 3.3.5), whether before deployment, through retraining (3.3.10), or through continuous learning during operation (3.1.9), and that serves as the specification of system behavior without a documented derivation constituting a correctness argument for that behavior. A trained model (3.3.14) is the paradigm case. A model designed for continuous learning is a learned artifact at every point in the operation phase, including prior to its first operational update. Data-fitted parameters within explicitly written model structures whose derivation constitutes a correctness argument are not learned artifacts (Table 1-2b). Framework-defined to make the Section 1.2.1 determination decidable, drawing on ISO/IEC 22989 process terminology.

Learned Component: A component whose behavior is specified by one or more learned artifacts. Framework requirements attach to learned components within the determination boundary and to artifacts whose correctness criterion is fidelity to a learned component (Section 1.2.1). Framework-defined.

Machine Learning: Process of optimizing model parameters through computational techniques, such that the model’s behavior reflects the data or experience. (ISO/IEC 22989:2022 3.3.5)

Machine Learning Model: Mathematical construct that generates an inference or prediction based on input data. (ISO/IEC 22989:2022 3.3.7)

Model: Physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data. (ISO/IEC 22989:2022 3.1.23)

Parameter (Model Parameter): Internal variable of a model that affects how it computes its outputs (weights in a neural network). (ISO/IEC 22989:2022 3.3.8)

Pruning: A model transformation that removes model weights, neurons, or layers determined to contribute minimally to model output, reducing model size and computational cost. Subject to revalidation per AI-4.7.

Quantization: A model transformation that reduces the numerical precision of model weights and/or activations (e.g., FP32 to INT8) to reduce memory footprint and computational cost, typically for execution on resource-constrained hardware. Subject to revalidation per AI-4.7.

Retraining: Updating a trained model by training with different training data. (ISO/IEC 22989:2022 3.3.10)

Trained Model: Result of model training. (ISO/IEC 22989:2022 3.3.14)

Training / Model Training: Process to determine or to improve the parameters of a machine learning model, based on a machine learning algorithm, by using training data. (ISO/IEC 22989:2022 3.3.15)

## SYSTEM CLASSIFICATION TERMS

**Mission-Critical AI:** AI outputs affect operational success but not human safety. (Framework-defined, Section 1.2.2)

**Operational Support AI:** AI supports operations but does not drive critical decisions. (Framework-defined, Section 1.2.2)

**Safety-Critical AI:** AI outputs directly affect human safety or system survivability. (Framework-defined, Section 1.2.2)

## OPERATIONAL DESIGN DOMAIN TERMS

ODD Declaration: The formal documentation of an AI system's Operational Design Domain, including environmental conditions, input characteristics, operational scenarios, and constraints under which the system is designed to function as validated. Per AI-1.0. (Framework-defined; aligned with SAE J3016 ODD concept.)

ODD Evolvability: The declared property of an Operational Design Domain that specifies whether and how the ODD may change over the system's operational life through continuous learning, retraining, or reconfiguration. Systems with non-evolvable ODD declarations operate within a frozen specification; evolvable ODDs require additional controls per AI-13. (Framework-defined.)

Operational Claim: Single-sentence summary statement that opens an ODD declaration, summarizing the function, subject population, operational conditions, user role, and regulatory regime of the certified system. (Framework-defined, AI-1.0)

Operational Design Domain (ODD): The operational conditions, inputs, users, subjects, regulatory context, and constraints under which an AI system is designed and validated to operate. The ODD defines the scope of the certification claim and the envelope within which reliable behavior is expected. (Framework-defined, Section 1.2.3 and AI-1.0; concept aligned with ISO 34503 and SAE J3016) See also ODD Declaration; ODD Evolvability; Operational Claim.

Subject Population: The population of persons or entities that an AI system's decisions act upon, distinct from the user population that operates the system. Examples include patients in a medical decision support system, pedestrians and cyclists in a vehicle perception system, or obstacle classes in an autonomous platform. (Framework-defined, AI-1.0) See also User Population.

User Population: The set of human operators, controllers, supervisors, or other authorized personnel who interact with the AI system to direct, monitor, or oversee its operation. Distinct from Subject Population, which refers to entities the AI system acts upon. (Framework-defined; see Subject Population.)

## DATA TERMS

Ground Truth: Value of the target variable for a particular item of labelled input data. (ISO/IEC 22989:2022 3.2.7)

Input Data: Data for which an AI system calculates a predicted output or inference. (ISO/IEC 22989:2022 3.2.9)

Label: Target variable assigned to a sample. (ISO/IEC 22989:2022 3.2.10)

Production Data: Data acquired during the operation phase of a deployed AI system for which a deployed system calculates a predicted output or inference. (ISO/IEC 22989:2022 3.2.12)

Test Data: Data used to assess the performance of a final model. (ISO/IEC 22989:2022 3.2.14)

Training Data: Data used to train a machine learning model. (ISO/IEC 22989:2022 3.3.16)

Validation Data: Data used to compare the performance of different candidate models. (ISO/IEC 22989:2022 3.2.15)

## RISK AND FAILURE MODE TERMS

Bias: Systematic difference in treatment of certain objects, people, or groups in comparison to others. (ISO/IEC 22989:2022 3.5.4)

Concept Drift: Change in the relationship between input data and the target variable over time (i.e., the decision boundary moves), requiring relabeling of training data and model retraining. (ISO/IEC 5338:2023 6.4.14 / ISO/IEC 22989:2022 5.11.9, tailored) See also Continuous Validation.

Data Drift: Change in the statistical characteristics of production data over time compared to the training data distribution, which can degrade model prediction accuracy. (ISO/IEC 5338:2023 6.4.14 / ISO/IEC 22989:2022 5.11.9, tailored) See also Continuous Validation.

