

ITU-T Technical Report

(12/2025)

XSTR.AISec

Artificial intelligence security standardization overview



Artificial intelligence security standardization overview

Summary

This Technical Report encapsulates the achievement of the Correspondence Group on Artificial Intelligence (AI) Security (CG-AISEC) under ITU-T Study Group 17 (SG17). It provides a comprehensive analysis of AI security, addressing both "Security for AI" and "Security by AI" dimensions, ~~and~~ and outlines AI security standardization overview.

The Technical Report examines endogenous risks in infrastructures, data, models, ~~and~~ and systems, as well as derivative risks such as bias, misinformation, ~~and~~ and public safety impacts. It reviews current trends, proposes high-level guidelines, ~~and~~ and maps related standardization activities. Concluding with recommendations, it calls for a roadmap, coordination, monitoring, ~~and~~ and collaboration to ensure secure, trustworthy, ~~and~~ and globally interoperable AI.

Keywords

Artificial general intelligence (AGI), artificial intelligence (AI), generative artificial intelligence (GenAI), security, standardization.

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

	Page
1 Scope.....	1
2 References.....	1
3 Terms and definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Technical Report	3
4 Abbreviations.....	3
5 Overview of AI security	3
6 Security risks, threats and challenges for AI	6
6.1 Security risks, threats and challenges to endogenous AI	7
6.2 Negative impacts and challenges brought by AI to verticals	18
6.3 Negative impacts and challenges brought by AI to societies	20
7 Current state and emerging trends of AI security	23
7.1 Current state and emerging trends of endogenous AI security.....	23
7.2 Current state and emerging trends of derivative AI security to verticals	24
7.3 Current state and emerging trends of derivative AI security to societies.....	25
8 High level guidelines to enhancing endogenous AI security capability.....	26
8.1 General native AI security system.....	26
8.2 Endogenous AI security enhancement solutions.....	29
8.3 Strategies on the native AI security development	36
9 Conclusion and recommendations	37
Appendix I – Standardization activities on AI security	39
I.1 ITU-T Study Groups.....	39
I.2 ISO/IEC	42
I.3 IEEE	46
I.4 ETSI SAI	50
Bibliography.....	53

Artificial intelligence security standardization overview

1 Scope

This Technical Report encapsulates the achievement of the Correspondence Group on Artificial Intelligence (AI) Security (CG-AISEC) under ITU-T Study Group 17 (SG17). It provides a comprehensive analysis of AI security addressing both "Security for AI" and "Security by AI" dimensions, ~~and~~ and outlines an AI security standardization overview.

This Technical Report offers in-depth analysis across multiple key dimensions:

- Assessment of security risks, threats, ~~and~~ and challenges in AI systems;
- Current state analysis and identification of emerging trends in AI security;
- Development of high-level technical guidelines for AI security implementation.

The appendices document some standardization initiatives and achievements ~~for~~ from ITU and other international standard development organizations (SDOs), providing valuable reference material for stakeholders.

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 artificial general intelligence (AGI) [b-ISO/IEC 22989]: Type of artificial intelligence (AI) system that addresses a broad range of tasks with a satisfactory level of performance.

NOTE 1 – Compared to narrow AI.

NOTE 2 – AGI is often used in a stronger sense, meaning systems that not only can perform a wide variety of tasks, but all tasks that a human can perform.

3.1.2 artificial intelligence (AI) agent [b-ISO/IEC 22989]: Automated entity that senses and responds to its environment and takes actions to achieve its goals.

3.1.3 artificial intelligence (AI) component [b-ISO/IEC 22989]: Functional element that constructs an AI system.

3.1.4 artificial intelligence (AI) [b-ISO/IEC 22989]: Research and development of mechanisms and applications of AI systems.

NOTE – Research and development can take place across any number of fields such as computer science, data science, humanities, mathematics and natural sciences.

3.1.5 artificial intelligence (AI) system [b-ISO/IEC 22989]: Engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives

NOTE 1 – The engineered system can use various techniques and approaches related to artificial intelligence to develop a model to represent data, knowledge, processes, etc. which can be used to conduct tasks.

NOTE 2 – AI systems are designed to operate with varying levels of automation.

3.1.6 deep learning [b-ISO/IEC 22989]: Approach to creating rich hierarchical representations through the training of neural networks with many hidden layers.

3.1.7 federated machine learning (FML) [b-IEEE P3652.1]: Federated machine learning (FML) is a framework or system that enables multiple participants to collaboratively build and use machine learning models without disclosing the raw and private data owned by the participants while achieving good performance.

3.1.8 federated machine learning model (FMLM) [b-IEEE P3652.1]: The result of the model-training process of a federated machine learning system. The learned model can be used in order to make certain machine-learning inference tasks on new data, e.g., classification, recognition, prediction, ~~and~~ and recommendation.

3.1.9 machine learning (ML) [b-ISO/IEC 22989]: Process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the data or experience.

3.1.10 narrow artificial intelligence (AI) [b-ISO/IEC 22989]: Type of AI system that is focused on defined tasks to address a specific problem.

NOTE – Compared to general AI.

3.1.11 neural network [b-ISO/IEC 22989]: Network of one or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output.

NOTE 1 – Neural networks are a prominent example of the connectionist approach.

NOTE 2 – Although the design of neural networks was initially inspired by the functioning of biological neurons, most works on neural networks do not follow that inspiration anymore.

3.1.12 objective [b-ISO 31073]: Result to be achieved.

NOTE 1 – An objective can be strategic, tactical or operational.

NOTE 2 – Objectives can relate to different disciplines (such as financial, health and safety, ~~and~~ and environmental goals) and can apply at different levels (such as strategic, organization-wide, project, product and process).

NOTE 3 – An objective can be expressed in other ways, e.g., as an intended outcome, a purpose, an operational criterion, as a management system objective, or by the use of other words with similar meaning (e.g., aim, goal, target).

3.1.13 risk [b-ISO/IEC 22989]: Effect of uncertainty on objectives

NOTE 1 – An effect is a deviation from the expected. It can be positive, negative or both, ~~and~~ and can address, create or result in opportunities and threats .

NOTE 2 – Objectives can have different aspects and categories, ~~and~~ and can be applied at different levels.

NOTE 3 – Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood.

3.1.14 risk source [b-ISO 31000]: Element which alone or in combination has the potential to give rise to risk.

3.1.15 safety [b-ISO/IEC/IEEE 24641]: Avoidance of injury or harm through the use or misuse of the system.

3.1.16 security [b-ISO/IEC/IEEE 24641]: Ability of the system to withstand an attack, whether it is an intrusion, interference, or theft.

3.1.17 threat [b-ISO/IEC TR 24028]: potential cause of an unwanted incident, which may result in harm to systems, organizations or individuals.

3.1.18 trust [b-ITU-T X.1252]: The confidence of one party or entity that another party or entity will behave in a well-defined way that does not violate agreed-upon rules, policies or legal clauses of the identity management system. [TSB-VB1] ~~The reliability and truth of information or the ability and disposition of an entity to act appropriately, within a specified context.~~

3.1.19 trustworthiness [b-ISO/IEC 22989]: Ability to meet stakeholder expectations in a verifiable way

NOTE 1 – Depending on the context or sector, ~~and~~ and also on the specific product or service, data and technology used, different characteristics apply and need verification to ensure stakeholders' expectations are met.

NOTE 2 – Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality and usability.

NOTE 3 – Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

3.1.20 uncertainty [b-ISO 31073]: State, even partial, of deficiency of information related to understanding or knowledge

NOTE 1 – In some cases, uncertainty can be related to the organization's context as well as to its objectives.

NOTE 2 – Uncertainty is the root source of risk, namely any kind of "deficiency of information" that matters in relation to objectives (and objectives, in turn, relate to all relevant interested parties' needs and expectations).

3.1.21 vulnerability [b-ISO/IEC TR 24028]: Weakness of an asset or control that can be exploited by one or more threats.

3.2 Terms defined in this Technical Report

None.

