ETSI TS 104 223 V1.1.1 (2025-04)

TECHNICAL SPECIFICATION

Securing Artificial Intelligence (SAI);
Baseline Cyber Security Requirements for
AI Models and Systems

Reference

DTS/SAI-0014

Keywords

artificial intelligence, security

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Securing Artificial Intelligence (SAI).

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

Artificial Intelligence (AI) is transforming our daily lives. As the technology continues to evolve and be embedded in people's lives, it is crucial that efforts are taken to protect AI systems from growing cyber security threats. Special focus on the cybersecurity of Artificial Intelligence (AI) is important due to its distinct differences compared to traditional software. These characteristics include security risks such as data poisoning, model obfuscation, indirect prompt injection and operational differences associated with data management.

The present document utilizes existing good practice in the AI and cyber security landscape alongside novel measures to provide a set of targeted high-level principles and provisions at each stage of the AI lifecycle. The objective of the present document is to provide stakeholders in the AI supply chain with clear baseline security requirements to help protect AI systems.

Information on the AI lifecycle and implementation examples are given in ETSI TR 104 128 [i.1].

1 Scope

The present document defines baseline security requirements for AI models and systems. This includes systems that incorporate deep neural networks, such as generative AI. For consistency, the term "AI systems" throughout the present document when framing the scope of provisions and "AI security" which is considered a subset of cyber security. The present document is not designed for academics who are creating and testing AI systems only for research purposes (AI systems which are not going to be deployed).

The present document separates principles and requirements into five phases. These are secure design, secure development, secure deployment, secure maintenance and secure end of life. Relevant standards and publications are signposted at the start of each principle to highlight links between the various documents and the present document. This is not an exhaustive list.

NOTE: The principles can also be mapped to the life cycle stages described in ISO/IEC 22989 [i.3]. The secure design and development principles can be applied during the Design and development life cycle stage. Similarly, the secure deployment principles can be applied during the Deployment stage, secure maintenance to the Operations and monitoring stage and secure end of life during the Retirement stage.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found in the ETSI docbox.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

[1] ISO/IEC 27001:2022: "Information security, cybersecurity and privacy protection — Information security management systems — Requirements".

[2] CISA: "Software Bill of Materials (SBOM)".

[3] NIST: "AI Risk Management Framework: Second Draft", 2022.

[4] NIST AI 100-1: "AI Risk Management Framework 1.0", 2023.

[5] Australian Signals Directorate: "An introduction to Artificial Intelligence", 2023.

[6] World Economic Forum, IBM: "Presidio AI Framework: Towards Safe Generative AI Models", 2024.

[7] OWASP: "OWASP AI Exchange".

[8] MITRE ATLAS™: "Mitigations".

[9] Google®: "Secure AI Approach Framework: A quick guide to implementing the Secure AI Framework (SAIF)", 2023.

[10] ELSA: "ELSA - European Lighthouse on Secure and Safe AI", 2023.

[11] Cisco: "The Cisco Responsible AI Framework", 2024.

[12] Amazon: "AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI", Amazon White Paper, 2024.

[13] NIST AI 100-2 E2023: "Adversarial Machine Learning Taxonomy: A Taxonomy and Terminology of Attacks and Mitigations".

[14] ENISA: "Multilayer Framework for Good Cybersecurity Practices for AI", 2023.

[15] NCSC: "Guidelines for secure AI system development", 2023.

[16] Federal Office in Information Security: "AI Security Concerns in a Nutshell", 2023.

[17] G7 Hiroshima Summit: "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems", 2023.

[18] United States Department of Health and Human Services: "Trustworthy AI (TAI) Playbook: Executive Summary", 2021.

[19] OpenAI: "Preparedness Framework (Beta)", 2023.

[20] Information Commissioner's Office (ICO): "Guidance on the AI Auditing Framework", 2020.

[21] Nvidia: "NeMo-Guardrails", 2023.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI TR 104 128: "Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems".

[i.2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[i.3] ISO/IEC 22989:2022: "Information technology — Artificial intelligence — Artificial intelligence concepts and terminology".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

adversarial attack: attempt to manipulate an AI model by introducing specially crafted inputs to cause the model to produce errors or unintended outcomes

AI system: engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives

Application Programming Interface (API): set of tools and protocols that allow different software systems to communicate and interact

data poisoning: type of adversarial attack where malicious data is introduced into training datasets to compromise the AI system's performance or behaviour

guardrails: predefined constraints or rules implemented to control and limit an AI system's outputs and behaviours, ensuring safety, reliability, and alignment with ethical or operational guidelines

inference: reasoning by which conclusions are derived from known premises

NOTE 1: In AI, a premise is either a fact, a rule, a model, a feature or raw data.