Hallucination: The production of confidently stated but erroneous or false content. (NIST AI 600-1 §2.2 (Confabulation))

Model Drift: Degradation in model performance over time due to data drift, concept drift, or other factors affecting the relationship between inputs and outputs. (Tailored)

Out-of-Distribution: Input data that falls outside the statistical distribution of the training data, for which model predictions may be unreliable or undefined. (Tailored from ISO/IEC 23053:2022) See ISO/IEC 23053:2022 §8.7 (Operation), which discusses data drift, concept drift, and generalization affecting model performance.

Overfitting: Creating a model which fits the training data too precisely and fails to generalize. (ISO/IEC 23053:2022 3.2)

Performance Drift: Degradation of AI model performance metrics over operational time, typically arising from data drift or concept drift. Continuous monitoring per AI-4.4 and AI-4.6 detects performance drift; periodic revalidation per AI-4.5 quantifies it. (Framework-defined; complements Data Drift and Concept Drift.) See also Continuous Validation.

## VERIFICATION AND VALIDATION TERMS

Certified Configuration: The set of elements declared by the applicant and recorded by Safety Critical Labs on the certificate, to which certification attaches: the determination boundary and classification, the Operational Design Domain, the threshold values, the attack pattern sets declared under AI-7.1 and AI-7.2, and the foundational cybersecurity baseline. A change to any element is a change to the certified configuration. (Framework-defined, Section 1.6) See also Conformity Basis; Operational Design Domain.

Conformity Basis: The requirement statements of this framework on which the certification determination rests: each [R.AI-x.y] requirement statement verified per its paired [V.AI-x.y] verification statement, together with the procedural requirements of Sections 1 and 4. No document outside the framework forms part of the conformity basis unless a requirement statement names it and states the extent of the invocation. (Framework-defined, Section 1.6) See also Certified Configuration.

Continuous Validation: Ongoing monitoring process to verify that AI models continue performing satisfactorily during operation, including detection of data drift, concept drift, and performance degradation. (ISO/IEC 5338:2023 6.4.14, tailored) See also Data Drift; Concept Drift; Performance Drift.

Control Point: A defined verification or authorization gate in the AI system lifecycle through which an artifact, decision, or change must pass before proceeding to the next stage. Used in cross-references between AI-4.3 (Model Maintenance Criteria), AI-4.7 (Deployment Format Validation), and AI-7.4 (Model Integrity). (Framework-defined.)

Deployment-Format Model: An AI model that has been transformed from its trained form for execution on operational hardware through processes such as quantization, pruning, knowledge distillation, or framework conversion. Subject to revalidation per AI-4.7. (Framework-defined.)

Fairness Metrics: Quantitative measures used to assess whether an AI system produces equitable outputs across different groups, operational contexts, and mission scenarios. Includes metrics for disparate impact, equal opportunity, demographic parity, and performance consistency across operational conditions. (Tailored from ISO/IEC TR 24027)

System Safety Process: The system-level safety analysis and assessment activity through which a project identifies hazards and produces the hazard analyses that software-level standards consume, conducted per applicable domain practice (e.g., SAE ARP4754/ARP4761 safety assessment in civil aviation; ISO 26262-3 hazard analysis and risk assessment in automotive; ISO 14971 risk management for medical devices; MIL-STD-882E system safety in defense; the NASA safety program hazard analysis process, from which NASA-STD-8739.8 derives software safety criticality). Distinct from the software standard this framework supplements (e.g., DO-178C, NPR 7150.2D, ISO 26262-6), which consumes system safety outputs to determine software criticality but does not itself conduct the system-level hazard analysis. (Framework-defined, AI-1.9 and Section 1.2.5)

Test: Execution of the system or component under controlled conditions with defined inputs, producing quantitative measurements evaluated against pass/fail criteria.

Validation: Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled. (ISO/IEC 22989:2022 3.5.18)

Verification: Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled. (ISO/IEC 22989:2022 3.5.17)

## OPERATIONAL TERMS

Autonomous Action: System action taken without requiring human approval or intervention at the time of execution. (Tailored)

Controllability: Property of an AI system that allows a human or another external agent to intervene in the system's functioning. (ISO/IEC 22989:2022 3.5.6)

Fallback Behavior: Predefined safe system response invoked when AI outputs are unreliable, unavailable, or flagged as out-of-distribution. (Tailored)

Human Override: Capability for human operators to supersede or reject AI recommendations or autonomous actions, with documented rationale. (Tailored)

Reliability: Property of consistent intended behavior and results. (ISO/IEC 22989:2022 3.5.9)

Resilience: Ability of a system to recover operational condition quickly following an incident. (ISO/IEC 22989:2022 3.5.10)

Robustness: Ability of a system to maintain its level of performance under any circumstances. (ISO/IEC 22989:2022 3.5.12)

## SECURITY TERMS

Adversarial Attack: Deliberate manipulation of AI system inputs, training data, or model parameters intended to cause incorrect or harmful outputs. (ISO/IEC 5338:2023 6.4.3)

Adversarial Example: An input crafted by perturbing a legitimate input with imperceptible or low-magnitude modifications designed to cause AI model misclassification or incorrect output. Distinct from Adversarial Attack, which refers to the broader class of attack techniques. Per AI-7 and AI-12.6.

