International Telecommunication Union

ITU-T Technical Report

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(05/2022)

XSTR-SEC-AI

Guidelines for security management of using artificial intelligence technology

Technical Report ITU-T XSTR-SEC-AI

Guidelines for security management of using artificial intelligence technology

Summary

As a new generation of information and communication technology (ICT) infrastructure, artificial intelligence (AI) has been widely used in various fields of social economy. In the development and application of AI technology, some security threats might arise, which may run through the whole life cycle of AI products, applications and services from design, through development and on to retirement. Organizations need to identify the source of security threats according to the life cycle of AI technology so as to deploy targeted security strategies.

This Technical Report focuses on security threats faced by current use of AI technology, puts forward AI security management suggestions, and provides a useful reference for organizations to improve the security protection ability in the use of AI technology.

Keywords

AI, artificial intelligence, security threats, using AI technology.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Table of Contents

Technical Report ITU-T XSTR-SEC-AI

Guidelines for security management of using artificial intelligence technology

1 Scope

This Technical Report provides guidelines for organizations within the information and communication technology (ICT) industry for security management of using artificial intelligence (AI) technology.

The scope of the work includes to:

– describe the life cycle of using AI technology in products, applications and services;

– analyse security threats in the life cycle of using AI technology;

– provide guidelines to organizations for security management of using AI technology.

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 application [b-ITU-T Y.2091]: A structured set of capabilities, which provide value-added functionality supported by one or more services, which may be supported by an API interface.

3.1.2 artificial intelligence [b-ISO/IEC 22989]: Set of methods or automated entities that together build, optimize and apply a model (see clause 3.1.15) so that the system can, for a given set of predefined tasks, compute predictions (see clause 3.1.18), recommendations, or decisions.

3.1.3 AI system [b-ISO/IEC 22989]: Engineered system featuring artificial intelligence (see clause 3.1.2)

3.1.4 bias [b-ISO/IEC 22989]: Systematic difference in treatment of certain objects, people, or groups in comparison to others.

3.1.5 capability [b-ITU-R M.1224-1]: The ability of an item to meet a service demand of given quantitative characteristics under given internal conditions.

3.1.6 continuous learning [b-ISO/IEC 22989]: Incremental training of an AI system (see clause 3.1.3) that takes place on an ongoing basis during the operation phase of the AI system life cycle.

3.1.7 control [b-ISO/IEC 22989]: Purposeful action on or in a process to meet specified objectives

3.1.8 data security [b-ISO/IEC 29182-2]: Preservation of data to guarantee availability, confidentiality and data integrity.

3.1.9 explainability [b-ISO/IEC 22989]: Property of an AI system (see clause 3.1.3) to express important factors influencing the AI system (see clause 3.1.3) results in a way that humans can understand.

3.1.10 inference [b-ISO/IEC 22989]: Reasoning by which conclusions are derived from known premises.

3.1.11 label [b-ISO/IEC 22989]: The target variable assigned to a sample.

3.1.12 machine learning [b-ITU-T Y.3172]: Process that enable computational system to understand data and gain knowledge from it without necessarily being explicitly programmed.

3.1.13 machine learning algorithm [b-ISO/IEC 22989]: Algorithm to establish parameters (see clause 3.1.16), according to a given criteria, of a machine learning model (see clause 3.1.14) from data.

3.1.14 machine learning model [b-ITU-T Y.3172]: Model created by applying machine learning techniques to data to learn from.

3.1.15 model [b-ISO/IEC 22989]: Physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, process or data.

3.1.16 parameter [b-ISO/IEC 22989]: Internal variable of a model (see clause 3.1.15) that affects how it computes its outputs.

3.1.17 performance [b-ISO/IEC 22989]: Measurable result.

3.1.18 prediction [b-ISO/IEC 22989]: Output of a machine learning model (see clause 3.1.14) when provided with input data.

3.1.19 reliability [b-ISO/IEC 22989]: Property of consistent intended behaviour and results.

3.1.20 robustness [b-ISO/IEC 22989]: Ability of a system to maintain its level of performance under any circumstances.

3.1.21 sample [b-ISO/IEC 22989]: Atomic data element processed in quantities by a machine learning algorithm (see clause 3.1.13).

3.1.22 service [b-ITU-T Y.2091]: A set of functions and facilities offered to a user by a provider.

3.1.23 test data [b-ISO/IEC 22989]: Data used to assess the performance of a final machine learning model (see clause 3.1.14).