NOTE 2: The term "inference" refers both to the process and its result.

input data: data for which an AI system calculates a predicted output or inference

machine learning algorithm: algorithm to determine parameters of a machine learning model from data according to given criteria

machine learning model: mathematical construct that generates an inference or prediction based on input data or information

model inversion: privacy attack where an adversary infers sensitive information about the training data by analysing the AI model's outputs

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

prediction: primary output of an AI system when provided with input data or information

NOTE 1: Predictions can be followed by additional outputs, such as recommendations, decisions and actions.

NOTE 2: Prediction does not necessarily refer to predicting something in the future.

NOTE 3: Predictions can refer to various kinds of data analysis or production applied to new data or historical data (including translating text, creating synthetic images or diagnosing a previous power failure).

prompt: input provided to an AI model, often in the form of text, that directs or guides its response.

NOTE: Prompts can include questions, instructions, or context for the desired output.

risk assessment: process of identifying, analysing and mitigating potential threats to the security or functionality of an AI system

sanitisation: process of cleaning and validating data or inputs to remove errors, inconsistencies and malicious content, ensuring data integrity and security

system prompt: predefined input or set of instructions provided to guide the behaviour of an AI model, often used to define its tone, rules, or operational context

threat modelling: process to identify and address potential security threats to a system during its design and development phases

training: process of teaching an AI model to recognize patterns, make decisions, or generate outputs by exposing it to labelled data and adjusting its parameters to minimize errors

training data: data used to train a machine learning model

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AI	Artificial Intelligence
API	Application Programming Interface

4 Audience

This clause defines the stakeholder groups that form the AI supply chain. An indication is given for each principle on which stakeholders are primarily responsible for its implementation. Importantly, a single entity can hold multiple stakeholder roles in the present document as well as responsibilities from different regulatory regimes.

NOTE: Examples include under data protection law, when processing personal data organizations can have a role of controller and/or joint controller and/or processor, depending on their role in creating and setting up AI systems.

All stakeholders included in Table 4-1 should note that they can have data protection obligations. Additionally, senior leaders in an organization also have responsibilities to help protect their staff and infrastructure. Some provisions for Developers in the present document are less applicable to AI systems involving open-source models. Further information and guidance about different types of AI systems can be found in ETSI TR 104 128 [i.1].

Table 4-1: Stakeholder definition

Stakeholder	Definitions
Developers	This encompasses any type of business or organization across any sector as well as individuals that are responsible for creating or adapting an AI model and/or system. This applies to all AI technologies, including proprietary and open-source models. For context, a business or organization that creates an AI model and who is also responsible for embedding/deploying that model/system in their organization are defined in the present document to be both a Developer and a System Operator. Developers can be AI providers under the EU AI Act [i.2], when putting a system into service or placing it on the market.
System Operators	This includes any type of business or organization across any sector that has responsibility for embedding/deploying an AI model and system within their infrastructure and/or its ongoing maintenance. This applies to all AI technologies, including proprietary and open-source models. This term also includes those businesses that provide a contractual service to organizations to embed/deploy an AI model and system for business purposes. System operators can be deployers under the EU AI Act [i.2] and can also be AI providers if they make changes to the system.
Data Custodians	This includes any type of business, organization or individual that controls data permissions and the integrity of data that is used for any AI model or system to function. This stakeholder group also includes those entities that set the policies for how data is used and managed for an AI model and/or system. In the context of an AI system, there could be multiple data custodians involved because some data used to create a model could come from the organization that is deploying/embedding the system in their infrastructure and other data could be from public databases and other sources.
End-users	This encompasses any employee within an organization or business and consumers who use an AI model and system for any purpose, including to support their work and day-to-day activities. This applies to all AI technologies and both proprietary and open-source models. This stakeholder group has been created because the voluntary Code has placed expectations on Developers, System Operators and Data custodians to help inform and protect end-users.

Stakeholder	Definitions
Affected entities	Encompasses all individuals and technologies, such as apps and autonomous systems, that are not directly affected by AI systems or decisions based on the output of AI systems. These individuals do not necessarily interact with the deployed system or application.