Adversarial Robustness: Ability of an AI system to maintain correct and safe operation despite adversarial inputs or attacks. (ISO/IEC 22989:2022 3.5.12, tailored)

Data Poisoning: Injection of malicious data into training datasets to influence model behavior in attacker-controlled ways. (ISO/IEC 5338:2023 6.4.3)

Input Manipulation: Crafting of inputs designed to cause model misclassification or incorrect outputs while appearing valid to human observers. (ISO/IEC 5338:2023 6.4.3)

Model Inversion: Attack technique that reconstructs sensitive training data by analyzing model outputs. (ISO/IEC 5338:2023 6.4.3)

Model Theft: Extraction of model parameters or behavior through systematic querying to replicate proprietary AI capabilities. (ISO/IEC 5338:2023 6.4.3)

## EXPLAINABILITY AND TRUST TERMS

Confidence: Quantified measure of certainty or probability associated with an AI output or prediction. (Tailored)

Explainability: Property of an AI system to express important factors influencing the AI system results in a way that humans can understand. (ISO/IEC 22989:2022 3.5.7)

Graceful Degradation: System capability to maintain reduced functionality when AI is impaired. (Tailored)

Predictability: Property of an AI system that enables reliable assumptions by stakeholders about the output. (ISO/IEC 22989:2022 3.5.8)

Transparency: Property of a system that appropriate information about the system is made available to relevant stakeholders. (ISO/IEC 22989:2022 3.5.15)

Trust Calibration: Alignment between operator trust in AI and actual AI reliability. (Human factors)

## PRIVACY AND DATA PROTECTION TERMS

Anonymization: The process of rendering personal data such that the data subject is no longer identifiable, by any means reasonably likely to be used. Regulatory standards for anonymization vary across jurisdictions, including GDPR Recital 26, HIPAA Safe Harbor at 45 CFR 164.514(b)(2), and HIPAA Expert Determination at 45 CFR 164.514(b)(1). (Framework-defined; aligned with GDPR Recital 26.) See also De-identification.

Data Controller: The natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data. (GDPR Article 4(7))

Data Minimization: The principle and practice of collecting, processing, and retaining only the personal data strictly required for the AI system's operational task, with the minimum sufficient subset kept and the remainder discarded or replaced with derived features. Per AI-10 and Section 5.4.1. (Aligned with GDPR Article 5(1)(c) and ISO/IEC 27559.)

Data Processor: A natural or legal person, public authority, agency, or other body which processes personal data on behalf of the controller. (GDPR Article 4(8))

Data Subject: The identified or identifiable natural person to whom personal data relates. (GDPR Article 4(1))

De-identification: The process of modifying personal data to reduce the likelihood of associating it with specific individuals, encompassing techniques such as masking, generalization, perturbation, and aggregation. Distinct from Anonymization in regulatory frameworks (HIPAA, CCPA) where de-identification implies a quantified risk threshold. Per AI-10 and Section 5.4.2. See also Anonymization; Re-identification Risk.

Differential Privacy: A mathematical framework for quantifying the privacy risk of a computation over a dataset, characterized by a privacy budget parameter (epsilon) that bounds the contribution of any individual record to the output. (Framework-defined; see Dwork et al. 2006 foundational literature) Full citation: Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). "Calibrating Noise to Sensitivity in Private Data Analysis," in Theory of Cryptography Conference (TCC 2006), Lecture Notes in Computer Science, vol. 3876, Springer, pp. 265-284.

DPIA (Data Protection Impact Assessment): A structured assessment of the privacy risks of a data processing activity, required under GDPR Article 35 and analogous regulations for high-risk processing. For AI systems processing personal data, DPIA is typically performed at design time and updated when the data processing materially changes. See also Privacy Impact Assessment (PIA).

Lawful Basis: The legal justification for processing personal data under an applicable privacy framework. Examples include the six bases enumerated in GDPR Article 6, HIPAA permitted uses and disclosures under 45 CFR 164.502, and CCPA business purposes. (Aligned with GDPR Article 6.)

Membership Inference: An attack class in which an adversary determines whether a specific data record was included in the training set of an AI model, typically by

exploiting model confidence patterns or output behavior. Per AI-10 and Section 5.4.5.

Personal Data: Any information relating to an identified or identifiable natural person. The precise scope varies by regulatory framework. Examples include GDPR Article 4(1), HIPAA Protected Health Information under 45 CFR 160.103, and CCPA "personal information" under Cal. Civ. Code 1798.140. (Aligned with GDPR Article 4(1).)

Privacy Impact Assessment (PIA): Systematic assessment of the privacy risks associated with a processing activity and the mitigation measures applied. Under GDPR Article 35 and for high-risk processing, referred to as a Data Protection Impact Assessment (DPIA). (Framework-defined, AI-10.6; GDPR Article 35)

Privacy-Enhancing Technique (PET): Technical method designed to reduce privacy risk through data or computation transformation, including anonymization, pseudonymization, differential privacy, federated learning, secure multi-party computation, and homomorphic encryption. (Framework-defined, AI-10.4) See ISO/IEC 27559:2022, "Information security, cybersecurity and privacy protection - Privacy enhancing data de-identification framework."

Pseudonymization: The processing of personal data in such a manner that the data can no longer be attributed to a specific data subject without the use of additional information, where that additional information is kept separately and subject to technical and organizational measures. (GDPR Article 4(5))

Re-identification Risk: The likelihood that personal data which has been de-identified or anonymized can be associated back to specific individuals through techniques such as auxiliary data linkage, attribute inference, or longitudinal tracking. Per AI-10 and Section 5.4.2.

Special Categories of Personal Data: Categories of personal data subject to additional protection under GDPR Article 9 (racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for identification, health data, sex life or sexual orientation) or analogous categories under other frameworks. Examples include HIPAA PHI and CCPA "sensitive personal information." (Aligned with GDPR Article 9.)

## Appendix B: AI CLASSIFICATION CHECKLIST

This checklist supports determination of AI system classification per Section 1.2.