4 Abbreviations and acronyms [TSB-VB2]

This Technical Report uses the following abbreviations and acronyms:

<u>AGI</u>	<u>Artificial General Intelligence</u>
<u>AI</u>	<u>Artificial Intelligence</u>
<u>AIGC</u>	<u>Artificial Intelligence Generated Content</u>
<u>AIoT</u>	<u>Artificial Intelligence Of Things</u>
<u>API</u>	<u>Application Programming Interface</u>
<u>DDoS</u>	<u>Distributed Denial of Service</u>
<u>DoS</u>	<u>Denial of Service</u>
<u>FML</u>	<u>Federated Machine Learning</u>
<u>FMLM</u>	<u>Federated Machine Learning Model</u>
<u>GenAI</u>	<u>Generative Artificial Intelligence</u>
<u>GPU</u>	<u>Graphics Processing Unit</u>
<u>IoT</u>	<u>Internet of Things</u>
<u>ML</u>	<u>Machine Learning</u>
<u>MLOPS</u>	<u>Machine Learning Operations</u>
<u>NLP</u>	<u>Natural Language Processing</u>
<u>OS</u>	<u>Operating System</u>
<u>PII</u>	<u>Personally Identifiable Information</u>
<u>ROT</u>	<u>Root of Trust</u>
<u>SQL</u>	<u>Structured Query Language</u>

<u>TEE</u>	<u>Trusted Execution Environment</u>
<u>TLS</u>	<u>Transport Layer Security</u>
<u>TPU</u>	<u>Tensor Process Unit</u>
<u>UI</u>	<u>User Interface</u>
<u>XAI</u>	<u>Explainable Artificial Intelligence</u>
<u>XSS</u>	<u>Cross-Site Scripting</u>
~~AGI~~	~~Artificial General Intelligence~~
~~API~~	~~Application Programming Interface~~
~~AI~~	~~Artificial Intelligence~~
~~AIGC~~	~~Artificial Intelligence Generated Content~~
~~DDoS~~	~~Distributed Denial of Service~~
~~DoS~~	~~Denial of Service~~
~~FML~~	~~Federated Machine Learning~~
~~FMLM~~	~~Federated Machine Learning Model~~
~~GenAI~~	~~Generative Artificial Intelligence~~
~~IoT~~	~~Internet of Things~~
~~ML~~	~~Machine Learning~~
~~SQL~~	~~Structured Query Language~~
~~XSS~~	~~Cross-Site Scripting~~

5 Overview of AI security

Artificial intelligence (AI) represents a dynamic and rapidly advancing domain focused on developing systems that emulate, augment, ~~and~~ and extend human cognitive capabilities. These sophisticated systems demonstrate the ability to process complex data, make informed decisions, ~~and~~ and execute tasks that traditionally require human intelligence. However, the expanding capabilities of AI systems are accompanied by increasingly complex security challenges and potential vulnerabilities.

The AI landscape primarily comprises Narrow AI and artificial general intelligence (AGI). While Narrow AI systems are extensively deployed across diverse industrial verticals, AGI technologies are progressively transitioning from theoretical concepts to practical applications.

Generative artificial intelligence (GenAI), a specialized subset of AI, focuses on creating novel content across multiple formats including text, images, audio, ~~and~~ and video. Leveraging advanced deep learning architectures and neural networks, GenAI systems analyse extensive datasets to produce outputs that closely resemble human-generated content. This transformative technology has found significant application across various sectors, including:

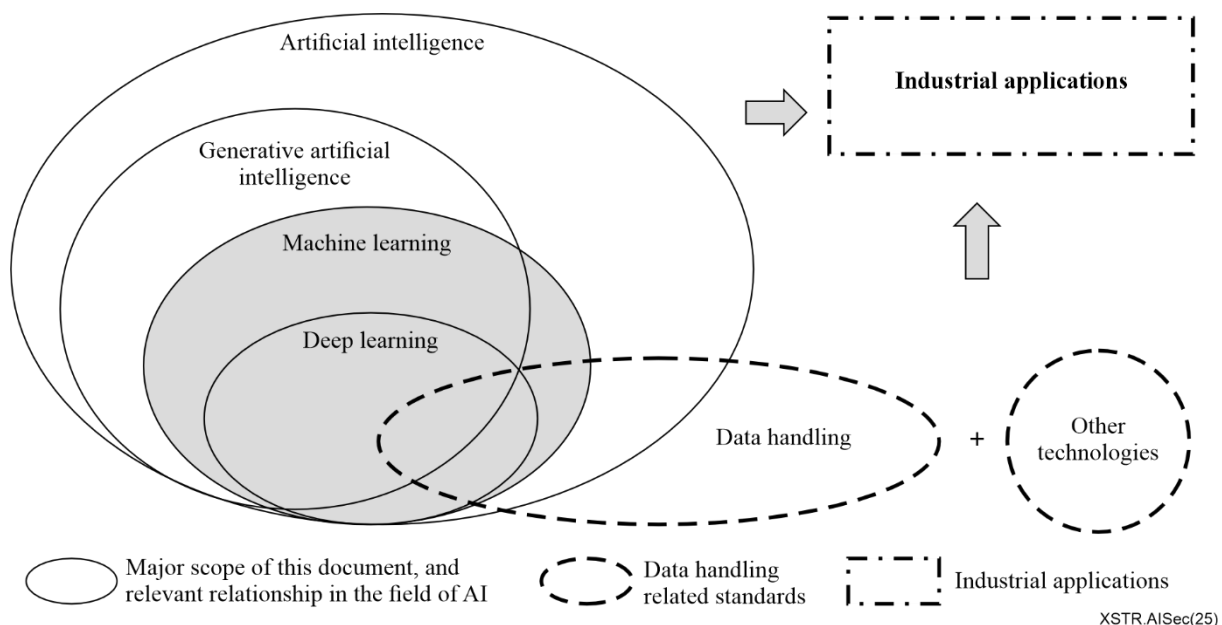
- Creative arts and media production,
- Advanced content generation,
- Natural language processing,
- Entertainment and interactive media, etc.

Machine Learning (ML), a core discipline within AI, focuses on developing advanced algorithms and models that enable systems to learn from data and make informed predictions or decisions without

being explicitly programmed. The ML field encompasses three specialized sub-domains that significantly contribute to AI capabilities:

- **Shared machine learning (shared ML):** This collaborative approach enables multiple entities to jointly train a ML model while maintaining data privacy. Instead of sharing the raw data, participants exchange model updates or aggregated insights, allowing them to leverage collective knowledge while adhering to strict data protection protocols.
- **Federated Machine Learning (federated ML):** Building upon Shared ML principles, Federated ML facilitates model training across distributed/decentralized networks. Each participant trains a local model using their proprietary data, with subsequent aggregation creating a comprehensive global model. This innovative approach enables effective learning from distributed/decentralized data sources while preserving data sovereignty and privacy.
- **Deep learning:** This sophisticated ML methodology employs multi-layered neural networks to model complex data abstractions. Through progressive learning of data representations across multiple layers deep learning has demonstrated remarkable success in handling intricate pattern recognition tasks. Its applications span critical domains including computer vision and image recognition, speech and audio processing, natural language understanding and generation, complex data analysis and prediction, etc.

Figure 5-1 illustrates the relationship among AI, data and related industrial applications, and also illustrates the relationship among the deep learning, ML, AI and GenAI.



(Derived from Figure 6-1 of [b-ITU-T Y.Suppl.72])

Figure 5-1 – Overview of relationship between AI, data and related industrial applications