3.1.24 threat [b-ISO/IEC 27000]: Potential cause of an unwanted incident, which can result in harm to a system or organization.

3.1.25 training data [b-ISO/IEC 22989]: Subset of input data samples used to train a machine learning model (see clause 3.1.14).

3.1.26 validation [b-ISO/IEC 22989]: Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.

3.1.27 verification [b-ISO/IEC 22989]: Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled.

3.1.28 vulnerability [b-NIST-SP-800-30]: A weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

AI Artificial Intelligence

CPU Central Processing Unit

ICT Information and Communications Technology

5 Life cycle of using AI technology in products, applications and services

Life cycle of using AI technology in products, applications and services can be divided into four phases:

– **Design and manufacturing**: It refers to the design, implementation and testing process of AI products, applications and services.

– **Deployment and application**: It refers to the installation in target environments and actual operational process of AI products, applications and services.

– **Maintenance and upgrading**: It refers to the process of modifying AI products, applications and services after they are delivered for use in order to correct the errors in their operation, adapt to the changes in the operating environment, meet the new needs of users, and lay the foundation for future improvement.

– **Abandonment and destruction**: It refers to the process of abandoning and destroying the data, algorithm model or the whole AI system.

6 Security threats

The analysis of security threats through the process across all life cycle phases of using AI technology could help organizations to figure out the sources of security threats, so as to research and deploy targeted security defence strategies. Table 1 summarizes the security threats in the life cycle of using AI technology from two dimensions: threat sources and threat performance.

Table 1 – Security threats in the process of using AI technology

Threat source	Threat performance
Design and manufacturing	
Infrastructure imperfection	– Algorithm backdoor embedding – Code security vulnerability – Unbalanced training data – Training data poisoning – Training data leakage
AI technology vulnerability	– Algorithm weak robustness – Algorithm unexplainability – Algorithm bias discrimination
Design and research error	– Uncontrollable behaviour
Test validation inadequacy	– Untimely problem detect and fix
Deployment and application	
Untrusted deployment environment	– Unauthorized access – Unauthorized use
Security attack	– Adversarial examples attack – Algorithm backdoor attack – Model stealing attack – Model feedback misdirection – Reverse data recovery – Member reasoning attack – Attribute inference attack – Code vulnerability exploitation
Insecure usage	– Abuse and malicious use

Table 1 – Security threats in the process of using AI technology

Threat source	Threat performance
Maintenance and upgrading	
Test validation data untimely updated	– Failure to detect and fix security problems brought by continuous learning
Incorrect objective	– Non-compliance with national laws, regulations and ethics
Abandonment and destruction	
Incomplete destruction	– Personal privacy leakage

7 Security measures

7.1 Security objectives

Based on the security threat analysis and considering the actual use of AI technology, Table 2 gives the security objectives of organizations using AI technology from six aspects.

Table 2 – Security objectives of using AI technology

Security objectives	Description
Legal compliance	Ensure that the use of AI technology complies with national laws, regulations and social ethics.
Reliable and controllable functions	Ensure that all functions of the AI system produce expected behaviour and results within the specified operating conditions and time cycle, and are under the control of human operators all the time.
Data security and reliability	Ensure that collected, used and stored data is prevented from stealing, privacy-leaking and unauthorized modifying.
Fairness and justice in decision-making	Ensure that AI applications take into account the characteristic information of various groups and will not make discriminatory and biased decisions.
Behaviour explainability	Ensure that AI applications provide reasonable explanations of their behaviours and results in a way that humans can understand.
Incident traceability	Ensure that the traceability system is improved and deploy technical measures to trace the cause, behaviour and subject of security incidents.

7.2 Security management

7.2.1 Organization construction

The organization could set up relevant departments to manage and execute the security of using AI technology, and establish an effective work assessment mechanism. Meanwhile, the organization could establish virtual collaborative mechanisms with designated security experts in AI related departments, responsible for formulating unified security management policies for the use of AI technology, and implement relevant processes in the department. In addition, the organization might set up security supervision agencies, which are responsible for regular supervision and inspection of the implementation of security systems and the effectiveness of technical tools.

7.2.2 System process

The organization could consider and design a series of AI security systems according to their own use of AI technology, including clarifying the objectives, vision, strategies, basic principles of AI security, security management methods corresponding to each life cycle phase of using AI technology, and security operation process specifications.

7.2.3 Personnel capacity

The organization could expand the team of AI security professionals and strengthen the professional skills of employees. AI security personnel ability mainly includes AI security management ability, AI security operation ability, AI security technology ability, etc.

