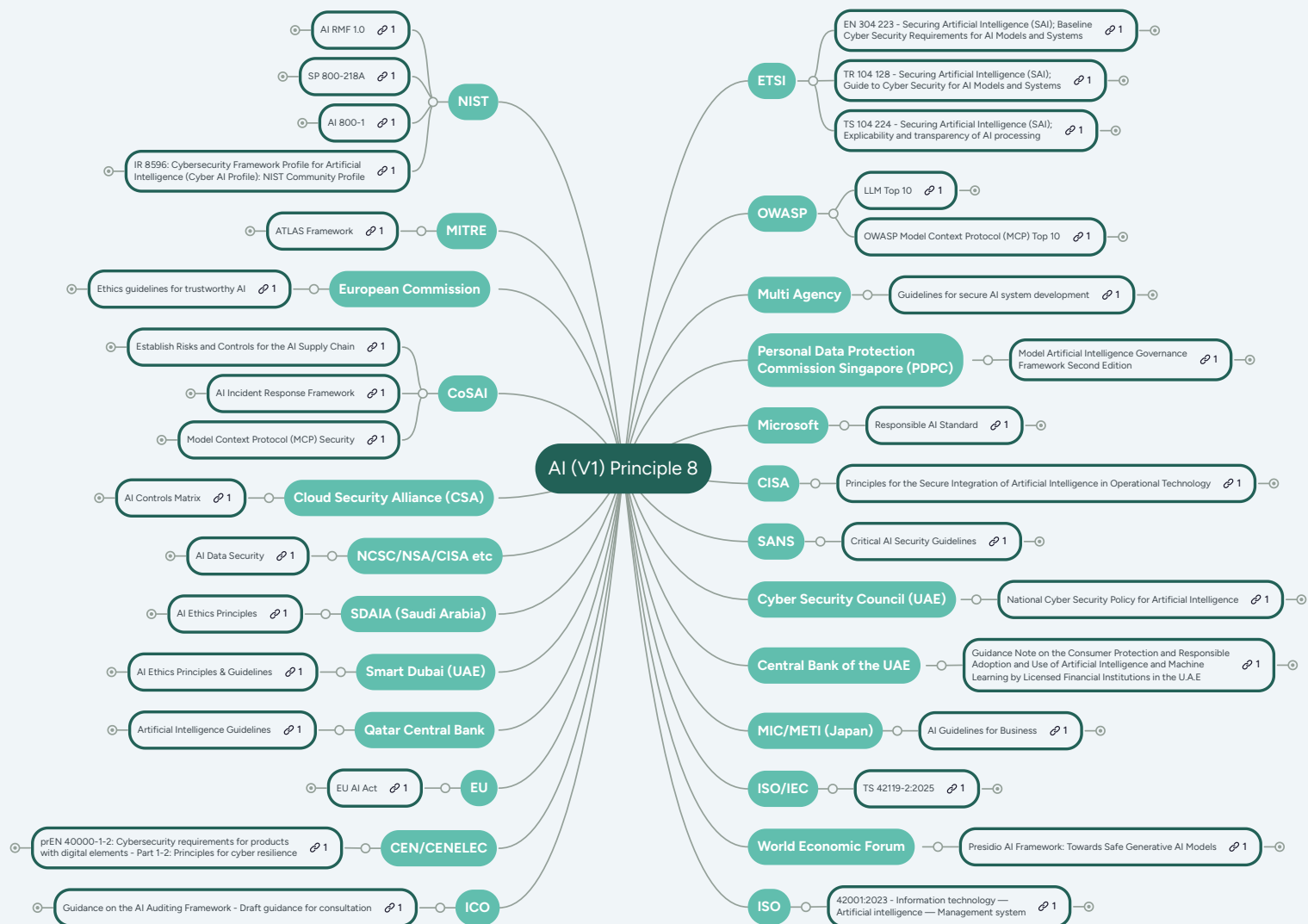




AI (V1) Principle 8

1. ETSI

1.1. EN 304 223 - Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

1.1.1. Provision 5.2.4-1

Developers shall document and maintain a clear audit trail of their system design and post-deployment maintenance plans. Developers should make the documentation available to the downstream System Operators and Data Custodians.

1.1.2. Provision 5.2.4-1.1

Developers should ensure that the document includes security-relevant information, such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, retention time, suggested review frequency and potential failure modes.

1.1.3. Provision 5.2.4-1.2

Developers shall release cryptographic hashes for model components that are made available to other stakeholders to allow them to verify the authenticity of the components.

1.1.4. Provision 5.2.4-2

Where training data has been sourced from publicly available sources, there is a risk that this data might have been poisoned. As discovery of poisoned data is likely to occur after training (if at all), Developers shall document how they obtained the public training data, where it came from and how that data is used in the model.

1.1.5. Provision 5.2.4-2.1

The documentation of training data should include at a minimum the source of the data, such as the URL of the scraped page, and the date/time the data was obtained. This will allow Developers to identify whether a reported data poisoning attack was in their data sets.

1.1.6. Provision 5.2.4-3

Developers should ensure that they have an audit log of changes to system prompts or other model configuration (including prompts) that affect the underlying working of the systems. Developers can make this available to any System Operators and End-Users that have access to the model.