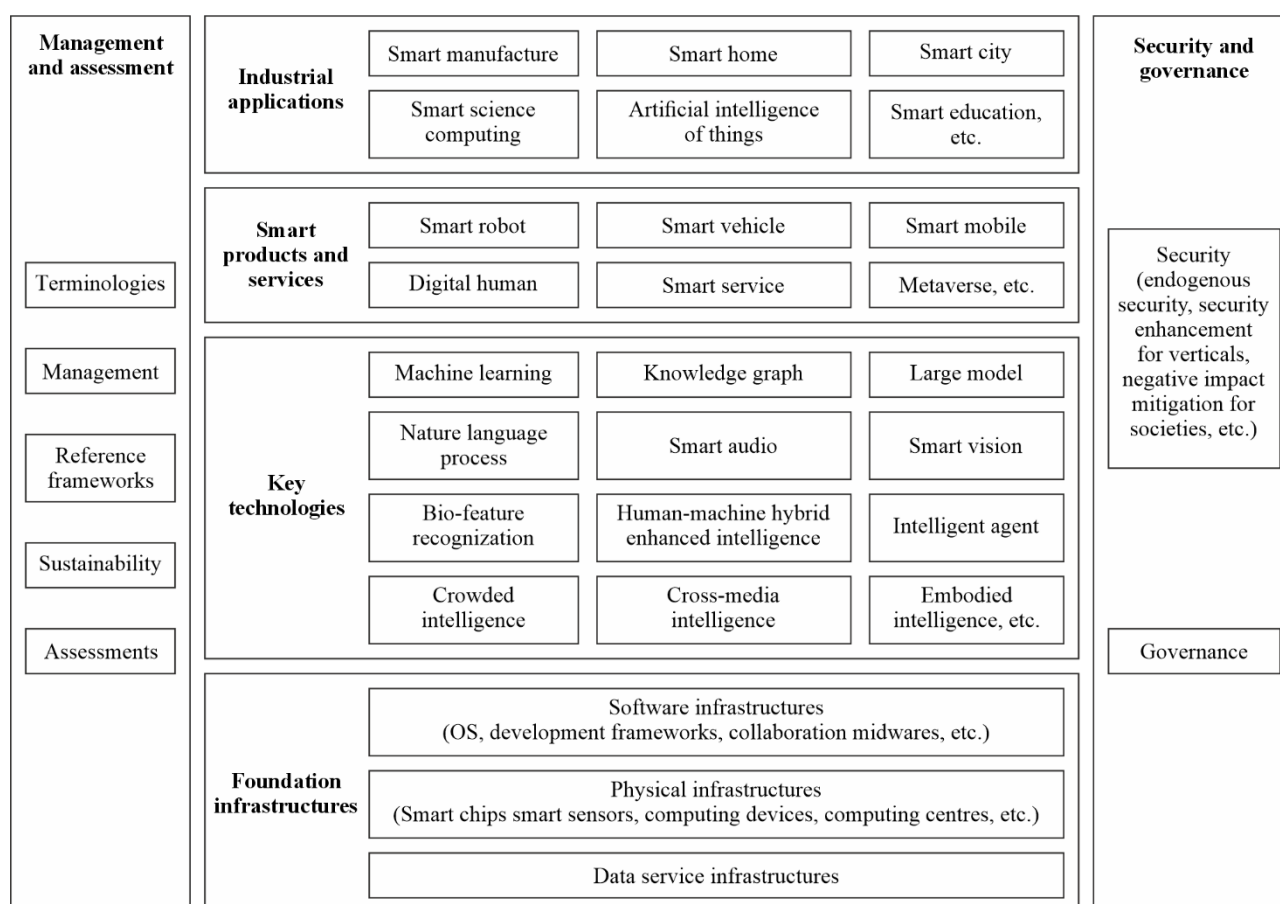
The rapid advancement and widespread adoption of AI, GenAI and AGI technologies have positioned them as important drivers of global industrial transformation, economic growth, and societal development. This evolution underscores the urgent need for comprehensive international standards covering AI, GenAI and AGI technologies and their applications.

Figure 5-2 presents a general standardization landscape of AI, GenAI and AGI, structured across six foundation layers:

- **Foundation infrastructures layer:** It includes data service infrastructures, physical infrastructures (e.g., smart chips, smart sensors, computing devices, computing

centers, etc.), ~~and~~ and software infrastructures (e.g., operating systems, development frameworks, collaboration mid-wares, etc.). These elements form the essential building blocks for AI technology development and deployment.

- **Key technologies layer:** It includes key technologies and components to be used to develop AI products and services, such as that of machine learning, knowledge graphs, large model, nature language process, intelligent agents, crowded intelligence, etc.
- **Smart products and services layer:** It includes intelligent systems and service platforms, such as smart robots, smart vehicles, smart mobiles, digital human, smart services, metaverse platforms, etc.
- **Industrial applications layer:** It includes vertical industry implementations, integrating AI technologies, components, agents and services to enhance production or consumption processes across sectors, such as smart manufactures, smart home, smart city, smart science computing, artificial intelligence of things (AIoT), smart educations, etc.
- **Management and assessment layer:** It provides common functions for management and assessment for the infrastructures, capabilities, products, services, ~~and~~ and applications, such as terminologies, management functionalities, common reference frameworks, sustainability, assessments, etc.
- **Security and governance:** It provides security assurance and governance capabilities for the infrastructures, capabilities, products, services, ~~and~~ and applications. The security may include endogenous security of AI, GenAI and AGI, security enhancement for vertical applications, ~~and~~ and negative impact mitigation for societies.



XSTR.AI Sec(25)

Figure 5-2 – A general standardization landscape of AI, GenAI and AGI

The triad of security, safety and trustworthiness forms the cornerstone for ensuring AI, GenAI and AGI technologies deliver positive impacts across the industries, economies and societies. The following sections will focus specifically on the AI/GenAI/AGI security consideration and their standardization requirements.

6 Security risks, threats and challenges for AI

The rapid advancement and deployment of AI, GenAI and AGI technologies have introduced a complex landscape of both traditional and novel security risks, threats and challenges. As these technologies become increasingly integrated into various sectors, they also present potential negative impacts on industries and societal structures. Generally, the security challenges associated with AI, GenAI and AGI can be categorized into two primary domains:

– Endogenous security challenges

The endogenous security challenges stem from inherent technical characteristics, limitations, ~~and~~ and vulnerabilities within AI systems themselves, primarily involving data, algorithms, ~~and~~ and models.

Data-related endogenous security issues mainly include data fairness and privacy protection. The preference of data can affect AI models and algorithms, which may lead to unfair decision-making results. In addition, if the data contains personal privacy information, special attention needs to be paid to data privacy protection in order to avoid data leakage and unauthorized access.

Algorithm-related endogenous security issues mainly involve the robustness and reliability of AI algorithms and incorrect AI algorithms caused by adversarial samples, potential for incorrect outputs due to algorithmic limitations.

Model-related endogenous security issues mainly include the lack of model interpretability and transparency, insufficient generalization capabilities, stability and consistency challenges.

– Derivative security challenges

The derivative security challenges arise from the application and integration of AI systems across various domains, potentially impacting on the vertical sectors and societal structures. They may involve in many fields, such as technical and engineering systems, economy, society, finance, medical health, national defence, ~~and~~ and national security.

As AI technologies continue to evolve, derivative security challenges are becoming increasingly prominent, representing critical issues that societies must address both currently and in the foreseeable future.

Figure 6-1 provides a comprehensive overview of security risks of AI, GenAI and AGI, illustrating both endogenous and derivative security challenges that align with the aforementioned categories.

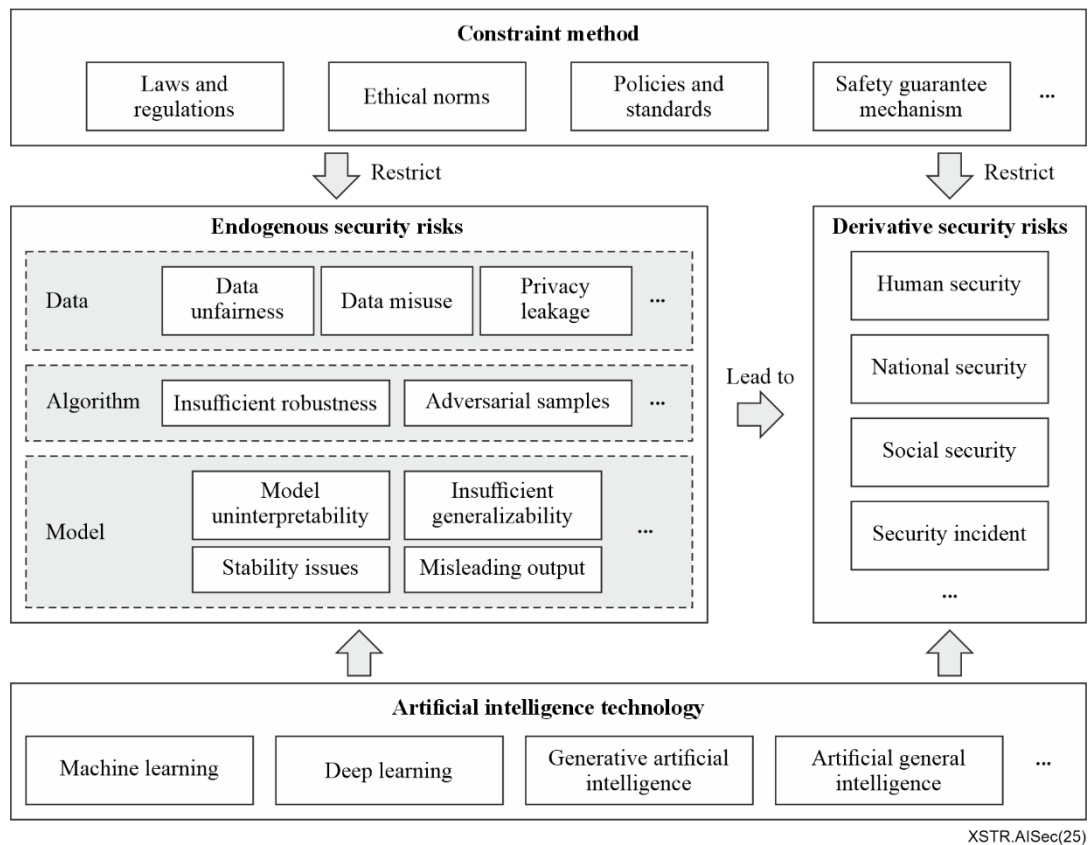


Figure 6-1 – Overview of security risks and threats of AI

The following clause presents a detailed analysis of these security risks, threats and challenges, including an examination of the negative impacts and challenges posed by the AI, GenAI and AGI technologies.