5 AI Security Principles and Provisions

5.1 Secure Design

5.1.1 Principle 1: Raise awareness of AI security threats and risks

Primarily applies to: System Operators, Developers, and Data Custodians

References: [3], [4], [5], [6], [7], [8], [9], [10], [11].

Provision 5.1.1-1 Organizations' cyber security training programme shall include AI security content which shall be regularly reviewed and updated, such as if new substantial AI-related security threats emerge.

Provision 5.1.1-1.1 AI security training shall be tailored to the specific roles and responsibilities of staff members.

Provision 5.1.1-2 As part of an Organization's wider staff training programme, they shall require all staff to maintain awareness of the latest security threats and vulnerabilities that are AI-related. Where available, this awareness shall include proposed mitigations.

Provision 5.1.1-2.1 These updates should be communicated through multiple channels, such as security bulletins, newsletters, or internal knowledge-sharing platforms. This will ensure broad dissemination and understanding among the staff.

Provision 5.1.1-2.2 Organizations shall provide developers with training in secure coding and system design techniques specific to AI development, with a focus on preventing and mitigating security vulnerabilities in AI algorithms, models, and associated software.

5.1.2 Principle 2: Design the AI system for security as well as functionality and performance

Primarily applies to: System Operators and Developers

References: [7], [8], [6], [14], [15], [16], [11], [17], [18], [19], [5], [20].

Provision 5.1.2-1 As part of deciding whether to create an AI system, a System Operator and/or Developer shall conduct a thorough assessment that includes determining and documenting the business requirements and/or problem they are seeking to address, along with associated AI security risks and mitigation strategies.

Provision 5.1.2-1.1 Where the Data Custodian is part of a Developer's organization, they shall be included in internal discussions when determining the requirements and data needs of an AI system.

Provision 5.1.2-2 Developers and System Operators shall ensure that AI systems are designed and implemented to withstand adversarial AI attacks, unexpected inputs and AI system failure.

Provision 5.1.2-3 To support the process of preparing data, security auditing and incident response for an AI system, Developers shall document and create an audit trail in relation to the AI system. This shall include the operation, and lifecycle management of models, datasets and prompts incorporated into the system.

Provision 5.1.2-4 If a Developer or System Operator uses an external component, they shall conduct an AI security risk assessment and due diligence process in line with their existing software development processes, that assesses AI specific risks.

Provision 5.1.2-5 Data Custodians shall ensure that the intended usage of the system is appropriate to the sensitivity of the data it was trained on as well as the controls intended to ensure the security of the data.

Provision 5.1.2-5.1 Organizations should ensure that employees are encouraged to proactively report and identify any potential security risks in AI systems and ensure appropriate safeguards are in place.

Provision 5.1.2-6 Where the AI system will be interacting with other systems or data sources, (be they internal or external), Developers and System Operators shall ensure that the permissions granted to the AI system on other systems are only provided as required for functionality and are risk assessed.

Provision 5.1.2-7 If a Developer or System Operator chooses to work with an external provider, they shall undertake a due diligence assessment and should ensure that the provider is adhering to the present document.

5.1.3 Principle 3: Evaluate the threats and manage the risks to the AI system

Primarily applies to: Developers and System Operators

References: [1], [7], [6], [21], [14], [9], [17], [15], [8], [13]

Provision 5.1.3-1 Developers and System Operators shall analyse threats and manage security risks to their systems. Threat modelling should include regular reviews and updates and address AI-specific attacks, such as data poisoning, model inversion, and membership inference.

Provision 5.1.3-1.1 The threat modelling and risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented or updated at any stage of the AI lifecycle.

Provision 5.1.3-1.2 Developers shall manage the security risks associated with AI models that provide superfluous functionalities, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function.

Provision 5.1.3-1.3 System Operators shall apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance.

Provision 5.1.3-2 Where AI security threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators so they can threat model their systems. System Operators shall communicate this information to End-users, so they are made aware of these threats. This communication should include detailed descriptions of the risks, potential impacts, and recommended actions to address or monitor these threats.

Provision 5.1.3-3 Where an external entity has responsibility for AI security risks identified within an organizations infrastructure, System Operators should attain assurance that these parties are able to address such risks.

Provision 5.1.3-4 Developers and System Operators should continuously monitor and review their system infrastructure according to risk appetite. It is important to recognize that a higher level of risk will remain in AI systems despite the application of controls to mitigate against them.