7.3 Security technologies

According to different targets to which security threats are directed, security technologies are analysed from the following four perspective of key protection objects of AI security: service security, algorithm security, data security and platform security.

1) Service security technology: It refers to security technology related to AI service.

2) Algorithm security technology: It refers to security technology related to AI algorithm.

3) Data security technology: It refers to security technology of AI training data deployment.

4) Platform security technology: It refers to security technology of machine learning framework platform deployment.

7.3.1 Service security technology

1) Service compliance assessment

Self-assessment: It refers to the compliance assessment of using AI technology in the organization. The organization assesses if the use of AI technology complies with national laws, regulations and social ethics.

Third-party assessment: It refers to the compliance assessment of using AI technology by the third-party assessment agency. The third-party can independently, or on the basis of the suspected items found in the self-assessment, carry out assessment.

2) Service security mechanisms

Service access control: It can avoid illegal or malicious frequent access through measures including authentication, restriction to access times and frequency, which could effectively prevent attackers from using a large number of access results to estimate the internal information to implement conducts of attacking, such as model theft and member reasoning.

Service security isolation: By decoupling the logical relationship between the service function modules of using AI technology, the service function modules are isolated to guarantee the normal operation of other services when some services are abnormal.

Service security redundancy: Multiple AI algorithm models that complete the same function are deployed in the key service links of using AI technology, so that the final service decision would not be affected when a single algorithm model is wrong, and the reliability of the whole application would be improved.

Service security fusing: Through the pre-set security policy, the use of AI technology can adjust the decision mechanism according to the different confidence levels of the output results in the dynamic environment.

Service security monitoring: Deploy a real-time monitoring system to monitor the operation and security status of using AI technology, analyse and judge the current security threat and give timely warnings of abnormal operation.

3) Security attack detection

Adversarial examples detection: Through training the classifier to detect whether the input data is an adversarial example with potential for malicious disturbance, and provide timely warnings against sample attacks.

Algorithm backdoor detection: Since an algorithm model subjected to a backdoor attack will trigger malicious behaviour only when faced with the input data of an embedded backdoor trigger, it is very challenging to detect whether the algorithm model of using AI technology is subjected to backdoor attack.

7.3.2 Algorithm security technology

1) Algorithm robustness enhancement

Data enhancement: By simulating various situations that may occur in natural scenes or adversarial scenes, the supporting algorithm model learns relevant features from the data to improve the robustness of the algorithm, so as to always maintain the normal performance level in various scenarios. The data enhancement method could be used to improve the natural robustness of the algorithm.

Robust feature learning: Enhance the robustness level of the algorithm model by making the model learn features that are not easy to disturb in the natural scene or by reducing the dependence on easily disturbed features.

Model randomization: It refers to an attack optimization method by introducing randomness in the model operation processes so that the attacker cannot obtain accurate information, and so as to ensure that the algorithm model can still maintain a normal performance level under active attack.

Model regularization: The loss function is made smoother by adding model constraints, so as to reduce the possibility of the attacker finding the algorithm vulnerability, thus ensuring that the algorithm model can still maintain the normal performance level under active attack.

Training data sampling: By extracting from the original data subset of a training set, restricting the greatest amount of data could be contributed to by each user, to ensure that the special sample model does not occupy a large proportion of the training data, thus effectively avoiding a sharp decline of model recognition performance after intentional and unintentional factors cause an imbalance of training data.

2) Algorithm fairness guarantee

Algorithm fairness constraint: By expressing the algorithm fairness as a model constraint and adding it to the model optimization process, the trained model can meet the fairness requirement, so that the model can make a fair decision for any input data.

Bias discrimination post-processing: It refers to modifying the prediction results of the pre-training model for any input to meet the requirements of fairness.

3) Algorithm interpretability improvement

Model self-explanation: It refers to the ability to understand the decision-making process and basis of the model without additional information by directly using the model with its own interpretability.

Algorithm global interpretation: It refers to the overall interpretation of the decision logic and internal working mechanisms behind the model in a manner that can be understood by humans.

Algorithm partial interpretation: It refers to help in understanding the decision-making process and decision-making basis of the model for each specific input sample by analysing the influence degree of each one-dimensional feature of the input sample on the final decision-making result of the model.

4) Algorithm intellectual property protection

Model watermark: The watermark is embedded in the model file during training to avoid the loss of intellectual property due to the theft of the model.