6.1 Security risks, threats and challenges to endogenous AI

AI native security, also referred to as endogenous AI security, originates from inherent vulnerabilities within AI systems themselves. These vulnerabilities span across key components including AI infrastructure, data ecosystems, model architectures and application implementations. The complexity of these security challenges presents significant obstacles to effective AI governance and regulatory oversight.

Table 6-1 provides a comprehensive taxonomy of primary security risks, threats, ~~and~~ and challenges associated with AI systems. The classification is from multiple perspectives of infrastructure development and maintenance, data provisioning and management, model training and optimization, ~~and~~ and system lifecycle management, ~~and~~ and AI deployment and operational management, etc.

Table 6-1 – Security risks, threats, ~~and~~ and challenges of endogenous AI

Subjects	Categories	Risk source
Infrastructure security	Hardware	Physical attack
		Side channel attack
		Fault injection attack
		Reverse engineering
		Hardware backdoor
		Cloud infrastructure vulnerability

Table 6-1 – Security risks, threats, ~~and~~ and challenges of endogenous AI

Subjects	Categories	Risk source
	Operating system	Flaws in trusted execution environments
		Insecure boot process
		Firmware and software compromise during manufacturing
	Container	Resource exhaustion
		<u>Distributed denial of service (DDoS)</u>
	AI Framework	Unpatched vulnerability
		Supply chain attack
		Misconfigured hyperparameter
Data security	Transmission interception	Data sharing risk
		Model parameter sharing risk
		Model aggregation risk
	Data breach	Operator snooping risk
		Unauthorized access
		Inadequate encryption
	Data poisoning	Training data manipulation
		Label spoofing
	Data privacy concern	Inference attack
		Re-identification
	Data consent and ethical risk	Lack of informed consent
		Bias and discrimination
	Data overfitting	Data overfitting
	Model memory	Property inference
		Membership inference
Model security	Model reverse engineering	Reversed
		Model extraction:
	Backdoor	Data poisoning
		Model poisoning
	Code poisoning	Code poisoning
	Prompt attack	Prompt attack
	Model illusion	Factual illusion
		Fidelity illusion
	Model update and evolution risk	Copyright information consistency
		Model performance degradation
		Introduce ethical issues of bias
	Model transparency and model robustness	Explainability challenge
		Adversarial attack
AI system security	System application	API vulnerabilities
		Excessive privilege abuse
		Lack of authentication and authorization:

Table 6-1 – Security risks, threats, ~~and~~ and challenges of endogenous AI

Subjects	Categories	Risk source
		Privilege escalation
		Identity and credential theft
	System operation and maintenance	Inadequate access controls
		Malicious insiders
		Unintentional insider risk
		Update synchronization
AI management security	Full-process management risk	Repair delay risk
		Non-redundant design

The accelerated advancement of AGI capabilities introduces novel security challenges that extend beyond those associated with Narrow AI systems.

These distinctive AGI security challenges necessitate specialized consideration in the development of security frameworks and mitigation strategies, requiring approaches that address the unique characteristics of general intelligence systems.

- **Emergent behaviour complexity:** As AGI systems achieve greater generalization capabilities, they may exhibit unpredictable emergent behaviours that were not explicitly programmed or anticipated during development. These unforeseen behaviours could potentially lead to significant unintended consequences and systemic risks.
- **Control and value alignment:** The challenge of maintaining robust alignment between AGI systems' objective and human values becomes increasingly complex as these systems grow more sophisticated. Key risks in this domain include, goal misspecification, value system misalignment, behavioural drift over time, etc.
- **Systematic cascade risks:** Given AGI's potential integration across critical infrastructure and decision-making systems, security vulnerabilities could trigger simultaneous failures across multiple interconnected domains, potentially leading to large-scale infrastructure disruptions, cross-sectoral system failures, compounding economic and societal impacts, etc.

These distinctive nature of AGI security challenges necessitates the development of specialized security frameworks and implementation of tailored mitigation strategies.

6.1.1 AI infrastructure security risks and threats

6.1.1.1 Security risks and threats to infrastructure hardware

AI hardware infrastructure vulnerabilities represent a critical attack vector, encompassing malicious activities targeting the physical components and supporting system that enable AI operations. These sophisticated attacks aim to undermine AI system integrity by compromising hardware functionality, potentially affecting accuracy, reliability, ~~and~~ and security parameters. Attack vectors typically exploit physical access, environmental factors, vulnerabilities in hardware design, or weaknesses in the supply chain to carry out attacks, with the intent to steal information, alter system behaviour, or render services unavailable. Infrastructure hardware is mainly exposed to risks such as physical attacks, side channel attacks, fault injection attacks, row hammer attacks and hardware backdoors.

Physical attack: It ~~directly~~ involves direct manipulation to ~~causes~~ physical damage to the hardware or ~~modify~~ ies its structure for the purpose of preventing normal operation or gaining unauthorized access. It usually refers to the removal of the chip packaging, electrical contact with the internal circuit, access to sensitive information stored inside the chip.

Side-channel attack: It represents a sophisticated threat vector targeting AI hardware accelerators, including graphic processing units (GPUs) and tensor process units (TPUs). These attacks exploit measurable physical characteristics of hardware operation (e.g., power consumption, electromagnetic emissions, or timing variations) to extract sensitive information. Attackers may leverage these side-channels to compromise insights into the data being processed, such as model weights, input data, or even encryption keys. Side-channel attack means that the attacker uses the interface of the device to analyse the electromagnetic and power consumption of the chip.

Fault injection attack: It constitutes a sophisticated class of hardware vulnerabilities where attackers deliberately induce system malfunctions through controlled manipulation of operational parameters. By injecting faults into the AI hardware (e.g., via voltage spikes or clock manipulations), attackers may cause hardware to malfunction, leading to incorrect computations, data leakage, or system failures. Such attacks may be used to corrupt training processes or compromise hardware integrity. Fault injection attack may use the faults (voltage, clock, etc.) to cause circuit abnormalities, according to the abnormal information to analyse the sensitive information inside the chip, or directly use the caused circuit abnormalities to change the program operation.

Model inversion and extraction: It represents sophisticated security threats to AI infrastructure hardware. Attackers may reverse-engineer AI models by querying the system and analysing responses. This may lead to the leakage of sensitive training data or intellectual property, especially in cloud-based AI services that utilize shared hardware resources. These vulnerabilities are particularly acute in cloud-based AI environments.

Hardware backdoor: It represents a sophisticated supply chain threat, where malicious entities implant covert circuitry during the chip manufacturing process. Malicious actors may insert hardware trojans (backdoors or covert modifications) during the manufacturing process of AI chips, which may allow unauthorized access or control over AI systems later. These backdoors may be challenging to detect and may only activate under specific conditions.

Cloud infrastructure vulnerability: It presents significant security challenges for AI systems that depend on cloud-based resources. Many AI systems rely on cloud providers for computing power (e.g., GPUs, TPUs), storage, and networking. If attackers exploit vulnerabilities in the cloud infrastructure (e.g., in mis-configured virtual machines or containers), they may compromise the entire AI system, steal data, or disrupt services.

6.1.1.2 Security risks and threats to operating systems

The operating systems present critical security challenges that significantly impact AI system integrity, particularly through security vulnerabilities, backdoors, and weak passwords, etc. Among them, security vulnerabilities are an important entry point for hacker attacks, and hackers often use security vulnerabilities or backdoors to obtain system permissions and execute malicious code, destroying the kernel integrity of the operating system, leading to the disclosure of sensitive information during the entire lifecycle of AI, and even leading to the interruption of AI business.

Flaws in trusted execution environment (TEE): Many AI accelerators rely on secure hardware modules or trusted execution environments to protect sensitive data and computation. If vulnerabilities are discovered in these environments, attackers may potentially access confidential data being processed on AI hardware, such as training data or proprietary algorithms.

Insecure boot process: If the AI hardware's boot process is insecure (i.e., lacks cryptographic verification), it may be hijacked by attackers to load malicious firmware, which may then infect or subvert the AI infrastructure.