5.1.4 Principle 4: Enable human responsibility for AI systems

Primarily applies to: Developers and System Operators

References: [7], [8], [16]

Provision 5.1.4-1 When designing an AI system, Developers and/or System Operators should incorporate and maintain capabilities to enable human oversight.

Provision 5.1.4-2 Developers should design systems to make it easy for humans to assess outputs that they are responsible for in said system (such as by ensuring that models outputs are explainable or interpretable).

Provision 5.1.4-3 Where human oversight is a risk control, Developers and/or System Operators shall design, develop, verify and maintain technical measures to reduce the risk through such oversight.

Provision 5.1.4-4 Developers should verify that the security controls specified by the Data Custodian have been built into the system.

Provision 5.1.4-5 Developers and System Operators should make End-users aware of prohibited use cases of the AI system.

5.2 Secure Development

5.2.1 Principle 5: Identify, track and protect the assets

Primarily applies to: Developers, System Operators and Data Custodians

References: [7], [21], [15], [16], [11], [12], [17], [20]

Provision 5.2.1-1 Developers, Data Custodians and System Operators shall maintain a comprehensive inventory of their assets (including their interdependencies/connectivity).

Provision 5.2.1-2 As part of broader software security practices, Developers, Data Custodians and System Operators shall have processes and tools to track, authenticate, manage version control and secure their assets due to the increased complexities of AI specific assets.

Provision 5.2.1-3 System Operators shall develop and tailor their disaster recovery plans to account for specific attacks aimed at AI systems.

Provision 5.2.1-3.1 System Operators should ensure that a known good state can be restored.

Provision 5.2.1-4 Developers, System Operators, Data Custodians and End-users shall protect sensitive data, such as training or test data, against unauthorized access (see clause 5.2.3 for details on securing data).

Provision 5.2.1-4.1 Developers, Data Custodians and System Operators shall apply checks and sanitisation to data and inputs when designing the model based on their access to said data and inputs and where those data and inputs are stored. This shall be repeated when model revisions are made in response to user feedback or continuous learning. See clause 5.2.2 for relevant provisions for open source.

Provision 5.2.1-4.2 Where training data or model weights could be confidential, Developers shall put proportionate protections in place.

5.2.2 Principle 6: Secure the infrastructure

Primarily applies to: Developers and System Operators

References: [7], [8], [6], [15], [20]

Provision 5.2.2-1 Developers and System Operators shall evaluate their organization's access control frameworks and identify appropriate measures to secure APIs, models, data, and training and processing pipelines.

Provision 5.2.2-2 If a Developer offers an API to external customers or collaborators, they shall apply controls that mitigate attacks on the AI system via the API. For example, placing limits on model access rate to limit an attacker's ability to reverse engineer or overwhelm defences to rapidly poison a model.

Provision 5.2.2-3 Developers shall also create dedicated environments for development and model tuning activities. The dedicated environments shall be backed by technical controls to ensure separation and principle of least privilege. In the context of AI, this is particularly necessary because training data shall only be present in the training and development environments where this training data is not based on publicly available data.

Provision 5.2.2-4 Developers and System Operators shall implement and publish a clear and accessible vulnerability disclosure policy.

Provision 5.2.2-5 Developers and System Operators shall create, test and maintain an AI system incident management plan and an AI system recovery plan.

Provision 5.2.2-6 Developers and System Operators should ensure that, where they are using cloud service operators to help to deliver the capability, their contractual agreements support compliance with the above requirements.

5.2.3 Principle 7: Secure the supply chain

Primarily applies to: Developers, System Operators and Data Custodians

References: [2], [7], [15], [5]

Provision 5.2.3-1 Developers and System Operators shall follow secure software supply chain processes for their AI model and system development.

Provision 5.2.3-2 System Operators that choose to use or adapt any models, or components, which are not well-documented or secured shall be able to justify their decision to use such models or components through documentation (for example if there was no other supplier for said component).

Provision 5.2.3-2.1 In this case, Developers and System Operators shall have mitigating controls and undertake a risk assessment linked to such models or components.

Provision 5.2.3-2.2 System Operators shall share this documentation with End-users in an accessible way.

Provision 5.2.3-3 Developers and System Operators shall re-run evaluations on released models that they intend on using.