1.2. TR 104 128 - Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

1.2.1. Provision 5.2.4-1

"Developers shall document and maintain a clear audit trail of their system design and post-deployment maintenance plans. Developers should make the documentation available to the downstream System Operators and Data Custodians." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without thorough documentation and audit trails, AI system operations can lack transparency, limiting traceability and increasing risks of security gaps, regulatory non-compliance, and operational inefficiencies.

Example Measures/Controls:

Develop Comprehensive System Design and Maintenance Documentation: Document the system design and postdeployment maintenance plans with audit trail of design decisions, architectural diagrams, and maintenance schedules. Ensure the documentation is accessible to downstream System Operators and Data Custodians, highlighting responsibilities, version control, and any dependencies.

1.2.2. Provision 5.2.4-1.1

"Developers should ensure that the document includes security-relevant information, such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, retention time, suggested review frequency and potential failure modes." (ETSI TS 104 223 [i.1])

Related threats/risks:

Lack of detailed security-relevant documentation can lead to unmitigated risks from unverified data sources, misuse of the AI system, and challenges in addressing failure modes.

Example Measures/Controls:

Include Relevant Security - Information in System Documentation: Document AI systems including intended scope, limitations, known failure modes, prompts, guardrails training data sources, data retention policies, review schedules. Use Model Cards to capture transparent summaries of a model's capabilities, ethical considerations, performance metrics, and limitations, consider using ML Bills of Materials (ML BOMs) to document in machinereadable way the model and its dependencies, with component versions, and licensing.

1.2.3. Provision 5.2.4-1.2

"Developers shall release cryptographic hashes for model components that are made available to other stakeholders to allow them to verify the authenticity of the components." (ETSI TS 104 223 [i.1])

Related threats/risks:

Absence of cryptographic hashes for model components increases the risk of tampering, unauthorized modifications, and integrity issues, compromising trust and security.

Example Measures/Controls:

Include Cryptographic Hashes for Models: Release cryptographic hashes for all models made available to stakeholders, enabling verification of component authenticity.

1.2.4. Provision 5.2.4-2

"Where training data has been sourced from publicly available sources, there is a risk that this data might have been poisoned. As discovery of poisoned data is likely to occur after training (if at all), Developers shall document how they obtained the public training data, where it came from and how that data is used in the model." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without comprehensive documentation of training data sources and collection timestamps, organizations can be unable to determine whether their AI models have been affected by data poisoning, fraud, or insider abuse and whether they are compliant to data protection legislation. If data poisoning attacks are revealed to have occurred on public websites, producers or users of models will only know if they are affected if they have robust documentation of what data the model was trained on.

Example Measures/Controls:

Document the process of sourcing public training data: Detailing how data was collected, processed, and used in the model (e.g. pretraining, fine-tuning). Include poisoning mitigation measures such as data validation and anomaly detection. Maintain an audit trail to trace datasets if poisoning is suspected.

1.2.5. Provision 5.2.4-2.1

"The documentation of training data should include at a minimum the source of the data, such as the URL of the scraped page, and the date/time the data was obtained. This will allow Developers to identify whether a reported data poisoning attack was in their data sets." (ETSI TS 104 223 [i.1])

Related threats/risks:

Failure to record detailed metadata, such as data sources and timestamps, can hinder efforts to trace and respond to data poisoning incidents.

Example Measures/Controls:

Document metadata for all publicly sourced training data: Include the exact source (e.g. URLs) and date/time of collection. Ensure this metadata is stored in an accessible format to facilitate traceability in case of reported data poisoning.

1.2.6. Provision 5.2.4-3

"Developers should ensure that they have an audit log of changes to system prompts or other model configuration (including prompts) that affect the underlying working of the systems. Developers can make this available to any System Operators and End-Users that have access to the model." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without an audit log for configuration changes, tracking and understanding modifications to prompts or model configurations become difficult, increasing risks of unintended system behaviour or accountability issues.

Example Measures/Controls:

Maintain an Audit Log of Prompt and Model Configuration Changes: Implement an audit log that captures all changes to system prompts and configuration settings, recording details such as change date, user ID, and a description of modifications. Provide access to System Operators and Data Custodians.

1.3. TS 104 224 - Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing

Link: https://www.etsi.org/deliver/etsi_ts/104200_104299/104224/01.01.01_60/ts_104224v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_ts/104200_104299/104224/01.01.01_60/ts_104224v010101p.pdf

1.3.1. 5.2 (a)

The statement of system purpose is critical in allowing a layperson to clearly understand the intent of the system and the role of AI in achieving that purpose or intent.

1.3.2. 5.2 (b)

The statement of system purpose should be written in natural language and be concise as well as precise (i.e. not open to variations in interpretation).

The following characteristics shall be identifiable in the statement of system purpose:

- Unambiguous: it should be impossible to interpret the system purpose in more than one way.
- Complete: the system purpose should contain all the information necessary to understand it without requiring reference to other documents.
- Precise: the system purpose should be worded clearly and exactly, without unnecessary detail that might confuse the reader.
- Well-structured: any individual elements of the system purpose should be included in an appropriate and easy-to-read manner.

1.3.3. 6.4.1

The system documentation shall clearly identify those events that are logged for future analysis. The content of the log record shall be sufficient to identify the trigger conditions of the event and should include the following:

- The time that the event occurred (including the metric and basis of the time).

EXAMPLE: Time may be system clock time, e.g. UTC time, or may be relative to some datum (e.g. clock cycles from a known datum point).