5) Algorithm security evaluation

Robustness evaluation: Currently the robustness evaluation of the use of AI technology mainly includes the two aspects. The first is natural robustness evaluation. This kind of method simulates abnormal mutations in normal scenarios by means of input data transformation, collection and rerelease of abnormal data with small probability, etc., to evaluate the natural robustness of the use of AI technology in normal environments. The second is the evaluation of countermeasure robustness. This kind of method evaluates the robustness of the use of AI technology in the case of a malicious attack by simulating the disturbance of a malicious attacker to the input data by using a countermeasure sample attack and other methods.

Fairness evaluation: Currently there are mainly two kinds of fairness evaluation methods of AI. The first is fairness evaluation based on static data sets. This kind of method evaluates the fairness of the use of AI technology by analysing and comparing the performance differences on data sets with different sensitive attribute values in a static test data set. The second is fairness evaluation based on dynamic simulation. This kind of method simulates the interaction between the use of AI technology and the operating environment through reinforcement learning and other techniques, and is used to evaluate the long-term fairness of the use of AI technology that are seriously affected by the feedback of the operating environment.

Interpretability evaluation: Due to the difference in the definition of interpretability under different application scenarios, the current interpretability evaluation still mainly relies on qualitative methods such as subjective observation to judge whether the decision principle and reasons are transparent and understandable.

7.3.3 Data security technology

1) Data privacy computing

Multi-party security computing: It refers to the cryptography technology based on multi-party data cooperation to complete the calculation goal, so as not to leak the private data of the parties except for the calculation results and the information that can be deduced.

Homomorphic encryption: It refers to a cryptography technology based on the computational complexity theory of mathematical problems, which can guarantee that the results of the calculation of the homomorphic encrypted data are consistent with the output results of the same calculation and encryption processing of the unencrypted original data.

Zero-knowledge certification: It refers to the ability of the prover to convince the verifier that an assertion is true without providing any useful information to the verifier.

Differential privacy: By adding interference noise to the data, the attacker can avoid analysing the data set and reverse cracking the corresponding relationship between the data and the individuals, so as to protect the privacy information in the data.

Trusted execution environment: It refers to an area of the central processing unit (CPU) on the server and mobile side that provides a secure space for data and code execution. Trusted applications running in the trusted execution environment have access to the full functionality of the device's main processor and memory, while hardware isolation protects these components from other applications running in the main operating system.

Federated learning: It refers to the process of establishing a shared machine learning model through data joint training by parameter exchange under an encryption mechanism under the condition that the data of each participant is not out of the local area. There are several modes for federated learning to choose from for different application scenarios. Horizontal federated learning can be used when user feature dimensions overlap more and users overlap less. Vertical federated learning can be used when users overlap more and feature dimensions overlap less. The federated transfer learning method is adopted in the case of less overlap between users and user feature dimensions.

2) Data tracing

Data security label: It refers to the integration and encryption of important information related to original data such as data collection source, collection time, provider, hash value, etc., to generate a data security label. Then, through digital watermarking and other methods, the security label is hidden in the original data in a way that does not destroy the use value of the source data. In the traceability process, the security label can be extracted to trace the problem data provider, and the hash value can be compared to determine whether the data has been tampered.

Block-chain traceability: The data traceability information such as data identification, collection source, collection time, provider, and each processing behaviour and processor for the data can be stored in the block-chain to realize the traceability of each processing behaviour of the data. Moreover, the characteristics of block-chain technology, such as decentralization and traceability, can guarantee the authenticity and reliability of the traceability information on the chain.

3) Problem data cleaning

Abnormal data detection and deletion: It refers to the discovery and deletion of potential abnormal data by analysing the difference between abnormal data and normal data. Two kinds of abnormal data that need to be detected in service operations are poisoning data and adversarial data.

Problem data reconstruction: It refers to the reconstruction of the input data, on the premise of retaining the original sample semantics, the adversarial disturbance added by the attacker on the real sample is destroyed, so as to prevent the adversarial example attack, algorithm back door attack, etc.

Problem data repair: It refers to abnormal data that can be repaired into usable normal data by analysing the difference between abnormal data and real data.

4) Data fairness enhancement

Data distribution modification: It refers to the machine learning algorithm that is trained on the data set with balanced distribution by modifying the distribution of training data, so that the prediction results of the algorithm model on any input data are fair.

5) Data security evaluation

Data leakage security evaluation: It refers to the simulation of the data theft behaviour by simulating data theft, member reasoning attack, data reverse restoration and other methods to evaluate the data security of using AI technology.