Provision 5.2.3-4 System Operators shall communicate their intention to update models to End-users in an accessible way prior to models being updated.

5.2.4 Principle 8: Document data, models and prompts

Primarily applies to: Developers

References: [7], [6], [15], [11], [20]

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.

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.

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.

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.

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.

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.

5.2.5 Principle 9: Conduct appropriate testing and evaluation

Primarily applies to: Developers and System Operators

References: [7], [6], [21], [15], [14], [9], [17]

Provision 5.2.5-1 Developers shall ensure that all models, applications and systems that are released to System Operators and/or End-users have been tested as part of a security assessment process.

Provision 5.2.5-2 System Operators shall conduct testing prior to the system being deployed with support from Developers.

Provision 5.2.5-2.1 For security testing, System Operators and Developers should use independent security testers with technical skills relevant to their AI systems.

Provision 5.2.5-3 Developers should ensure that the findings from the testing and evaluation are shared with System Operators, to inform their own testing and evaluation.

Provision 5.2.5-4 Developers should evaluate model outputs to ensure they do not allow System Operators or End-users to reverse engineer non-public aspects of the model or the training data.

Provision 5.2.5-4.1 Additionally, Developers should evaluate model outputs to ensure they do not provide System Operators or End-users with unintended influence over the system.

5.3 Secure Deployment

5.3.1 Principle 10: Communication and processes associated with End-users and Affected Entities

NOTE: As part of an organization's wider deployment practices, pre-deployment testing of AI systems is to be considered alongside the requirements below.

Provision 5.3.1-1 System Operators shall convey to End-users in an accessible way where and how their data will be used, accessed and stored (for example, if it is used for model retraining, or reviewed by employees or partners). If the Developer is an external entity, they shall provide this information to System Operators.

Provision 5.3.1-2 System Operators shall provide End-users with accessible guidance to support their use, management, integration, and configuration of AI systems. If the Developer is an external entity, they shall provide all necessary information to help System Operators.

Provision 5.3.1-2.1 System Operators shall include guidance on the appropriate use of the model or system, which includes highlighting limitations and potential failure modes.

Provision 5.3.1-2.2 System Operators shall proactively inform End-users of any security relevant updates and provide clear explanations in an accessible way.

Provision 5.3.1-3 Developers and System Operators should support End-users and Affected Entities during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with End-users.

5.4 Secure Maintenance

5.4.1 Principle 11: Maintain regular security updates, patches and mitigations

Primarily applies to: Developers and System Operators

References: [20]

Provision 5.4.1-1 Developers shall provide security updates and patches, where possible, and notify System Operators of the security updates. System Operators shall deliver these updates and patches to End-users.

Provision 5.4.1-1.1 Developers shall have mechanisms and contingency plans to mitigate security risks, particularly in instances where updates cannot be provided for AI systems.

Provision 5.4.1-2 Developers should treat major AI system updates as though a new version of a model has been developed and therefore undertake a new security testing and evaluation process to help protect users.

Provision 5.4.1-3 Developers should support System Operators to evaluate and respond to model changes, (for example by providing preview access via beta-testing and versioned APIs).

5.4.2 Principle 12: Monitor the system's behaviour

Primarily applies to: Developers and System Operators

References: [7], [6], [21], [14], [16], [11], [12], [17], [20]

Provision 5.4.2-1 System Operators shall log system and user actions to support security compliance, incident investigations, and vulnerability remediation.

Provision 5.4.2-2 System Operators should analyse their logs to ensure that AI models continue to produce desired outputs and to detect anomalies, security breaches, or unexpected behaviour over time (such as due to data drift or data poisoning).

Provision 5.4.2-3 System Operators and Developers should monitor internal states of their AI systems where this could better enable them to address security threats, or to enable future security analytics.

Provision 5.4.2-4 System Operators and Developers should monitor the performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security.

5.5 Secure End of Life

5.5.1 Principle 13: Ensure proper data and model disposal

Primarily applies to: Developers and System Operators

Provision 5.5.1-1 If a Developer or System Operator decides to transfer or share ownership of training data and/or a model to another entity they shall involve Data Custodians and securely dispose of these assets. This will protect AI against security issues that can transfer from one AI system instantiation to another.

Provision 5.5.1-2 If a Developer or System Operators decides to decommission a model and/or system, they shall involve Data Custodians and securely delete applicable data and configuration details.

History

Document history		
V1.1.1	April 2025	Publication