- The software versions of the processes involved or impacted in the event.
- The hardware elements involved or impacted in the event.
- The data, and its supply chain, involved or impacted in the event.

In addition the liable party for resolving the impact of the event should be clearly identified.

NOTE: In the AI Act [i.5] it is stated that high risk systems have the capability to allow for automatic recording of events over the lifetime of the AI System and this is consistent with the requirements outlined in the present document.

2. OWASP

2.1. LLM Top 10

Link: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>

Link: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>

2.1.1. LLM01: Prompt Injection - 2

Specify clear output formats, request detailed reasoning and source citations, and use deterministic code to validate adherence to these formats

2.1.2. LLM08: Vector and Embedding Weaknesses - 3

When combining data from different sources, thoroughly review the combined dataset. Tag and classify data within the knowledge base to control access levels and prevent data mismatch errors.

2.2. OWASP Model Context Protocol (MCP) Top 10

Link: <https://owasp.org/www-project-mcp-top-10/>

Link: <https://owasp.org/www-project-mcp-top-10/>

2.2.1. MCP02:2025 - Privilege Escalation via Scope Creep - 7

Strong Change Management & Audit Trails All permission changes must be tracked, reviewed, and linked to a change request or ticket. Keep immutable, tamper-evident logs tying actions to agent identity and session.

2.2.2. MCP10:2025 – Context Injection & Over-Sharing - 7

Context Access Logging

Log:

Agent ID

Context ID

Read/write events

TTL + purge events Integrate context logs into SIEM/XDR for monitoring.

3. Multi Agency

3.1. Guidelines for secure AI system development

Link: <https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf>

Link: <https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf>

3.1.1. Document your data, models and prompts

You document the creation, operation, and life cycle management of any models, datasets and meta- or system-prompts. Your documentation includes security-relevant information such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, cryptographic hashes or signatures, retention time, suggested review frequency and potential failure modes. Useful structures to help do this include model cards, data cards and software bills of materials (SBOMs). The production of comprehensive documentation supports transparency and accountability.

4. Personal Data Protection Commission Singapore (PDPC)

4.1. Model Artificial Intelligence Governance Framework Second Edition

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf>

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf>

4.1.1. Data for Model Development - a) Understanding the lineage of data

This means knowing where the data originally came from, how it was collected, curated and moved within the organisation, and how its accuracy is maintained over time. Data lineage can be represented visually to trace how the data moves from its source to its destination, how the data gets transformed along the way, where it interacts with other data, and how the representations change. There are three types of data lineage:

- i. Backward data lineage looks at the data from its end-use and backdating it to its source.
- ii. Forward data lineage begins at the data's source and follows it through to its end-use.
- iii. End-to-end data lineage combines the two and looks at the entire solution from both the data's source to its end-use and from its end-use to its source. Keeping a data provenance record allows an organisation to ascertain the quality of the data based on its origin and subsequent transformation, trace potential sources of errors, update data, and attribute data to their sources.

In some instances, the origin of data could be difficult to establish. One example could be datasets obtained from a trusted third-party which may have commingled data from multiple sources. It would be prudent for organisations to assess the risks of using such data and manage them accordingly.

4.1.2. ALGORITHM AND MODEL - a)

Model training and selection are necessary for developing an intelligent system (i.e. a system that contains AI technologies). Documenting how the model training and selection processes are conducted, the reasons for which decisions are made, and measures taken to address identified risks will enable the organisation to provide an account of the decisions subsequently.

In this regard, the field of Automated Machine Learning aims to automate a significant portion of machine learning workflows, including feature engineering, feature selection, model selection and hyper-parameter tuning. Organisations using these types of tools can consider the transparency, explainability and traceability of the automated machine learning approach, as well as the models selected.

4.1.3. ALGORITHM AND MODEL - b)

Incorporating descriptions of the solutions' design and expected behaviour into product or service descriptions and system technical specifications documentation demonstrates accountability to individuals and/or regulators. This could also include design decisions in relation to why certain features, attributes or models are selected in place of others. These steps can help provide greater clarity on an AI model by giving understandable and digestible insights into how the model operates.

Where an organisation's AI system was obtained or procured from a third-party AI solution provider, the organisation can consider requesting assistance from the AI solution provider as they may be better placed to explain how the solution functions.

5. Microsoft

5.1. Responsible AI Standard

Link: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us>

Link: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us>

5.1.1. A4.4

Define and document methods for evaluating data to be used by the system against the requirements defined in A4.1.

5.1.2. A5.5

Define and document Responsible Release Criteria to achieve this Goal. (Human oversight and control)

5.1.3. T1.1

Identify:

- 1) stakeholders who will use the outputs of the system to make decisions, and
- 2) stakeholders who are subject to decisions informed by the system.

Document these stakeholders using the Impact Assessment template.

5.1.4. T1.4

Define and document a Responsible Release Plan, to include Responsible Release Criteria to achieve this Goal.

5.1.5. T2.1

Identify:

- 1) stakeholders who make decisions about whether to employ a system for particular tasks, and
- 2) stakeholders who develop or deploy systems that integrate with this system.

Document these stakeholders in the Impact Assessment template.

