Using the AISVS

The Artificial Intelligence Security Verification Standard (AISVS) defines security requirements for modern AI applications and services, focusing on aspects within the control of application developers.

The AISVS is intended for anyone developing or evaluating the security of AI applications, including developers, architects, security engineers, and auditors. This chapter introduces the structure and use of the AISVS, including its verification levels and intended use cases.

Artificial Intelligence Security Verification Levels

The AISVS defines three ascending levels of security verification. Each level adds depth and complexity, enabling organizations to tailor their security posture to the risk level of their AI systems.

Organizations may begin at Level 1 and progressively adopt higher levels as security maturity and threat exposure increase.

Definition of the Levels

Each requirement in AISVS v1.0 is assigned to one of the following levels:

Level 1 requirements

Level 1 includes the most critical and foundational security requirements. These focus on preventing common attacks that do not rely on other preconditions or vulnerabilities. Most Level 1 controls are either straightforward to implement or essential enough to justify the effort.

Level 2 requirements

Level 2 addresses more advanced or less common attacks, as well as layered defenses against widespread threats. These requirements may involve more complex logic or target specific attack prerequisites.

Level 3 requirements

Level 3 includes controls that are typically harder to implement or situational in applicability. These often represent defense-in-depth mechanisms or mitigations against niche, targeted, or high-complexity attacks.

Role (D/V)

Each AISVS requirement is marked according to the primary audience:

- D – Developer-focused requirements
- V – Verifier/auditor-focused requirements
- D/V – Relevant to both developers and verifiers

C1 Training Data Governance & Bias Management

Control Objective

Training data must be sourced, handled, and maintained in a way that preserves provenance, security, quality, and fairness. Doing so fulfils legal duties and reduces risks of bias, poisoning, or privacy breaches that show up during training that could effect the entire AI lifecycle.

C1.1 Training Data Provenance

Maintain a verifiable inventory of all datasets, accept only trusted sources, and log every change for auditability.

#1.1.1 Level: 1 Role: D/V
Verify that an up-to-date inventory of every training-data source (origin, steward/owner, licence, collection method, intended use constraints, and processing history) is maintained.

#1.1.2 Level: 1 Role: D/V
Verify that training data processes exclude unnecessary features, attributes, or fields (e.g., unused metadata, sensitive PII, leaked test data).

#1.1.3 Level: 2 Role: D/V
Verify that all dataset changes are subject to a logged approval workflow.

#1.1.4 Level: 3 Role: D/V
Verify that datasets or subsets are watermarked or fingerprinted where feasible.

C1.2 Training Data Security & Integrity

Restrict access to training data, encrypt it at rest and in transit, and validate its integrity to prevent tampering, theft, or data poisoning.

#1.2.1 Level: 1 Role: D/V
Verify that access controls protect training data storage and pipelines.

#1.2.2 Level: 2 Role: D/V
Verify that all access to training data is logged, including user, time, and action.

#1.2.3 Level: 2 Role: D/V
Verify that training datasets are encrypted in transit and at rest, using industry-standard cryptographic algorithms and key management practices.

#1.2.4 Level: 2 Role: D/V
Verify that cryptographic hashes or digital signatures are used to ensure data integrity during training data

storage and transfer.

#1.2.5 Level: 2 Role: D/V
Verify that that automated detection techniques are applied to guard against unauthorized modifications or corruption of training data.

#1.2.6 Level: 2 Role: D/V
Verify that obsolete training data is securely purged or anonymized.

#1.2.7 Level: 3 Role: D/V
Verify that all training dataset versions are uniquely identified, stored immutably, and auditable to support rollback and forensic analysis.

C1.3 Training Data Labeling Quality, Integrity, and Security

Protect labels and require technical review for critical data.

#1.3.1 Level: 2 Role: D/V
Verify that cryptographic hashes or digital signatures are applied to label artifacts to ensure their integrity and authenticity.

#1.3.2 Level: 2 Role: D/V
Verify that labeling interfaces and platforms enforce strong access controls, maintain tamper-evident audit logs of all labeling activities, and protect against unauthorized modifications.

#1.3.3 Level: 3 Role: D/V
Verify that sensitive information in labels is redacted, anonymized, or encrypted at the data field level at rest and in transit.

C1.4 Training Data Quality and Security Assurance

Combine automated validation, manual spot-checks, and logged remediation to guarantee dataset reliability.

#1.4.1 Level: 1 Role: D
Verify that automated tests catch format errors and nulls on every ingest or significant data transformation.

#1.4.2 Level: 2 Role: D/V
Verify that LLM training and fine-tuning pipelines implement poisoning detection & data integrity validation (e.g., statistical methods, outlier detection, embedding analysis) to identify potential poisoning attacks (e.g., label flipping, backdoor trigger insertion, role-switching commands, influential instance attacks) or unintentional data corruption in training data.

#1.4.3 Level: 2 Role: D/V
Verify that automatically generated labels (e.g., via LLMs or weak supervision) are subject to confidence thresholds and consistency checks to detect hallucinated, misleading, or low-confidence labels.

#1.4.4 Level: 3 Role: D/V
Verify that appropriate defenses, such as adversarial training (using generated adversarial examples), data augmentation with perturbed inputs, or robust optimization techniques, are implemented and tuned for relevant models based on risk assessment.
#1.4.5 Level: 3 Role: D
Verify that automated tests catch label skews on every ingest or significant data transformation.

C1.5 Data Lineage and Traceability

Track the full journey of each data point from source to model input for auditability and incident response.

#1.5.1 Level: 2 Role: D/V
Verify that the lineage of each data point, including all transformations, augmentations, and merges, is recorded and can be reconstructed.
#1.5.2 Level: 2 Role: D/V
Verify that lineage records are immutable, securely stored, and accessible for audits.
#1.5.3 Level: 2 Role: D/V
Verify that lineage tracking covers synthetic data generated via privacy-preserving or generative techniques and that all synthetic data is clearly labeled and distinguishable from real data throughout the pipeline.

References

- NIST AI Risk Management Framework
- EU AI Act – Article 10: Data & Data Governance
- CISA Advisory: Securing Data for AI Systems
- OpenAI Privacy Center – Data Deletion Controls