## B.1 AI PRESENCE DETERMINATION

Before answering, declare the determination boundary per Section 1.2.1 (the bounded collection of hardware and software items under determination and its declared interfaces) and record it in Appendix C.1. The declaration shall name the encompassing system and enumerate every candidate learned component within it; every candidate learned component shall lie within a declared determination boundary, and a negative determination discharges only the boundary for which it was made. The determination is evidence-cited: each answer shall record the artifact or artifacts inspected (system or software requirements, design descriptions, interface control documents, model provenance records, data flow documentation). Answer question 1 first. Questions 2 and 3 record supporting indicators and do not establish AI presence on their own.

1. (Anchor) Within the declared boundary, is any system behavior specified by an artifact induced from data rather than by an explicitly written specification against which correctness can be fully verified by established methods? For behavior implemented through data-fitted parameters, apply the deletion test (Section 1.2.1): delete every value derived from data, or designed to be derived from data during operation, and inspect what remains; a surviving behavioral specification (written structure bounding behavior, values re-derivable by a documented derivation constituting a correctness argument, or generated at runtime by an explicitly specified procedure whose established analysis bounds the correctness of the resulting behavior) means the anchor criterion is not met, and no surviving specification means it is met. Parameter determination or improvement by machine learning, including retraining and continuous learning during operation, meets the anchor criterion regardless of when the optimization runs. Cite the evidence inspected. If 'No' → not AI for framework purposes; standard software assurance applies; record the negative determination and its evidence in Appendix C.1 (see Table 1-2b for exclusion examples). If 'Yes' → AI is present; answer questions 2 and 3, then proceed to B.2.
2. (Supporting indicator) Are system outputs probabilistic or non-deterministic? If 'Yes' → note verification emphasis on confidence indication and calibration (AI-8.3). Cite the evidence inspected.
3. (Supporting indicator) Can system behavior change or degrade over time without code modification? If 'Yes' → the system shall be classified no lower than Mission-Critical per Section 1.2.2. Cite the evidence inspected.

**Evaluation:** Question 1 alone determines AI presence. If question 1 is 'Yes' → AI is present; proceed to B.2. If question 1 is 'No' → this framework does not apply

regardless of the answers to questions 2 and 3; standard software assurance applies. A determination is complete only when the determination boundary is declared and each answer cites the evidence inspected; the record is maintained in Appendix C.1.

## B.2 SAFETY-CRITICAL AI DETERMINATION

1. Do AI outputs directly affect human safety or system survivability?
2. Could AI malfunction result in loss of life or critical equipment?
3. Is the AI in the control path for safety-critical functions?

**Evaluation:** If any of questions 1 through 3 above is 'Yes' → Safety-Critical AI. If all are 'No' → proceed to B.3.

## B.3 MISSION-CRITICAL AI DETERMINATION

1. Do AI outputs affect operational success (but not human safety)?
2. Could AI malfunction result in loss of operational objectives?
3. Is the AI in the decision path for mission-critical functions?

**Evaluation:** If any of questions 1 through 3 above is 'Yes' → Mission-Critical AI. If all are 'No' → Operational Support AI.

## B.4 OPERATIONAL DESIGN DOMAIN DECLARATION

Following classification determination, declare the ODD per AI-1.0. Confirm the following:

- Operational Claim summary statement drafted
- All eight ODD attribute categories addressed (Operational Environment, System Inputs, User Population, Subject Population, Operational Phases, Performance Envelope, Regulatory and Operational Authorization, Exclusions)
- Exclusions enumerated across all five sub-categories (Environments, Use Cases, Operator Conditions, Subject Conditions, System States)
- Declaration approved at the authority level specified for the system's classification per Section 1.2.3 Table 1-4

## B.5 PRIVACY APPLICABILITY DETERMINATION

- Does the system process personal data as defined under any applicable privacy framework (GDPR, HIPAA, CCPA/CPRA, PIPEDA, or analogous)?

- Does the system process special categories of personal data (e.g., GDPR Article 9, HIPAA PHI, CCPA sensitive personal information)?
- Does the system transfer personal data across jurisdictional boundaries?

**Evaluation:** If any answer is 'Yes' → AI-10 applies in full; conduct PIA/DPIA per AI-10.6. If all answers are 'No' → AI-10 sub-requirements may be documented as Not Applicable with rationale.

### B.6 REQUIREMENT APPLICABILITY

| Classification | Applicable Requirements |
|---|---|
| Safety-Critical AI | AI-1 through AI-10 (all, no tailoring without PSRB approval); plus AI-11 if multi-model, AI-12 if neural network (AI-12.2 and AI-12.4 also apply to other statistical machine learning models), AI-13 if continuous learning |
| Mission-Critical AI | AI-1 through AI-6, AI-8, AI-9, AI-10 (tailoring permitted with CE approval); AI-7 tailored; plus AI-11, AI-12, AI-13 conditionally per system design |
| Operational Support AI | AI-1 through AI-6 and AI-10 minimum; AI-7, AI-8, AI-9 as tailored; plus AI-11, AI-12, AI-13 conditionally per system design |

## Appendix C: VERIFICATION EVIDENCE TEMPLATES

### C.1 AI APPLICABILITY DETERMINATION RECORD

| Field | Entry |
|---|---|
| **Project Name** | |
| **System/Subsystem** | |
| **Determination Boundary (B.1)** | |
| **Document ID** | |
| **Date** | |
| **Prepared By** | |
| **AI Presence (B.1 Result)** | ☐ AI Present ☐ No AI |
| **Evidence Cited (B.1)** | |
| **AI Classification (B.2/B.3)** | ☐ Safety-Critical ☐ Mission-Critical ☐ Operational Support |

| Field | Entry |
|---|---|
| **Applicable Requirements** | ☐ AI-1 ☐ AI-2 ☐ AI-3 ☐ AI-4 ☐ AI-5 ☐ AI-6 ☐ AI-7 ☐ AI-8 ☐ AI-9 ☐ AI-10 ☐ AI-11 (conditional) ☐ AI-12 (conditional) ☐ AI-13 (conditional) |
| **Foundational Cybersecurity Baseline (Section 1.6)** | ☐ Attestation ☐ Existing conformity evidence (identify) |
| **Tailoring** | ☐ None ☐ See attached rationale |
| **Approved By** | |
| **Approval Date** | |