5.1.6. T2.2

Publish documentation for the system so that stakeholders defined in T2.1 can understand the system. Include:

- 1) capabilities,
- 2) intended uses,
- 3) uses that require extra care or guidance,
- 4) operational factors and settings that allow for effective and responsible system use,
- 5) limitations, including uses for which the system was not designed or evaluated, and
- 6) evidence of system accuracy and performance as well as a description of the extent to which these results are generalizable across use cases that were not part of the evaluation.

When the system is a platform service made available to external customers or partners, a Transparency Note is required.

5.1.7. T2.3

Review and update documentation annually or when any of the following events occur:

- 1) new uses are added,
- 2) functionality changes,
- 3) the product moves to a new release stage,
- 4) new information about reliable and safe performance becomes known as defined by requirement RS3.3, or
- 5) new information about system accuracy and performance becomes available.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

5.1.8. F1.8

Document the pre-release results from requirements F1.4, F1.5, and F1.6. Determine and document how often ongoing evaluation should be conducted to continue supporting this Goal.

5.1.9. F2.7

Identify and document any justifiable factors that account for any remaining differences between the rates at which resources and opportunities are allocated to identified demographic groups.

5.1.10. F2.8

Document the pre-release results for the evaluation described by requirements F2.4, F2.5, and F2.6. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal.

5.1.11. F3.6

Document the pre-release results from requirements F3.4 and F3.5. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal.

5.1.12. RS1.1

Document how:

- 1) reliable and safe behavior is defined for this system and,
- 2) what acceptable error rates are for overall system performance in the context of intended uses.

5.1.13. RS1.3

Determine and document the operational factors, including quality of system input, use, and operational context that are critical to manage for reliable and safe use of the system in its deployed context.

5.1.14. RS1.4

Define and document acceptable ranges for each operational factor important to support reliable and safe system use. Define and document an acceptable error rate for the system when operating within these ranges.

5.1.15. RS1.9

Provide documentation to customers and potential customers of the system that includes the outputs of requirements RS1.2, RS1.7 and RS1.8, and any unsupported uses defined in the Impact Assessment and in RS1.8. When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

5.1.16. RS3.1

Establish and document a detailed inventory of the system health monitoring methods to be used, to include:

- 1) data and insights generated from data repositories, system analytics, and associated alerts,
- 2) processes by which customers can submit information about failures and concerns, and
- 3) processes by which the general public can submit feedback.

5.1.17. RS3.8

Review and update documentation required by Goal T2 when any of the following events occur:

- 1) new uses are added,
- 2) functionality changes,
- 3) new information about reliable and safe performance becomes known as defined by requirement RS3.3, or
- 4) new information about system accuracy and performance becomes available.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

6. CISA

6.1. Principles for the Secure Integration of Artificial Intelligence in Operational Technology

Link: <https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf>

Link: <https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf>

6.1.1. 2.2.6 - Data Quality and Availability

Apply specialized domain knowledge that curates high-quality, comprehensive datasets for effective AI performance. This can be challenging in industrial settings; potential difficulties include the environment containing proprietary or outdated systems and bespoke solutions, and the difficulty of capturing safety-related edge cases. OT operator expertise is necessary for capturing this specialized knowledge. Additionally, operators should work with AI model developers to ensure a data integrity program covers AI systems.

7. SANS

7.1. Critical AI Security Guidelines

Link: <https://sansorg.egnyte.com/dl/bvkYQxrW8QMj>

Link: <https://sansorg.egnyte.com/dl/bvkYQxrW8QMj>

7.1.1. 3.1 Assess Model Hosting Options: Local vs. SaaS Models

When weighing where and how to host AI solutions, be sure to think carefully about and codify legal requirements in any contracts. For example, will your data ever be used or retained by the provider for training or refining a model? If the provider claims that they will not store or use your data, what steps have been taken to prevent your data from being logged when sent to and processed by the API endpoint? How are these logs controlled? How long are they stored? Who has access to them?

8. Cyber Security Council (UAE)

8.1. National Cyber Security Policy for Artificial Intelligence

Link:

https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

Link:

https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

8.1.1. 3.1.4 Change Management and Reporting - 1

The entity should establish and maintain a change management process for AI/ML systems, ensuring any manual or human-initiated modifications, upgrades, or alterations to models, algorithms, training data, or system configurations are properly documented, reviewed, and approved.

9. Central Bank of the UAE

9.1. Guidance Note on the Consumer Protection and Responsible Adoption and Use of Artificial Intelligence and Machine Learning by Licensed Financial Institutions in the U.A.E

Link: <https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence>

Link: <https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence>

9.1.1. 4. Transparency and Explain ability - b

Institutions should maintain documentation relating to the design, training data, model assumptions and functioning of AI to facilitate internal review and internal and external audit.

10. MIC/METI (Japan)

10.1. AI Guidelines for Business

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

10.1.1. Accountability - 1

Establish a situation that allows the origin of data and decisions made during the development, provision, or use of the AI system or service to be traced forward and backward to the extent that is reasonable and technically possible.

10.1.2. Accountability - 6

Document and store information on the items described above and keep them available for a prescribed period whenever and wherever required and able to be referenced in a manner appropriate for their use.

11. ISO/IEC

11.1. TS 42119-2:2025

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:ts:42119:-2:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:ts:42119:-2:ed-1:v1:en>

11.1.1. A.6

Test documentation (summarized from ISO/IEC/IEEE 29119-3)

12. World Economic Forum

12.1. Presidio AI Framework: Towards Safe Generative AI Models

Link: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf

Link: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf

12.1.1. Transparent documentation and use restriction

Transparent documentation is a collection of details (decisions, choices and processes) about the AI model, including the data. It mitigates the risk of lack of transparency, and therefore empowers downstream adapters and users to understand the model's limitations, evaluate its impact and make decisions on model use. This guardrail increases the auditability of the model and helps advance policy initiatives.