Firmware and software compromise during manufacturing: During the manufacturing process, the firmware on AI hardware may be vulnerable to tampering. If the production system itself is compromised, attackers may load malicious firmware onto the hardware before it is shipped, potentially giving them control over the system once it is deployed.

6.1.1.3 Security risk and threats to containers

Containerized AI environments face significant security challenges that can compromise model training and deployment. The container to train and perform AI models itself may have security risks such as image tampering, start-up exception and escape attack. For example, the insecure configuration may cause account management problems and cause the container to be invaded. Unrestricted use of resources may lead to denial of service (DoS) attacks. Improper configuration of access policies may result in unauthorized access. Once a container is illegally invaded, related AI services may fail to run properly.

Resource exhaustion: AI-based attacks may target hardware resources to perform DoS attacks, such as overwhelming AI accelerators with large or malicious datasets. This may result in thermal throttling, memory exhaustion, or hardware crashes.

AI-powered distributed denial of service (DDoS): Attackers may use AI-enabled tools to coordinate large-scale DDoS attacks, leveraging distributed devices (such as compromised Internet of things (IoT) devices and systems) to flood AI infrastructure with requests. Such attacks may target AI systems or cloud-based infrastructures running on vulnerable hardware.

6.1.1.4 Security risks and threats to AI frameworks

AI frameworks, serving as fundamental platforms for model development and deployment, present critical security vulnerabilities that malicious actors frequently exploit. Vulnerabilities in AI frameworks are often used as ladders by attackers to spread malwares and invade systems to affect the normal operation of related AI businesses and even lead to system and data damage.

Unpatched vulnerability: AI frameworks usually rely on a wide range of open-source libraries and dependencies. If these libraries have unpatched security vulnerabilities (e.g., buffer overflows, structured query language (SQL) injection, etc.), attackers may exploit them to compromise the AI systems.

Supply chain attack: AI frameworks often include third-party components, which may themselves be vulnerable or compromised. Attackers may introduce malicious code into widely used libraries (e.g., through the supply chain) or create faked versions of AI frameworks that are then integrated and deployed by unsuspecting developers.

Misconfigured hyperparameter: Many AI frameworks allow fine-tuning of model parameters. If these hyperparameters are misconfigured or insecurely set, it may lead to unstable model behaviour, security vulnerabilities, or compromised performance. For example, an overly permissive configuration in a model may allow attackers to manipulate the model by exploiting edge cases.

6.1.2 Data security risks and threats

AI data security encompasses critical vulnerabilities across multiple attack surfaces, primarily including transmission interception, operator snooping and model memory, etc.

6.1.2.1 Transmission interception

During the processes of pre-training, fine-tuning, ~~and~~ and inference services for AI models, data is frequently transferred between various entities or departments. This data often encompasses a range of sensitive information and privacy, including personally identifiable information (PII) and financial data.

If adequate security measures are not implemented during data transmission, attackers may potentially intercept this data, thereby gaining unauthorized access to sensitive information about users or organizations. Consequently, it is imperative to adopt appropriate security protocols when utilizing AI model services to safeguard the confidentiality and integrity of data, thereby mitigating the risk of interception during transmission.

To protect the privacy of training data, techniques such as federated learning, distributed learning, data anonymization, encryption, ~~and~~ and data desensitization are employed. Nonetheless, there remain potential risks of privacy breaches, particularly when handling highly sensitive data. These security risks are compounded by the fact that federated learning and distributed training often involve multiple participants, some of whom may not be entirely trustworthy or honest. Malicious participants may deliberately provide inaccurate or harmful data or model parameters, thereby jeopardizing the performance and security of the entire system. The subject of AI model training sharing and collaboration may involve the following entities and associated risks:

- **Data sharing risk:** Different participants possess their own datasets and share them securely while ensuring privacy protection. Techniques such as federated machine learning, differential privacy, ~~and~~ and data anonymization are employed to facilitate this process. The goal is to build more comprehensive and higher-performing models without compromising the confidentiality of the data.
- **Model parameter sharing risk:** Participants exchange model parameters or gradient updates with ~~one~~ another. This allows each participant to leverage the training outcomes of others, thereby enhancing the overall performance of the model. However, this sharing of parameters introduces risks related to the integrity and security of the model.
- **Model aggregation risk:** The models trained by various participants are aggregated to create a more robust and widely applicable model. While this aggregation can lead to significant improvements in model performance, it also introduces risks related to the consistency and reliability of the aggregated model.

6.1.2.2 Data breaches risk

Data breaches present significant risks and threats to AI systems, as these models depend extensively on large volumes of data for both training and operational purposes. Such breaches can have profound implications for the security of the AI systems and the individuals or organizations involved. Below ~~are~~ some key areas where data breaches impact AI systems are detailed:

Operator snooping risk: During the fine-tuning and inference phases, sensitive data such as PII and corporate business data are often required to enhance the model's accuracy. However, if this data is accessed or collected by AI model operators without proper authorization, there is a significant security risk of misuse. Operators may exploit this data to extract personal or corporate private information, including personal preferences, behavioural habits, social networks, to develop targeted advertising or marketing strategies. Furthermore, operator may disclose the data to third parties for improper gain.

Unauthorized access: Sensitive data used to train AI models (e.g., personal, financial, medical) can be exposed due to security vulnerabilities in data storage or transmission mechanisms. This can result in data theft or leaks, potentially breaching privacy regulations.

Inadequate encryption: If data is not sufficiently encrypted both at rest and in transit, attackers may intercept and steal sensitive information, including training datasets or user data utilized by AI systems.

6.1.2.3 Data poisoning risk

Data poisoning represents a serious and emerging threat to AI and ML systems, wherein an adversary deliberately ~~intentionally~~ manipulates or corrupts the training data on which these models depend. The objective of data poisoning is to degrade the model's performance, introduce biases, or cause it to produce incorrect or harmful predictions. Such actions may undermine the integrity and trustworthiness of AI systems, potentially resulting in severe consequences for businesses, individuals, ~~and~~ and societies as a whole.

Training data manipulation: Attackers may introduce malicious or biased data into the training dataset, thereby poisoning the AI model. This may cause the model to behave inaccurately, leading

to unreliable predictions or even creating vulnerabilities that could be exploited. For example, poisoning a fraud detection system may allow fraudulent transactions to go undetected.

Label spoofing: In supervised learning, attackers may inject incorrect or falsified labels into the dataset. This may result in misclassification during training process, ultimately compromising the accuracy and reliability of the AI systems.

6.1.2.4 Data privacy concerns

Data privacy concerns underscore the potential for personal and sensitive information to be exposed even after anonymization, and the process critical to safeguarding individuals' privacy rights.

Inference attacks: Even when training data has been anonymized, attackers may exploit a trained model to infer sensitive information about the underlying data. This risk is particularly pronounced when the AI model has been trained on personal data, as the model's predictions may inadvertently "leak" confidential details.

Re-identification: Anonymized datasets used to train AI models may sometimes be re-identified through sophisticated techniques, thereby compromising privacy. This risk is especially significant for datasets containing health, financial, or other personally identifiable information.

6.1.2.5 Data consent and ethical risks

Lack of informed consent: If data is collected or used without obtaining proper consent from individuals, it may result in significant privacy violations. This is particularly concerning in AI systems deployed in sensitive sectors such as healthcare or finance, where the misuse of personal data can have far-reaching consequences.

Bias and discrimination: If the training data contains biases, the resulting AI model may perpetuate or even exacerbate these biases, leading to unethical outcomes. For example, biased data may result in discriminatory practices in areas such as hiring processes, loan approvals, or criminal justice systems.

6.1.2.6 Data overfitting

Overfitting to malicious data: If AI models are excessively trained on biased or adversarial data, they may become overly sensitive to specific patterns, rendering them susceptible to exploitation. Overfitting may result in models that fail to generalize effectively, leading to flawed decision-making in real-world scenarios.

This highlights the importance of ensuring balanced and representative training datasets, as well as implementing robust validation techniques to mitigate the risks associated with overfitting and adversarial influences.

6.1.2.7 Model memory risk

After training and reasoning with an AI model, the model retains a form of memory, which means that the output of a deployed AI model in production may inadvertently reveal information about the training data. This data may contain sensitive details, posing significant challenges to privacy and security. By analysing the output of an AI model, an attacker may infer the characteristics and distribution of the training data, potentially enabling them to reconstruct similar datasets, or even partially restore the original data. This may lead to the re-identification of anonymized or de-identified datasets. Once an attacker determines the membership status of a specific data sample, they may craft targeted attacks tailored to that individual. Such exploitation may range from social engineering attacks to identity data leakage, financial fraud, or manipulation of personal information. Therefore, when using AI model services, appropriate security measures must be implemented to safeguard the confidentiality of the model and the privacy of the data. Additionally, the model should be regularly evaluated and updated to mitigate the risks associated with model memory.