## C.2 TEST REPORT TEMPLATE

| Field | Entry |
|---|---|
| **Test Report ID** | |
| **Requirement ID(s)** | |
| **Test Title** | |
| **Test Date** | |
| **Test Environment** | |
| **Test Conductor** | |
| **Test Objective** | |
| **Pass/Fail Criteria** | |
| **Declared Set and Derivation Basis (AI-7.1, AI-7.2, AI-10.4)** | |
| **Test Results** | |
| **Pass/Fail Determination** | ☐ Pass ☐ Fail |
| **Anomalies/ Observations** | |
| **Approved By** | |

## C.3 INSPECTION RECORD TEMPLATE

| Field | Entry |
|---|---|
| **Inspection Record ID** | |

| Field | Entry |
|---|---|
| **Requirement ID(s)** | |
| **Artifact Inspected** | |
| **Inspection Date** | |
| **Inspector** | |
| **Inspection Criteria** | |
| **Findings** | |
| **Compliance Determination** | ☐ Compliant ☐ Non-Compliant ☐ Partially Compliant |
| **Corrective Action (if any)** | |

## C.4 ODD DECLARATION RECORD TEMPLATE

| Field | Entry |
|---|---|
| **Project Name** | |
| **System/Subsystem** | |
| **Document ID** | |
| **Classification** | ☐ Safety-Critical ☐ Mission-Critical ☐ Operational Support |
| **Operational Claim** | *This AI system is designed and validated to [function] for [subject population] under [operational conditions], operated by [user role] within [regulatory regime].* |
| **Operational Environment** | |
| **System Inputs (including degradation behaviors)** | |
| **User Population** | |
| **Subject Population** | |
| **Operational Phases** | |
| **Performance Envelope** | |
| **Regulatory and Operational Authorization (including ODD evolvability)** | |
| **Excluded Environments** | |
| **Excluded Use Cases** | |

| Field | Entry |
|---|---|
| **Excluded Operator Conditions** | |
| **Excluded Subject Conditions** | |
| **Excluded System States** | |
| **Approved By** | |
| **Approval Date** | |
| **Review Cadence** | |

## C.5 PRIVACY IMPACT ASSESSMENT TEMPLATE

| Field | Entry |
|---|---|
| **Project Name** | |
| **System/Subsystem** | |
| **PIA ID** | |
| **Applicable Frameworks** | ☐ GDPR ☐ HIPAA ☐ CCPA/CPRA ☐ Other: _____ |
| **Personal Data Categories Processed** | |
| **Special Category Data** | ☐ Yes ☐ No (if Yes, categories: _____) |
| **Lawful Basis (per framework)** | |
| **Data Flow Map Reference** | |
| **Cross-Border Transfers** | ☐ None ☐ See attached transfer documentation |
| **Privacy-Enhancing Techniques Applied** | |
| **Re-Identification Risk Assessment** | |
| **Model Inversion and Membership Inference Risk** | |
| **Data Subject Rights Implementation Summary** | |
| **Identified Privacy Risks and Mitigations** | |
| **Approving Privacy Authority** | |
| **Approval Date** | |
| **Next Review Date** | |

### C.6 DATA SUBJECT RIGHTS REQUEST LOG TEMPLATE

| Field | Entry |
|---|---|
| **Request ID** | |
| **Receipt Date** | |
| **Request Type** | ☐ Access ☐ Rectification ☐ Erasure ☐ Restriction ☐ Portability ☐ Objection ☐ Automated Decision Explanation |
| **Applicable Framework** | ☐ GDPR ☐ HIPAA ☐ CCPA/CPRA ☐ Other: _____ |
| **Requestor Verification Method** | |
| **Applicable Response Timeline** | |
| **Processing Steps** | |
| **Model Retraining Assessment (erasure only)** | |
| **Disposition** | ☐ Granted ☐ Partially Granted ☐ Denied (with rationale) |
| **Response Date** | |
| **Response Delivery Method** | |
| **Approved By** | |

## Appendix D: RISK SCORE METHODOLOGY

### D.1 PURPOSE

This appendix defines the quantitative risk scoring methodology used in SCL certification assessments. The risk score is an independent output of the assessment process, issued alongside the certification determination. It quantifies assessed residual risk across four dimensions and provides a reproducible, documented measure of the system's risk posture at the time of assessment.

The risk score does not determine certification outcome. A system may receive a low risk score and still receive a Not Certified determination if requirements are not

met. The score reflects assessed risk level; the determination reflects compliance status. These are independent outputs.

## D.2 SCORING DIMENSIONS

The risk score is composed of four dimensions. Each dimension is scored on a 1 to 5 scale where 1 represents the lowest risk and 5 represents the highest risk.