13. ISO

13.1. 42001:2023 - Information technology — Artificial intelligence — Management system

Link: <https://www.iso.org/standard/42001>

Link: <https://www.iso.org/standard/42001>

13.1.1. 7.5.2

Creating and updating documented information

13.1.2. 7.5.3

Control of documented information

14. ICO

14.1. Guidance on the AI Auditing Framework - Draft guidance for consultation

Link: <https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>

Link: <https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>

14.1.1. Preventative Controls - 4

Document policy / process for the separation of the AI development environment from the rest of the IT network / infrastructure. Evidence that the separation has been adhered to / happened.

14.1.2. Preventative Controls - 9

Adhere to the policy / process.

14.1.3. Preventative Controls - 10

Have a model governance policy.

15. CEN/CENELEC

15.1. prEN 40000-1-2: Cybersecurity requirements for products with digital elements - Part 1-2: Principles for cyber resilience

Link: <https://genorma.com/en/standards/pren-40000-1-2>

Link: <https://genorma.com/en/standards/pren-40000-1-2>

15.1.1. 5.5

Transparency

16. EU

16.1. EU AI Act

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

16.1.1. 11.1 Technical Documentation

The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to date. The technical documentation shall be drawn up in such a way as to demonstrate that the high-risk AI system complies with the requirements set out in this Section and to provide national competent authorities and notified bodies with the necessary information in a clear and comprehensive form to assess the compliance of the AI system with those requirements. It shall contain, at a minimum, the elements set out in Annex IV. SMEs, including start-ups, may provide the elements of the technical documentation specified in Annex IV in a simplified manner. To that end, the Commission shall establish a simplified technical documentation form targeted at the needs of small and microenterprises. Where an SME, including a start-up, opts to provide the information required in Annex IV in a simplified manner, it shall use the form referred to in this paragraph. Notified bodies shall accept the form for the purposes of the conformity assessment.

16.1.2. 11.2 Technical Documentation

Where a high-risk AI system related to a product covered by the Union harmonisation legislation listed in Section A of Annex I is placed on the market or put into service, a single set of technical documentation shall be drawn up containing all the information set out in paragraph 1, as well as the information required under those legal acts.

16.1.3. 11.3 Technical Documentation

The Commission is empowered to adopt delegated acts in accordance with Article 97 in order to amend Annex IV, where necessary, to ensure that, in light of technical progress, the technical documentation provides all the information necessary to assess the compliance of the system with the requirements set out in this Section.

16.1.4. 18.1 Documentation Keeping

1. The provider shall, for a period ending 10 years after the high-risk AI system has been placed on the market or put into service, keep at the disposal of the national competent authorities:

- (a) the technical documentation referred to in Article 11;
- (b) the documentation concerning the quality management system referred to in Article 17;
- (c) the documentation concerning the changes approved by notified bodies, where applicable;
- (d) the decisions and other documents issued by the notified bodies, where applicable;
- (e) the EU declaration of conformity referred to in Article 47.

16.1.5. 18.2 Documentation Keeping

Each Member State shall determine conditions under which the documentation referred to in paragraph 1 remains at the disposal of the national competent authorities for the period indicated in that paragraph for the cases when a provider or its authorised representative established on its territory goes bankrupt or ceases its activity prior to the end of that period.

16.1.6. 18.3 Documentation Keeping

Providers that are financial institutions subject to requirements regarding their internal governance, arrangements or processes under Union financial services law shall maintain the technical documentation as part of the documentation kept under the relevant Union financial services law.

16.1.7. 53.5 Obligations for providers of general purpose AI models

For the purpose of facilitating compliance with Annex XI, in particular points 2 (d) and (e) thereof, the Commission is empowered to adopt delegated acts in accordance with Article 97 to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation.

16.1.8. 55.10 Obligations of Providers of General-Purpose AI models with Systemic Risk

Any information or documentation obtained pursuant to this Article, including trade secrets, shall be treated in accordance with the confidentiality obligations set out in Article 78.

17. Qatar Central Bank

17.1. Artificial Intelligence Guidelines

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

17.1.1. 8.2.2

Maintain, monitor, document, and review of the AI Models that have been deployed.

17.1.2. 10.1

An Entity must develop and maintain an updated Register of information on all its AI Systems arrangements.

17.1.3. 10.7

The Register of the Entity must include for each system:

- The classification of high-risk or not high-risk.
- The role of an Entity, that is, whether it a User of the system or a Provider of the system.
- A category assigned by an Entity that reflects the nature or functional use of the AI System.
- The Human Oversight protocol used.

In addition, if the AI System is purchased or licensed or outsourced:

- The name of the Provider and any outsourcer.
- Third party supplier assessment.
- The country of registration, local corporate registration number and LEI (where applicable).
- Registered address and relevant contact details.
- Name of the parent company (where applicable).
- Governing law of the AI licensing or purchase arrangement.
- The creation date of the system and all subsequent upgrades.
- The contract start date, as applicable.
- The next contract renewal date.
- For any high-risk system:
 - Assess the AI's substitutability as easy, difficult, or impossible.
 - Identify an alternate service Provider, (where applicable).
- An Entity must maintain a detailed description of all High-Risk AI Systems including Life Cycle (development history, data used, testing, tracking).
- The date of the most recent risk assessment or audit of all High-Risk AI Systems together with a summary of the main results, and the date of the next planned risk assessment/ audit.