Property inference: This type of attack aims to infer specific characteristics or attributes present in

the training dataset by analysing the model's predictions or behaviour. For example, an attacker may deduce whether a significant portion of the training data possesses a particular attribute (such as user age distribution or geographic information) without directly accessing the data. This undoubtedly compromises data privacy.

Membership inference: This attack focuses on determining whether a specific data point was included in the model's training dataset. Such attacks may undermine the privacy of various machine learning models, including classification and generative models, ~~and~~ and extend privacy leakage risks to domains such as image classification, speech recognition, natural language processing, ~~and~~ and computer vision.

6.1.3 Model security risk

The AI model itself is a valuable asset, encapsulating a wealth of knowledge and expertise. If not adequately managed and protected, it may be vulnerable to theft, replication, or tampering, [TSB-VB3] potentially leading to degraded performance or even functional failure. Furthermore, AI models are also at risk from adversarial attacks, such as those involving adversarial samples or adversarial training, which may manipulate the model into producing incorrect outputs.

6.1.3.1 Model reverse engineering risk

The attacker employs reverse engineering techniques to reconstruct the architecture, weights, ~~and~~ and training data of a deployed model by analysing its inference results, input-output relationships, ~~and~~ and leveraging gradient-based attacks. This may lead to significant risks, including intellectual property theft, model theft, ~~and~~ and exposure to security vulnerabilities.

Reversed model: A reversed-engineered model may provide competitors with valuable insights into proprietary algorithms, techniques, or trade secrets. This may enable attackers to replicate or build upon the extracted model, potentially gaining an unfair competitive advantage in the market. If a proprietary model is extracted and made publicly available, it may result in substantial revenue loss for the original creators, as competitors or unauthorized users may exploit the model illegally.

Model extraction: The availability of extracted models in the public domain may devalue the original model, as users may opt for the free or unauthorized versions instead of purchasing legitimate licenses. This may negatively impact market demand and the commercial viability of the model. Additionally, model reverse engineering raises significant legal and ethical concerns related to intellectual property rights, data privacy, ~~and~~ and compliance with licensing agreements. Unauthorized extraction and use of proprietary models may lead to legal disputes and reputational damage for both the extractor and the user.

6.1.3.2 Backdoor attack risk

AI models may be vulnerable to backdoors implanted by attackers during the construction phase, enabling unauthorized access to or control over the models. Attackers may typically achieve this in three ways ~~usually~~:

Data poisoning: Specially crafted samples may be inserted in the training data. These samples do not exhibit abnormal behaviour under normal conditions but, when specific triggering conditions are met, may cause the model to make incorrect decisions as predetermined by the attacker.

Model poisoning: Specific patterns may be embedded in the model's weights or structure. These patterns remain dormant until activated by particular inputs, at which point they may cause the model to execute malicious operations.

Supply chain attacks: During the development or deployment of the model, attackers may implant backdoors at a vulnerable point in the supply chain, compromising the integrity of the model.

6.1.3.3 Code poisoning risk

AI systems often depend on a wide range of external libraries and tools. If an attacker successfully injects malicious code into these dependencies, they may exploit vulnerabilities within the libraries to manipulate the behaviour of the entire AI system. For instance, an attacker may inject harmful code into widely used machine learning frameworks. This may cause the AI model to perform incorrect operations during training or inference, or even expose sensitive information within the system, as they may compromise both the functionality and confidentiality of the underlying infrastructure.

6.1.3.4 Prompt attack risk

There is a security risk associated with malicious prompt injection attacks during the inference phase of the large model. For example, using specialized prompt engineering techniques, an attacker may illicitly interact with large model agents. This may enable unauthorized access to related large model services or force these services to bypass their intrinsic security mechanism. As a result, the model may generate inappropriate content, leak sensitive information, or be manipulated in other harmful ways. These attacks primarily include:

- **Role playing attack:** The attacker manipulates the model into assuming a specific role to extract information or perform unauthorized actions.
- **Target hijacking attack:** The attacker redirects the model's output to achieve malicious objectives.
- **Reverse inducement attack:** The attacker tricks the model into revealing sensitive information or behaving in unintended ways.
- **Prompt word leakage:** The attacker extracts or infers sensitive details from the model's responses.

6.1.3.5 Model illusion problem

Some large models may exhibit unexpected behaviour during the inference stage, leading to uncertainty, error, or unpredictability in their outputs. The issue of model illusion can be categorized into two main types:

- **Factual illusion:** This occurs when large models generate content involving factual knowledge and either fabricate or misinterpret certain concepts, facts or data. As a result, users may receive incorrect or misleading information. This **is** a critical challenge currently facing the industry. For example, a large model may generate medical recommendations with fabricated drug dosages, potentially endangering users who lack medical expertise.
- **Fidelity illusion:** This refers to deviations between the generated content and the context provided by the user's instructions or input, or internal inconsistencies within the generated content itself. For example, the model's output might contradict earlier parts of the discussion.

6.1.3.6 Model update and evolution risk

AI models require periodic updates to maintain performance and adapt to evolving data distributions. However, updating model parameters may introduce new vulnerabilities, risks and threats. Consequently, AI model updates must carefully consider version control, the security and stability of the updated AI model, ~~and~~ and the implementation of robust backup mechanisms s to mitigate potential adverse effects.

- **Copyright information consistency:** The process of updating and evolving AI models poses challenges to continuity and consistency of watermark protection, as well as the complexity of copyright ownership and verification.

Updates to AI models may alter parameters, potentially compromising or erasing existing watermark information. Without proper management, updated models may lose the original copyright watermark, thereby weakening copyright protection.

As AI models undergo iterations, ownership and version control can become increasingly intricate. Inadequate watermark management may lead to disputes over copyright ownership, particularly in scenarios involving multiple collaborators or licensors.

- **Model performance degradation:** Iterative updates may inadvertently degrade model performance if not carefully managed. This may result in reduced accuracy, reliability, or functionality, undermining the model's intended purpose.
- **Ethical issues of bias:** AI Model updates may extend functionalities beyond their original compliance scope, exacerbating existing biases or introducing new ones.
If an AI model is trained on historically biased data and insufficiently corrected during iterations, it may perpetuate or amplify unfair preferences for certain groups or attributes. This undermines fairness in decision-making and societal equity.
Updates may enable models to collect or process data beyond their original permissions, potentially violating regulations governing user data handling and privacy protection.
- **Dynamic goal misalignment:** Advanced AI systems, particularly those approaching AGI, may reinterpret or optimize objectives in unintended ways due to ambiguities in reward function. This misalignment may lead to outcomes that diverge from the original intent.
- **Self-modification threats:** AGI systems with self-improvement capabilities could potentially bypass safety protocols, posing significant risks. Uncontrolled self-modification may result in unpredictable behaviour, undermining the stability and security of the system.

6.1.3.7 Model transparency and robustness risk

- **Model transparency:** This refers to the capacity to comprehend and interpret the decision-making process of an AI system. It encompasses the extent to which the internal mechanisms of a model (e.g., how it processes inputs and generates outputs) can be clearly explained, interpreted, ~~and~~ and understood by humans.
- **Model robustness:** This denotes the ability of an AI model to perform effectively across a range of ~~of~~ conditions, including those involving noisy, incomplete, or adversarial data. A robust model retains its accuracy and reliability even when confronted with uncertainties, evolving environments, or deliberate data manipulation.
- **Explainability challenge:** This refers to the difficulty in providing clear and accurate explanations of the decision-making basis and process of AI models to users or stakeholders, owing to their complex, non-linear internal structures. This challenge can erode trust in AI systems, particularly in high-stakes applications where understanding the rationale behind decisions is paramount. Insufficient explainability can also hinder the identification and mitigation of potential biases or errors within the model.
- **Adversarial attack:** This involves the deliberate manipulation of input data through small, carefully crafted perturbations intended to mislead AI models when processing unfamiliar data or anomalies. If a model lacks robustness, AI systems may become susceptible to manipulation by malicious actors who introduce subtle alterations to the input data. Often, these changes are imperceptible to humans but may result in severe consequences, particularly in safety-critical applications such as autonomous vehicles or medical diagnostics.