### D.2.1 Failure Mode Coverage

Assesses the extent to which AI specific failure modes have been identified, tested, and mitigated across all applicable requirement areas.

| Score | Criteria |
|---|---|
| 1 | All applicable failure modes identified, tested, and mitigated with verified controls. No untested failure modes remain. |
| 2 | All applicable failure modes identified and tested. Mitigations verified for safety critical and mission critical failure modes. Minor gaps in mitigation verification for lower severity modes. |
| 3 | Most applicable failure modes identified and tested. Some failure modes identified but not fully tested or mitigated. Gaps are documented with risk acceptance rationale. |
| 4 | Significant failure modes identified but testing or mitigation is incomplete. Gaps exist without documented risk acceptance. |
| 5 | Failure mode identification is incomplete. Critical failure modes may be unidentified, untested, or unmitigated. |

### D.2.2 Consequence Severity

Assesses the operational consequence if identified failure modes occur during operations, based on the system's classification level and operational context.

| Score | Criteria |
|---|---|
| 1 | Failure modes produce consequences limited to minor operational inconvenience. No impact on safety, mission success, or operational continuity. |
| 2 | Failure modes could degrade operational efficiency or produce incorrect advisory outputs. Human operators can detect and correct before operational impact. |

| Score | Criteria |
|---|---|
| 3 | Failure modes could affect mission success or operational objectives. Detection and correction is possible but depends on human oversight functioning as designed. |
| 4 | Failure modes could affect system survivability or create conditions requiring immediate human intervention to prevent escalation to safety impact. |
| 5 | Failure modes could directly affect human safety or create conditions where human intervention may be insufficient to prevent harm. |

### D.2.3 Likelihood Based on Evidence Quality

Assesses the likelihood that failure modes will occur during operations, derived from the quality, completeness, and rigor of the evidence examined during assessment.

| Score | Criteria |
|---|---|
| 1 | Evidence is comprehensive, traceable, and independently verified. Test coverage is thorough. Continuous validation mechanisms are demonstrated and operational. Strong confidence that failure modes are controlled. |
| 2 | Evidence is complete and traceable. Testing addresses primary operational scenarios. Continuous validation is operational. Minor gaps in edge case coverage or evidence traceability. |
| 3 | Evidence addresses most requirements but gaps exist in traceability, test coverage, or validation completeness. Some reliance on process controls rather than verified technical controls. |
| 4 | Evidence is incomplete or inconsistent. Significant gaps in test coverage or validation. Controls are documented but verification is limited. |
| 5 | Evidence is absent, untraceable, or contradicted by assessment findings. Little confidence that failure modes are controlled. |

### D.2.4 Human Oversight Effectiveness

Assesses the effectiveness of human oversight and override controls as documented in AI-9, including the ability of operators to detect AI errors, maintain situational awareness, and exercise decision authority.

| Score | Criteria |
|---|---|
| 1 | Human oversight is comprehensive. Decision authority is enforced through technical controls. Operators demonstrate calibrated trust, effective override capability, and graceful degradation to manual operations. |
| 2 | Human oversight is established and functional. Decision authority boundaries are defined. Minor gaps in trust calibration validation or degradation mode testing. |
| 3 | Human oversight exists but relies partially on procedural rather than technical controls. Override capability is available but has not been fully validated under operational conditions. |
| 4 | Human oversight is defined but implementation is incomplete. Operators may lack situational awareness of AI state or may not have effective override capability in all operational modes. |
| 5 | Human oversight is absent or ineffective. AI operates autonomously without functional human decision authority, or operators cannot detect AI errors in time to intervene. |

## D.3 SCORING PROCESS

The risk score is calculated by the Lead Assessor during Phase 3 of the assessment process. Each dimension is scored independently based on the evidence examined, findings documented, and system behavior observed during the assessment. Per dimension scores are documented with specific rationale referencing assessment findings.

The aggregation of dimension scores into the final risk score is defined in the SCL Assessment Process Documentation and is maintained separately from this framework to allow methodology refinement without framework revision.

## D.4 SCORE PROPERTIES

The risk score is:

**Reproducible.** Another qualified assessor working from the same evidence package and assessment findings must be able to arrive at the same per dimension scores. Scoring rationale is documented at sufficient detail to support independent verification.

**Independent.** The risk score does not determine the certification outcome. Requirements compliance determines the outcome. The risk score quantifies residual risk.

**Point in time.** The risk score reflects the assessed risk posture at the time of assessment. Operational changes, model updates, or environmental shifts after the assessment date may change the actual risk posture.

**Classification aware.** Consequence severity scoring inherently reflects the system's classification level. An Operational Support system operating in an advisory capacity will score differently on consequence severity than a Safety Critical system in the control path, even if both have similar technical implementations.

### D.5 SCORE REPORTING

The per dimension scores and aggregated risk score are reported in the Technical Assessment Report, the Certification Determination Document, and the Certificate (for Certified outcomes). For Not Certified outcomes, the per dimension scores and the aggregated risk score are reported in the Technical Assessment Report and the Certification Determination Document to inform remediation prioritization; no certificate is issued. Every assessment produces one of two outcomes, Certified or Not Certified; there is no intermediate outcome.

## DOCUMENT REVISION HISTORY

**Version numbering and change management.** This document is identified as SCL-ARF-001 independent of its version. Version increments are defined by conformity impact, not by the size of the edit: a MAJOR increment (X) when the meaning of an existing requirement or verification changes or identifiers renumber, so that existing evidence and traceability break and transition arrangements are required; a MINOR increment (X.Y) when normative content is added with every existing identifier stable, so that existing evidence remains valid and new evidence is needed only for the additions; and a PATCH increment (X.Y.Z) for editorial changes with zero change to normative meaning. Each row of the table below records the change class of the release: Editorial, Normative Addition, or Normative Modification. This policy was adopted at version 3.7; the classes shown for earlier versions are applied retrospectively for the reader and do not alter the numbers those versions were issued under. The framework is cited at its concept DOI, 10.5281/zenodo.19024420, which resolves to the latest version. Certificates cite the document identifier, the version assessed against, and that version's own DOI, never the concept DOI, so that the conformity basis of an issued certificate does not

change when a new version is published. Transition arrangements between versions are defined in the assessment methodology.