17.1.4. 10.8

An Entity must maintain a detailed description of all High-Risk AI Systems including Life Cycle (development history, data used, testing, tracking).

17.1.5. 10.9

The date of the most recent risk assessment or audit of all High-Risk AI Systems together with a summary of the main results, and the date of the next planned risk assessment/ audit.

17.1.6. 12.5

An Entity must obtain approval from its Board of Directors to outsource any function in relation to the use of AI and must document it.

17.1.7. 16.1

Where an Entity is the User or Provider of a High-Risk AI System it must put a framework in place that ensures compliance with this guideline. It must be documented in a systematic and orderly manner in the form of written policies, procedures, and instructions, and must include at least the following aspects:

- A framework for regulatory compliance, including compliance with QCB procedures for how to manage or modify High-Risk AI Systems.
- Techniques, procedures, and systematic actions to be used for the design, design control and design verification of a High-Risk AI System.
- Techniques, procedures, and systematic actions to be used for the development, quality control and quality assurance of the High-Risk AI System.
- Examination, testing and validation procedures to be carried out before, during and after the development of the High-Risk AI System, and the frequency of post-launch reviews.
- Systems and procedures for data management, including data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed before and for the purposes of the placing on the market or putting into service of High-Risk AI Systems.
- The setting-up, implementation and maintenance of a Monitoring System.
- Systems and procedures for record keeping of all relevant documentation and information.
- Resource management, including security of supply-related measures.
- An accountability framework setting out the responsibilities of the management and other staff.

17.1.8. 18.1

An Entity should maintain documentation outlining the design of the AI Model, whether in the role of Provider or User, including but not limited to, where applicable:

- The Input Data source and data description (types and use of data) as per the AI data governance above.
- The data quality checks and data transformations conducted.
- Reasons and justifications for specific model design and development choices.
- Methodology or numerical analyses and calculations conducted.
- Results and expected outcomes.
- Quantitative evaluation and testing metrics used to determine soundness of the model and its results.
- Model usage and implementation.
- Form and frequency of model validation, monitoring and review.
- Assumptions or limitations of the model with justifications.

17.1.9. 19.3.2

Develop and maintaining design documentation.

18. Smart Dubai (UAE)

18.1. AI Ethics Principles & Guidelines

Link: <https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf>

Link: <https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf>

18.1.1. 1.3.1.1

For AI systems which inform significant decisions, especially those with the potential to cause loss, harm or damage, AI developer organisations should consider building in traceability, i.e. the ability to trace the key factors leading to any specific decision.

18.1.2. 1.3.1.2

To facilitate the above, AI developer organisations and AI operator organisations should consider documenting the following information during the design, development and deployment phases, and retaining this documentation for a length of time appropriate to the decision type or industry:

- the provenance of the training data, the methods of collection and treatment, how the data was moved, and measures taken to maintain its accuracy over time;
- the model design and algorithms employed, and;
- changes to the codebase, and authorship of those changes

19. SDAIA (Saudi Arabia)

19.1. AI Ethics Principles

Link: <https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf>

Link: <https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf>

19.1.1. Principle 5 – Reliability & Safety - Plan and Design - 4

The documentation standards are essential to track the evolution of the system, foresee possible risks and fix vulnerabilities.

19.1.2. Principle 6 – Transparency & Explainability - Prepare Input Data - 1

The data sets and the processes that yield the AI system's decision should be documented to the best possible standard to allow for traceability and an increase in transparency.

19.1.3. Principle 7 – Accountability & Responsibility - Prepare Input Data - 3

The documentation of the process is necessary for auditing and risk mitigation. Data must be properly acquired, classified, processed, and accessible to ease human intervention and control at later stages when needed.

20. NCSC/NSA/CISA etc

20.1. AI Data Security

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

20.1.1. 1.1 Source Reliable Data and Track Data Provenance

Verify data sources use trusted, reliable, and accurate data for training and operating AI systems. To the extent possible, only use data from authoritative sources. Implement provenance tracking to enable the tracing of data origins, and log the path that data follows through an AI system. [7] [8] [9] Incorporate a secure provenance database that is cryptographically signed and maintains an immutable, append-only ledger of data changes. This facilitates data provenance tracking, helps identify sources of maliciously modified data, and helps ensure that no single entity can undetectably manipulate the data.

20.1.2. 1.2 Verify and maintain data integrity during storage and transport

Maintaining data integrity is an essential component to preserve the accuracy, reliability, and trustworthiness of AI data. [4] Use checksums and cryptographic hashes to verify that data has not been altered or tampered with during storage or transmission. Generating such unique codes for AI datasets enables the detection of unauthorized changes or corruption, safeguarding the information's authenticity.

20.1.3. 1.3 Employ digital signatures to authenticate trusted data revisions

Digital signatures help ensure data integrity and prevent tampering by third parties. Adopt quantum-resistant digital signature standards [5] [6] to authenticate and verify datasets used during AI model training, fine tuning, alignment, reinforcement learning from human feedback (RLHF), and/or other post-training processes that affect model parameters. Original versions of the data should be cryptographically signed, and any subsequent data revisions should be signed by the person who made the change. Organizations are encouraged to use trusted certificate authorities to verify this process.