7.3.4 Platform security technology

1) Vulnerability mining and repair

Code auditing: The static code audit technology can detect security vulnerabilities and non-standard coding problems in the code of a machine learning open source framework platform, and find the security threats in the open source framework platform in a timely manner.

Fuzzy testing: Fuzzy testing is one of the mainstream vulnerability mining techniques. Through fuzzy testing of modules such as file parsing and model loading in the open source framework

platform of machine learning, security vulnerabilities in the framework can be found and repaired in advance.

Security response mechanism: By establishing a fast security response mechanism, security problems can be found with the help of community forces such as white hat and a security research team, and security threats of a machine learning framework platform can be reduced.

2) Model verification

Model file verification: By checking and verifying the format, size, parameter range, network topology, node name, data dimension and other key information of the model file, security problems in the model file can be found before the model file is loaded, preventing the model file of a malicious AI algorithm from being loaded.

3) Framework platform security deployment

Trusted environment deployment: By deploying the machine learning framework platform in a trusted environment, the stability of the machine learning framework platform operating environment can be enhanced, and the potential harm to the framework platform can be cut off.

7.3.5 Security techniques in the life cycle of using AI technology

In order to facilitate the organization to choose and refer to the corresponding security technologies, this Technical Report classifies and sorts out the security technologies that may be used according to the life cycle of using AI technology. Table 3 summarizes security techniques in life cycle of using AI technology.

Table 3 – Security techniques in life cycle of using AI technology

Type	Security techniques
Design and manufacturing	
Service security technology	Service compliance assessment – Self-assessment – Third-party assessment
	Service security mechanism – Service access control – Service security isolation – Service security redundancy
Algorithm security technology	Algorithm robustness enhancement – Data enhancement – Robust feature learning – Model randomization – Model regularization – Training data sampling
	Algorithm intellectual property protection – Model watermark
	Algorithm fairness guarantee – Algorithm fairness constraints – Bias discrimination post-processing
	Algorithm interpretability improvement. – Model self-explanation – Algorithm global interpretation – Algorithm partial interpretation

Table 3 – Security techniques in life cycle of using AI technology

Type	Security techniques
	Algorithm security evaluation – Fairness evaluation – Robustness evaluation – Interpretability evaluation
Data security technology	Data privacy computing – Multi-party security computing – Homomorphic encryption – Zero-knowledge certification – Differential privacy – Federal learning
	Data tracing – Data security label – Block-chain traceability
	Problem data cleaning – Abnormal data detection and deletion – Problem data reconstruction – Problem data repair
	Data fairness enhancement – Data distribution modification
	Data security evaluation – Data leakage security evaluation
Platform security technology	Vulnerability mining and repair – Code auditing – Fuzzy testing
	Model verification – Model file verification
Deployment and application	
Service security technology	Service security mechanism – Service security fusing – Service security monitoring
	Security attack detection – Adversarial examples detection – Algorithm backdoor detection
Data security technology	Data privacy computing – Trusted execution environment
Platform security technology	Framework platform security deployment – Trusted environment deployment
Maintenance and upgrading	
Service security technology	Service compliance assessment – Self-assessment – Third-party assessment
Algorithm security technology	Algorithm security evaluation

Table 3 – Security techniques in life cycle of using AI technology

Type	Security techniques
	– Fairness evaluation – Robustness evaluation – Interpretability evaluation
Data security technology	Data security evaluation – Data leakage security evaluation
Abandonment and destruction	
Data security technology	Data security evaluation – Data leakage security evaluation

Bibliography

[b-ITU-T Y.2091] Recommendation ITU-T Y.2091 (2011), *Terms and definitions for next generation networks.*

[b-ITU-T Y.3172] Recommendation ITU-T Y.3172 (2019), *Architectural framework for machine learning in future networks including IMT-2020.*

[b-ITU-R M.1224-1] Recommendation ITU-R M.1224-1 (2012), *Vocabulary of terms for international mobile telecommunications (IMT).*

[b-ISO/IEC 22989] ISO/IEC 22989:2021, *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*.

[b-ISO/IEC 27000] ISO/IEC 27000:2012, *Information technology – Security technique – Information security management systems – Overview and vocabulary*.

[b-ISO/IEC 29182-2] ISO/IEC 29182:2013, *Information technology – Sensor networks: Sensor Network Reference Architecture (SNRA) – Part 2: Vocabulary and terminology*.

[b-NIST-SP-800-30] NIST Special Publication 800-30 (2012), *Guide for Conducting Risk Assessments.*