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| 1.0 | February 2025 | Safety Critical Labs | Initial release | Initial framework development |
| 2.0 | October 2025 | Safety Critical Labs | Normative Addition | Added AI-4.6 Operational Validation Without Ground Truth, AI-9.6 Operational Role Verification, Appendix D Risk Score Methodology |
| 2.1 | January 2026 | Safety Critical Labs | Normative Addition | Added AI-1.0 Operational Design Domain with Operational Claim and eight attribute categories; AI-1 renamed to Operational and Data Foundations; added Section 1.2.3 ODD procedural gate and Table 1-4 ODD Declaration Rigor; added AI-10 Privacy and Data Protection with seven sub-requirements; added Section 1.3.1 Domain Standards Citation Convention with anchor citations on all ten parent requirements and exemplar sub-requirement retrofits on AI-9.1, AI-8.3, AI-2.1; added cross-references for ODD, AI-10, and model integrity across Section 2; expanded Section 4.2 Verification Matrix to cover all sub-requirements; added Appendix C templates for ODD Declaration Record, Privacy Impact Assessment, and Data Subject Rights Request Log; added Appendix A glossary terms for ODD, Operational Claim, Subject Population, and privacy terminology; added Appendix B.4 ODD Declaration step and B.5 |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | Privacy Applicability Determination; removed Appendix E Source Traceability Matrix (traceability retained inline in Appendix A); renamed the Mission Support classification tier to Operational Support throughout the framework (tier definition, tailoring authority, applicability, and worked examples). |
| 2.1.1 | April 2026 | Safety Critical Labs | Normative Addition | Added Gap 6 human factors expansion to AI-9: AI-9.7 Operator Qualification, AI-9.8 Workload Management, AI-9.9 Training Program Requirements, each with full requirement, verification, success criteria, and Applicable Domain Standards citations. Added Gap 10 public disclosure support as new AI-8.5 Public Disclosure Support (optional by classification) covering external model cards, EU AI Act Article 13 user information, EO 14110 and OMB M-24-10 federal AI inventory, and EU AI Act Article 71 public database registration. Added corresponding rows to the Section 4.2 Verification Matrix. Sub-requirement-level Applicable Domain Standards retrofit for the approximately 53 remaining sub-requirements is scheduled as a Phase 2 pass within v2.1.1. |
| 2.1.3 | 05/04/2026 | Safety Critical Labs | Normative Addition | Added AI-4.7 Deployment Format Validation. Added Section 4.0 Overview, Section 4.4 Continuous and Periodic Verification, and Cadence column to Section 3.2 Verification Matrix. Editorial pass on |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | Sections 1-3 (terminology consistency, lexicon alignment, tracked-change cleanup). |
| 3.0 | 05/05/2026 | Safety Critical Labs | Normative Modification | Major structural revision. Added Section 3 (Architecture and Paradigm Requirements) with AI-11 Multi-Model Systems, AI-12 Neural Networks, AI-13 Continuous Learning and Adaptation. Renumbered prior Section 3 (Verification) to Section 4 and prior Section 4 (Implementation Guidance) to Section 5 (Implementation Patterns). Lifted normative content from prior Section 4.2/4.3/4.4 into new Section 3 architecture/paradigm requirements with proper [R.AI-X.Y] / [V.AI-X.Y] scaffolding. Resolves prior lexicon contradictions in Section 4 by separating normative architecture/paradigm requirements from informative implementation patterns. |
| 3.1 | 06/11/2026 | Safety Critical Labs | Editorial | Editorial and formatting revision. Bolded Section 3 verification statements ([V.AI-11.x] through [V.AI-13.x]) for consistency with Sections 1-2. Normalized sub-heading capitalization to title case. Removed residual editorial annotations. Replaced em and en dashes per house style. Corrected fused sentences, duplicated text in Table 1-3, and assorted typos. Normalized Rationale and Success Criteria formatting. Updated |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | Section 4.2 matrix cadence for AI-4.1 and AI-6.4 to Continuous. Extended Appendix C.1 requirement checklist through AI-13. Added Section 3 sub-requirements (AI-11.1 through AI-13.8) to the Section 4.2 Verification Matrix and the conditional AI-11/AI-12/AI-13 applicability to Appendix B.6, aligning both with Table 1-3. Added 1.3.3 to the table of contents. |
| 3.2 | 06/14/2026 | Safety Critical Labs | Editorial | Reformatted Section 3 (AI-11, AI-12, AI-13) verification blocks to the four-part Section 2 pattern (rationale, verification statement, method elaboration, success criteria) and aligned AI-4.7 phrasing. Verification format now uniform across AI-1 through AI-13. |
| 3.3 | 06/14/2026 | Safety Critical Labs | Editorial | Corrected the circular Test definition in the Section 4.1 verification-methods table; refined the Analysis and Demonstration definitions to sharpen the quantitative-Test versus qualitative-Demonstration distinction; added a Test entry to the Appendix A glossary; updated the IEEE 1012 citation to the 2024 edition; corrected a stale section cross-reference in the normative-references table. |
| 3.4 | 07/08/2026 | Safety Critical Labs | Normative Addition; Normative Modification | Redesigned the Section 1.2.1 applicability gate: Learned behavior is the anchor criterion, Probabilistic outputs and Potential drift are supporting indicators, Decision impact relocated to classification in |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | Section 1.2.2; added Table 1-2b negative criteria with worked examples; rewrote Appendix B.1 as a decision sequence and gave B.1 through B.3 self-contained question numbering. Added AI-1.9 AI Hazard Analysis Integration and AI-1.10 Retrieval Corpus Controls with verifications and matrix rows. Added Section 4.5 Verification Independence scaled by classification. Added classification severity anchoring and approval authority to Section 1.2.2. Added threshold documentation and adequacy language to Section 1.5. Extended AI-7.2 to prompt injection for generative and retrieval-augmented systems. Extended AI-12.2 and AI-12.4 to non-neural-network statistical machine learning models. Reworded AI-2.1 and AI-2.2 from bias-free to bias-bounded and bias-managed; corrected AI-2.3 scope to include Safety-Critical decisions; added fallback response paths to AI-2.11 and AI-5.3. Assigned the Demonstration method to AI-8.2 and AI-9.3 and aligned logging and calibration verification cadences. Corrected stale section references and miswired cross-references in Sections 1.4, 2, 4.0, 4.4, 5.3.2, AI-4.2, AI-8.5, AI-9.7, and AI-9.9. Added the Table 1-7 Retirement phase, the UL 4600 reference, Not Applicable provisions for AI-1.5 through AI-1.8, Table 1-4 rigor scaling in AI-1.0, and an input space |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | coverage exemplar replacing neuron activation coverage in Section 5.1.1. |