20.1.4. 1.5 Classify data and use access controls

Categorize data using a classification system based on sensitivity and required protection measures. [11] This process enables organizations to apply appropriate security controls to different data types. Classifying data enables the enforcement of robust protection measures like stringent encryption and access controls. [33] In general, the output of AI systems should be classified at the same level as the input data (rather than creating a separate set of guardrails).

20.1.5. 2.6 Raw Data Hashes

Data curators should attach a cryptographic hash to all raw data referenced in the dataset. This will enable follow-on data consumers to verify that the data has not changed since it was added to the list.

20.1.6. 2.7 Hash Verification

Data consumers should incorporate a hash check at time of download in order to detect any changes made to it, and the downloader should discard any data that does not pass the hash check.

20.1.7. 3.6 Metadata Management

Implement strong data governance practices to help ensure metadata is well-documented, complete, accurate, and secured.

20.1.8. 3.10 Representative Training Data

Ensure that training data is representative of the totality of the information relevant to any given topic to reduce the risk of statistical bias. Also ensure that AI data is properly divided into training, development, and evaluation sets without overlap to properly measure statistical bias and other measures of performance.

20.1.9. 3.12 Test and Correct for Statistical Bias

Create a repository with instances of observed model output bias. Leverage that information to improve training data audits and with reinforcement learning to “undo” some of the measured bias.

20.1.10. 4.1 Data Management

Employ a data management strategy in keeping with the best practices in this CSI to help ensure that it is easy to add and track new data elements for model training and adaptation. This management strategy enables identification of data elements causing drift for appropriate mitigation or action.

21. Cloud Security Alliance (CSA)

21.1. AI Controls Matrix

Link: <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>

Link: <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>

21.1.1. A&A-01

Establish, document, approve, communicate, apply, evaluate and maintain audit and assurance policies and procedures and standards. Review and update the policies and procedures at least annually or upon significant changes.

21.1.2. DSP-05

Create data flow documentation to identify what data is processed, stored or transmitted where. Review data flow documentation at defined intervals, at least annually, and after any change.

21.1.3. DSP-19

Define and implement, processes, procedures and technical measures to specify and document the physical locations of data, including any locations in which data is processed or backed up.

21.1.4. DSP-20

Define, implement and evaluate processes, procedures and technical measures to: 1) Document and trace data sources, and 2) Make the data source available according to legal and regulatory requirements

21.1.5. GRC-01

Establish, document, approve, communicate, apply, evaluate and maintain policies and procedures for an information governance program, which is sponsored by the leadership of the organization and related to AI systems as well. Review and update the policies and procedures at least annually.

21.1.6. GRC-13

Establish, document, and communicate the degree of explainability needed for the AI Services.

21.1.7. LOG-08

Generate audit records containing relevant security information.

21.1.8. MDS-03

Define, implement, enforce, approve, document, communicate, maintain and evaluate processes and procedures for model documentation. Regularly review and update the model documentation.

21.1.9. MDS-04

Establish and implement baseline requirements for Model documentation.

21.1.10. MDS-05

Define, implement, and evaluate processes, procedures, and technical measures for the validation of the Model documentation aligned with the current model.

21.1.11. MDS-08

Regularly calculate and compare checksums using cryptographic hashes of model checkpoints to detect unauthorized modifications. Apply at least annually based on the level of risk, or after any change of hands.

21.1.12. MDS-09

Sign models cryptographically and verify signatures to ensure model provenance and ownership, any time the model changes hands or is loaded from storage.

21.1.13. UEM-01

Establish, document, approve, communicate, apply, evaluate and maintain policies and procedures for all endpoints. Review and update the policies and procedures at least annually or upon significant system changes.

21.1.14. UEM-02

Define, document, apply and evaluate a list of approved services, applications and sources of applications (stores) acceptable for use by endpoints when accessing or storing organization-managed data.

22. CoSAI

22.1. Establish Risks and Controls for the AI Supply Chain

Link: <https://github.com/cosai-oasis/ws1-supply-chain/blob/main/risks-and-controls-for-the-ai-supply-chain-v1.md>

Link: <https://github.com/cosai-oasis/ws1-supply-chain/blob/main/risks-and-controls-for-the-ai-supply-chain-v1.md>

22.1.1. 3.1.1 Data Poisoning: Threats and Mitigations in AI Supply Chains - Inadequate Data

Using irrelevant data for model fine-tuning:
Comprehensive data provenance with mixture weight documentation

22.1.2. 3.1.1 Data Poisoning: Threats and Mitigations in AI Supply Chains - Invalid Transformations

Data corruption during cleaning processes:
Complete lineage tracking of data transformations

22.1.3. 3.1.4 AI Infrastructure Supply Chain - Opaque Infrastructure Provenance

Insufficient visibility into hosting infrastructure components and their security posture:
Request detailed infrastructure documentation, attestations, and third-party audit results

22.1.4. 3.2.2 - Lack of Model Provenance or Incorrect Provenance

The model's provenance information may be invalid, incomplete, or deliberately tampered with to conceal attacks. In some cases, provenance documentation may be entirely absent, creating significant security gaps:
Generate comprehensive model provenance using trusted build systems and ensure it exists in a tamper-proof format. Implement strict policies preventing the deployment of models lacking proper provenance documentation.