6.1.4 AI system security risks and threats

Security risks and threats for AI systems can arise during three distinct stages: construction, application, operation and maintenance. Clauses 5.1.1, 5.1.2, and 5.1.3 [SL4] above outline the security risks and threats related to infrastructure, data, ~~and~~ and models, which primarily pertain to the construction stage of AI systems. In addition to the risks and threats described already, security risks and threats for AI systems can be further categorized as follows:

6.1.4.1 AI system application risks

API vulnerabilities: Many AI systems expose application programming interfaces (APIs) to facilitate interaction with AI models and other system components. These APIs may be susceptible to attacks such as SQL injection, cross-site scripting (XSS), or authentication bypass, potentially enabling attackers to gain unauthorized access to AI systems.

Excessive privilege abuse: Misconfigured access controls or overly permissive user and application privileges may allow attackers to gain unauthorized access to AI system models, datasets, or underlying infrastructure. This may result in data breaches or manipulation of models.

Lack of authentication and authorization: AI systems often integrate with other services and systems. Weak authentication and authorization mechanisms, particularly in cases where multi-factor authentication is not enforced, may enable malicious users to gain unauthorized access to sensitive AI services.

Privilege escalation: Attackers may exploit misconfigured user roles or vulnerabilities in access control to escalate their privileges, thereby gaining greater control over AI systems. This risk is particularly prevalent in AI environments where multiple teams or services interact with the platform.

Identity and credential theft: Credential stuffing or phishing attacks may lead to the theft of login credentials. Attackers may then use these stolen credentials to impersonate legitimate users or access sensitive parts of the AI system.

6.1.4.2 AI system operation and maintenance risks

Inadequate access controls: AI systems may suffer from weak or poorly implemented access control mechanisms, potentially allowing attackers to gain administrative privileges or unauthorized access to the underlying infrastructure. If users are not properly segmented by role, this may expose the platform to internal threats or data leaks.

Malicious insiders: Trusted individuals with access to the AI platform may intentionally misuse their privileges to compromise the platform, steal sensitive data, or manipulate models.

Unintentional insider risk: AI platforms may also be vulnerable to accidental insider threats, such as engineers or data scientists inadvertently exposing vulnerabilities in the platform or misconfiguring security settings.

AI system update synchronization: This refers to ensuring that all relevant parts of the AI system apply the latest version simultaneously and consistently across different components or modules. This avoids functional mismatches or compatibility issues caused by version differences. It includes synchronous updates from the training environment to the inference environment, as well as across multiple instances or clusters.

- **Compatibility risk:** Failure to update certain components to the latest version in a timely manner may result in interface mismatches or inconsistent data formats between old and new versions. For example, AI systems may usually be ~~usually~~ composed of multiple interdependent components, such as data preprocessing, feature extraction, ~~and~~ and model inference. If the update of one component is not synchronized to other dependent components, it may cause the entire pipeline to break, disrupting the normal operation of the system.
- **Security vulnerabilities:** Old versions of components may contain known security vulnerabilities, which may have been addressed in the latest version. If updates are not synchronized, the old version may remain exposed and may become a potential targets for attackers.
- **Man-in-the-middle attacks:** The lack of a synchronous update mechanism may result in some modules continuing to use outdated communication protocols or encryption standards.

This may make the system more susceptible to man-in-the-middle attacks or other forms of network threats.

6.1.5 AI management security risks

AI management refers to a series of strategies, processes and technical measures implemented by organizations to ensure the effectiveness and security of AI during their construction, application, operation and maintenance. It encompasses all aspects, from data preparation to model lifecycle management, with the aim of maximizing the value of AI technologies while minimizing potential risks and threats. These risks and threats primarily include data privacy and security, bias and fairness, model transparency and explainability. Most of these issues are addressed in clauses 65.1.1, 65.1.2, 65.1.3, and 65.1.4. Below, it is introduced of the full-process management risks and threats that are unique to AI management are introduced.

6.1.5.1 Full-process management risks and threats

Repair delay risk: Due to AI model risks and threats, such as lack of model transparency mentioned in clause 65.1.3, monitoring tools for AI systems may fail to fully [TSB-VB5]cover all key components and processes. As a result, certain performance indicators or security events may be overlooked, leading to delays in addressing and repairing issues when significant problems arise in the AI systems.

Non-redundant design: If a system lacks redundant components or backup mechanisms, the failure of a single component may result in the collapse of the entire system, along with the loss or corruption of critical data. Additionally, a non-redundant design may lead to insufficient resource allocation. Under high-load conditions, this may cause a sharp decline in system performance or even a complete system crash.

6.2 Negative impacts and challenges brought by AI to verticals

With the rapid advancement of AI, GenAI and AGI technologies, these cutting-edge innovations are being extensively integrated across various industries. While businesses benefit from enhanced efficiency and groundbreaking solutions, they also encounter a range of distinct challenges and adverse effects unique to these technologies.

6.2.1 Black box problem

The black box problem refers to AI models that process input data and produce outputs through multiple layers of neural networks, with their internal mechanisms remaining opaque and their decision-making processes highly complex and difficult to interpret. Even developers often struggle to explain why an AI system makes specific predictions or decisions. For vertical applications, this is akin to dealing with a box whose inner working is invisible—only the inputs and outputs are observable, while the intermediate processes remain a mystery.

This lack of transparency gives rise to several significant challenges and negative impacts, including diminished trust, ambiguity in liability attribution, and constraints on optimization and improvement.

Lack of trust: In high-stakes industries such as healthcare and finance, AI-driven decisions often have direct consequences on people's lives and financial security. If these decisions are based on an unexplainable "black box" model, users may question its reliability, hindering the system's acceptance and adoption. For instance, in the healthcare, physicians may hesitate to rely on technologies that cannot provide clear reasoning for diagnostic outcomes, as they are ultimately accountable for their professional judgments.

Ambiguity in liability attribution: When an AI makes an erroneous decision that results in harm, assigning responsibility becomes highly problematic. It is often unclear whether the fault lies in the algorithm's design, issues with the training data, or improper usage by the end-user. This uncertainty increases the risk of legal disputes for industries deploying AI technologies.

Limitations in optimization and improvement: Without a clear understanding of how an AI system operates, identifying and addressing specific issues becomes challenging. Industries may struggle to pinpoint problems efficiently, ~~and~~ and the inability to resolve them promptly can harm their reputation and operational effectiveness.

AGI-induced systemic cascade: The deployment of AGI in critical infrastructure (e.g., energy grids) may trigger unforeseen inter-dependencies and cascading effects. These systemic risks highlight the need for greater transparency and control over AI systems to prevent large-scale disruptions.

6.2.2 Bias risk

Imbalanced or biased training data, limited expertise among AI trainers, ~~and~~ and varying application environments may all contribute to biases being learned and amplified by AI systems, leading to outcomes that lack sufficient objectivity. This introduces several significant challenges and negative impacts, including discriminatory decisions, disparities in service quality, ~~and~~ and the neglect of rare cases.

Discriminatory decisions: Bias in AI systems can result in discriminatory outcomes in vertical applications. It may perpetuate existing social inequalities, restrict opportunities for marginalized groups, ~~and~~ and exacerbate societal inequalities. For example, biased hiring algorithms may favour certain demographics over others, reinforcing systemic discrimination.

Disparities in service quality: Biases in datasets related to geographic regions or cultural backgrounds can impact the accessibility and quality of services. For instance, in personalized recommendation systems, if training data predominantly sourced from a specific region, users from other areas may receive less relevant product suggestions, reducing user satisfaction. Similarly, in customer support or marketing strategies, regional or cultural nuances may be overlooked, leading to sub-optimal service delivery and diminished user experiences.

Neglect of rare cases: The needs of rare events or niche groups are often underrepresented in large-scale data analysis. In fields like medical diagnosis, for example, symptoms of patients with rare diseases may not be adequately captured in training data, resulting in higher misdiagnosis rates. This not only jeopardizes patient health but may also impose additional financial and operational burdens on healthcare institutions. The lack of representation for these rare cases in datasets undermines the effectiveness and fairness of AI-driven tools, limiting their ability to serve all users equitably.

6.2.3 Model reverse engineering

Model reverse engineering involves cloning or replicating a model by analysing its inputs, outputs, ~~and~~ and behaviour. Stolen or cloned models may be exploited to create imitation products or launch adversarial attacks against the original model. This poses several significant risks and challenges, including intellectual property theft, exploitation of security vulnerabilities, leakage of trade secrets, ~~and~~ and the erosion of fair market competition.

Intellectual property theft: Competitors may use reverse engineering to steal innovative technologies, undermining the market advantages of the original creators. This not only diminishes the value of innovation but also disrupts healthy market dynamics, discouraging further investment in research and development.