| 3.5 | 07/29/2026 | Safety Critical Labs | Normative Modification | Replaced the undefined term "host standard" in AI-1.9 (requirement, rationale, and verification) with the project's system safety process; added the System Safety Process definition to the Appendix A glossary, distinguishing the system-level hazard analysis process from the software standard this framework supplements; amended Section 1.2.5 to identify AI-1.9 as the framework's single interface requirement to the system safety process. Findings from the 2026-07-28 correspondence review, verified against the v3.4 text. |
| 3.6 | 08/08/2026 | Safety Critical Labs | Normative Addition | Made the Section 1.2.1 applicability determination verifiable. Promoted the key-existence principle into the anchor criterion: applicability turns on whether correctness can be fully verified against explicitly written specification by established methods, with the Table 1-2 mechanism list illustrative rather than load-bearing. Defined the determination boundary as the unit of determination, adapted from the EASA AI Concept Paper Issue 2 AI/ML constituent concept, with an explicit propagation rule: the framework attaches to learned components and to artifacts whose correctness criterion is fidelity to a learned component, and ends at |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | declared interfaces. Adopted the deletion test for data-fitted parameters as a normative determination step in Appendix B.1 with a corresponding Table 1-2b exclusion row, incorporating runtime-generated-parameter and derivation-as-correctness-argument provisos validated by a seven-case battery (GRAM, wind tunnel aerodynamic databases, gain scheduling, PID autotune, adaptive Kalman filtering with online covariance estimation, table-driven FADEC, learned-then-frozen gain tables). Added a Monte Carlo simulation and dispersion analysis row to Table 1-2b. Reworked Appendix B.1 as an evidence-cited determination with declared boundary; added Determination Boundary and Evidence Cited fields to Appendix C.1. Added the Determination Boundary glossary entry, the EASA AI Concept Paper Issue 2 reference to Table 1-10, and ARP6983/ED-324 to the Section 1.3.3 unverified citations list. Same-day review pass: added the determination boundary coverage condition; added Learned Artifact and Learned Component glossary definitions grounded in ISO/IEC 22989 process terms, clause-verified against the source standard; made the deletion test timing-neutral to cover learning during operation; conditioned the runtime parameter proviso on |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | correctness-bounding analysis with an explicit machine learning and continuous learning carve-out; added an adaptive-variant pointer to the classical state estimators row; corrected the Appendix A Parameter definition to internal variable per 3.3.8 and re-attributed the Artificial Intelligence entry as an adaptation of 3.1.4. |
| 3.7 | 09/15/2026 | Safety Critical Labs | Normative Addition; Normative Modification (Appendix D.5) | Added Section 1.6 Conformity Basis, stating that the certification determination rests on the framework's requirement statements verified per their paired verification statements and on nothing outside the document, that domain standards crosswalks carry no conformity burden, that the foundational cybersecurity posture is a declared prerequisite carried on the certificate rather than assessed scope, and defining the certified configuration. Partitioned the references: Table 1-9 now lists only the normative references with the extent of each invocation (ISO/IEC 22989:2022, ISO/IEC 5338:2023, IEEE 1012-2024) and Table 1-10 the informative references; renamed every "Applicable Domain Standards" subsection to "Related Domain Standards" and restated the citation convention in Section 1.3.1, naming Safety Critical Labs as the party that performs the crosswalk during assessment. Replaced |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | "known attack patterns" in V.AI-7.1 and V.AI-7.2 with a project-declared attack pattern set documented with its derivation basis, adjudicated during assessment, and recorded as an element of the certified configuration; applied the same construction to the indirect prompt injection patterns in V.AI-1.10 and the cryptographic standard in V.AI-10.4; added the corresponding Appendix C.1 and C.2 template fields. Reframed the AI-7.4 rationale to name the two threat cases (tampering after validation; hostile origin surviving integrity verification). Added AI-7.7 Artifact Load Safety, binding serialization formats and loader configurations on every artifact load path, with verification by inspection and negative test and a Section 4.2 matrix row. Rewrote the Section 1.2 scope statement and aligned Section 1.1 so that classification selects requirements rather than gating entry, consistent with Table 1-1. Adopted the version numbering and change management policy stated above, added the Change Class column, and assigned the document identifier SCL-ARF-001. Removed "Conditionally Certified" from Appendix D.5 and stated the two-outcome rule, with Not Certified outcomes reporting per dimension and aggregated scores in the Technical Assessment Report |

| Version | Date | Author | Change Class | Description |
|---|---|---|---|---|
| | | | | and the Certification Determination Document. Corrected the Appendix A term count, added the Certified Configuration and Conformity Basis glossary entries, and settled the spelling of de-identification throughout. |