22.2. AI Incident Response Framework

Link: <https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md>

Link: <https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md>

22.2.1. 3.3.2. Detection and Analysis Phase - Investigation Procedures - Evidence Collection

- Interaction logs
- System configurations
- Model versions and parameters
- Network traffic and API calls
- Vector database queries

22.3. Model Context Protocol (MCP) Security

Link: <https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md>

Link: <https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md>

22.3.1. 3.2.6 Cryptographic Verification of Resources

Organizations developing MCP servers must provide cryptographic signatures and software bill of materials (SBOMs) for all server code to verify provenance. Organizations deploying MCP clients and servers should obtain and verify the contents and cryptographic signatures prior to deployment, and have policies restricting the approved sources and signing keys. TLS should be used to protect all data in transit. Remote attestation can further verify that servers are running expected code in a trusted environment. When supported, end-to-end cryptographic signatures can prove the authenticity of resources returned by MCP servers.

23. European Commission

23.1. Ethics guidelines for trustworthy AI

Link: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Link: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

23.1.1. 1.2.3 Accuracy

Accuracy pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions. When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives.

23.1.2. 1.3.2 Quality and integrity of data

The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems. Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.

23.1.3. 1.4.1 Traceability

The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

23.1.4. 2.1.3 Explanation Methods

For a system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation. A whole field of research, Explainable AI (XAI) tries to address this issue to better understand the system's underlying mechanisms and find solutions. Today, this is still an open challenge for AI systems based on neural networks. Training processes with neural nets can result in network parameters set to numerical values that are difficult to correlate with results. Moreover, sometimes small changes in data values might result in dramatic changes in interpretation, leading the system to e.g. confuse a school bus with an ostrich. This vulnerability can also be exploited during attacks on the system. Methods involving XAI research are vital not only to explain the system's behaviour to users, but also to deploy reliable technology.

23.1.5. 2.2.4 Certification

As it cannot be expected that everyone is able to fully understand the workings and effects of AI systems, consideration can be given to organisations that can attest to the broader public that an AI system is transparent, accountable and fair. These certifications would apply standards developed for different application domains and AI techniques, appropriately aligned with the industrial and societal standards of different contexts. Certification can however never replace responsibility. It should hence be complemented by accountability frameworks, including disclaimers as well as review and redress mechanisms.

24. MITRE

24.1. ATLAS Framework

Link: <https://atlas.mitre.org/mitigations>

Link: <https://atlas.mitre.org/mitigations>

24.1.1. AML.M0025 - Maintain AI Dataset Provenance

Maintain a detailed history of datasets used for AI applications. The history should include information about the dataset's source as well as a complete record of any modifications.

25. NIST

25.1. AI RMF 1.0

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

25.1.1. MAP 2.1

The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).

25.1.2. MAP 2.2

Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.

25.1.3. MEASURE 2.9

The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the MAP function – to inform responsible use and governance.

25.1.4. MEASURE 4.2

Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

25.1.5. MEASURE 4.3

Measurable performance improvements or declines based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics are identified and documented.

25.2. SP 800-218A

Link: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>

25.2.1. PO.3.3

Configure tools to generate artifacts of their support of secure software development practices as defined by the organization.

25.2.2. PS.2.1

Make software integrity verification information available to software acquirers.

Generate and provide cryptographic hashes or digital signatures for an AI model and its components, artifacts, and documentation.

Provide digital signatures for AI model changes.

25.2.3. PS.3.1

Securely archive the necessary files and supporting data (e.g., integrity verification information, provenance data) to be retained for each software release.

Perform versioning and tracking for infrastructure tools (e.g., pre-processing, transforms, collection) that support dataset creation and model training.

Include documentation of the justification for AI model selection in the retained information.

Include documentation of the entire training process, such as data preprocessing and model architecture.

25.2.4. PW.1.2

Track and maintain the software's security requirements, risks, and design decisions.

25.2.5. PW.3.2

Track the provenance, when known, of all training, testing, fine-tuning, and aligning data used for an AI model, and document which data do not have known provenance.

25.2.6. PW.9.2

Implement the default settings (or groups of default settings, if applicable), and document each setting for software administrators.

25.3. AI 800-1

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf>

25.3.1. Practice 7.1: Publish transparency reports - 1

Share the methodology and results of pre- and post-deployment evaluations of model capabilities, risks, and mitigations, including as much detail about the data and evaluation methodology as can be disclosed without introducing risks to public safety.

25.3.2. Practice 7.1: Publish transparency reports - 2

Share details regarding the safeguards in place for the model, including how they are applied across different distribution channels, with as much detail as can be shared without rendering the safeguards ineffective.

25.3.3. Practice 7.1: Publish transparency reports - 3

Share information about data used to build the model that is relevant to assessing the misuse risk, such as criteria for data filtering, criteria for data selection, and data sources, when possible.

25.4. IR 8596: Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile

Link: <https://csrc.nist.gov/pubs/ir/8596/iprd>

Link: <https://csrc.nist.gov/pubs/ir/8596/iprd>

25.4.1. PR.PS-01

Configuration management practices are established and applied

25.4.2. RS.AN-06

Actions performed during an investigation are recorded, and the records' integrity and provenance are preserved