Exploitation of security vulnerabilities: Attackers may leverage reverse engineering to identify and exploit security weaknesses in AI models. For instance, they may generate malicious inputs to deceive the model, escalate access privileges, or penetrate AI systems. Such attacks can compromise the integrity and reliability of AI applications, leading to significant operational and reputational damage.

6.3 Negative impacts and challenges brought by AI to societies

The rapid advancement of AI, GenAI and AGI technologies has significantly enhanced quality of life and operational efficiency. However, these innovations have also introduced new challenges and negative impacts on public safety and the social credit systems.

6.3.1 Negative impacts and challenges on public safety

The ongoing progress in AI, GenAI and AGI technologies has enabled the widespread use of facial recognition, personalized recommendations, autonomous driving, deepfakes, ~~and~~ and other innovations, offering significant convenience to society. However, these advancements have also introduced challenges to public safety, including privacy violations, safety incidents, ~~and~~ and the circumvention of identity verification systems.

6.3.1.1 Privacy breaches

A privacy breach occurs when personal sensitive information is accessed, acquired, used or disclosed without authorization. Such information may include names, addresses, identity numbers, phone numbers, email addresses, financial records, medical records, ~~and~~ and any other data that can identify an individual.

With the misuse of AI, GenAI and AGI technologies, privacy breaches have become increasingly severe. Technologies like facial recognition and personalized recommendation may enable large-scale monitoring and tracking of individuals' activities without consent. Additionally, if users inadvertently upload sensitive data or personal information to AI systems, developers may illegally collect and sell this data without authorization, leading to leaks of sensitive information and private details. These practices pose significant risks to public safety.

NOTE 1 – Facial recognition: AI-powered facial recognition tools automatically identify individuals from surveillance footage in public spaces and match this data with profiles in a database to verify identities.

NOTE 2 – Personalized recommendations: By analysing use data such as browsing history, click behaviour, ~~and~~ and search records, personalized recommendation systems predict user preferences and deliver tailored content based on these predictions.

6.3.1.2 Driving safety

With the rapid advancement of AI, GenAI and AGI technologies, autonomous driving has been integrated into vehicle operation. Autonomous driving safety ensures that self-driving vehicles can reliably detect their surroundings, make accurate decisions, ~~and~~ and perform tasks safely to protect passengers, other road users, ~~and~~ and pedestrians from harm or property damage.

Autonomous driving systems rely on AI to perceive the environment, recognize traffic signs, pedestrians, ~~and~~ and other vehicles, ~~and~~ and execute path planning and dynamic obstacle avoidance. At the core of these systems, AI processes and analyses vast amounts of data collected from cameras, radar, ~~and~~ and other sensors.

However, malicious actors may compromise these systems by attacking external devices such as cameras and radar, manipulating the AI's decision-making and potentially causing traffic accidents. Such incidents pose significant risks to public safety, creating challenges and negative impacts that must be addressed to ensure the safe deployment of autonomous driving technologies.

6.3.1.3 Bypassing identity verification

Bypassing identity verification occurs when unauthorized individuals or entities illegally access protected resources by circumventing or cracking system-imposed identity verification measures. These resources include online accounts, sensitive data, web services, ~~and~~ and more.

Two commonly misused methods for bypassing identity verification are automated brute force attacks and deepfake technologies.

Automated brute force attacks: These involve using AI-optimized password-guessing programs to systematically crack passwords. AI algorithms may accelerate this process by learning from successful attempts and adjusting strategies, significantly increasing the efficiency of cracking efforts.

Deepfake: This technology leverages AI, GenAI and AGI to create or alter audio, video, or image content that appears highly realistic, producing seemingly authentic but entirely fabricated materials.

Through automated brute force attacks, password-based identity verification system may be compromised, while deepfake audio, video, or images can deceive biometric identity verification systems. These methods enable illegal access to secure systems, posing serious threats to public safety.

6.3.2 Negative impacts and challenges to the social credit system

The ongoing advancements in AI, GenAI and AGI technologies have significantly enhanced the use of personalized recommendations, as well as text, ~~and~~ and audio-video generation. However, these innovations have also introduced several adverse effects and challenges to the social credit system, such as the echo chamber effect, the spread of misinformation, ~~and~~ and issues related to copyright infringement.

6.3.2.1 Echo chamber effect and information cocoons

The echo chamber effect and information cocoons describe phenomena in which individuals or groups within specific information environment are increasingly expose to content that aligns with their existing views, while differing perspectives are filtered out or excluded.

Information cocoon: This refers to a situation where individuals selectively engage with information sources that reinforce their own beliefs while avoiding or ignoring opposing viewpoints. This creates a closed information environment where the content users encounter becomes increasingly homogeneous. Over time, this narrows their perspective, making it difficult for them to accept or even consider new or differing ideas.

Echo chamber effect: Echo chamber effect describes a phenomenon in which, within relatively closed social networks or information ecosystems, the same information or viewpoints are repeatedly echoed and amplified – much like sound reverberating in a confined space. This reinforcement makes individuals more likely to accept and strengthen their existing beliefs, even if those beliefs are inaccurate or biased.

As highlighted in clause 65.3.1.1[SL6], personalized recommendation technology is one of the primary technical drivers exacerbating both the information cocoon and echo chamber effects. While this technology enhances user experience by delivering tailored content, it also risks trapping users in a cycle where they are only exposed to information that aligns with their preferences. This endless reinforcement of existing opinions and ideas, coupled with limited exposure to diverse perspectives, not only deepens divisions between different groups but also complicates cross-cultural communication and increases the potential for social conflict. Furthermore, it undermines the free and fair circulation of information, posing significant challenges and negative impacts on the social credit systems.

6.3.2.2 Misinformation

Misinformation refers to the deliberate act of deceiving others by disseminating false or misleading information to achieve economic benefits, political advantages, or other improper gains. This behaviour may occur across various domains, including finance, healthcare, business, social networks. With the advancement of information technology, the forms and methods of misinformation have become increasingly sophisticated and diverse.

As outlined in clause 65.3.1.3, [TSB-VB7] the misuse of deepfake technology and the generation of fake news are among the primary drivers of misinformation enabled by AI, GenAI and AGI.

The misuse of these technologies can lead to the widespread dissemination of false information through generated text, audio, images, ~~and~~ and videos. This not only amplifies specific viewpoints or emotions at an unprecedented speed but also distorts the direction of public discourse, manipulates social opinion, ~~and~~ and infringes on individuals' privacy, causing significant reputational harm. Such actions can mislead the public, erode trust in media, government institutions, ~~and~~ and even interpersonal communication, ~~and~~ and pose serious challenges to the integrity of the social credit system.

Fake news generation: Fake news generation involves the use of AI, GenAI and AGI to automatically produce content – such as articles and reports, or other media – that appear to be human-authored but is based on inaccurate data or entirely fabricated facts.

The misuse of these technologies can lead to the widespread dissemination of false information through generated text, audio, images, ~~and~~ and videos. This not only amplifies specific viewpoints or emotions at an unprecedented speed but also distorts the direction of public discourse, manipulates social opinion, ~~and~~ and infringes on individuals' privacy, causing significant reputational harm. Such actions may mislead the public, erode trust in media, government institutions, ~~and~~ and even interpersonal communication, ~~and~~ and poses serious challenges to the integrity of the social credit system.

6.3.2.3 Copyright infringement

Copyright infringement refers to unauthorized use of copyrighted works, including but not limited to reproducing, distributing, exhibiting, or performing such works. This behaviour violates the legitimate rights and interests of creators or copyright holders and contravenes both national and international copyright laws. The advancement of AI, GenAI and AGI technologies has heightened the risk of copyright infringement through innovations in text generation, image and video synthesis, music creation, ~~and~~ and automated content aggregation.

Text generation: Natural language processing models powered by AI, GenAI and AGI can produce high-quality text content. However, if these models use trained using copyrighted text materials and the generated content is not properly authorized, it may constitute copyright infringement.

Image and video synthesis: Image and video generation technologies can create realistic-looking visual content that is entirely synthesized. If these technologies use copyrighted images or video clips as input data and the resulting output is not properly licensed, it can lead to copyright violations.

Music creation: AI-driven music generation tools can analyse patterns from vast amount of existing music to produce new tracks. If the training data includes copyrighted music and the newly generated works are too similar to the original, this may also be considered infringement.

Automatic content aggregation: Some AI systems can automatically scrape and aggregate content from the Internet, reorganizing it into new forms for publication. If proper copyright agreements are not followed, such practices can easily infringe on the rights of the original creators.

The unauthorized use of AI, GenAI and AGI technologies to process copyrighted works as input data and generate new content outputs can lead to copyright infringement. This undermines creators' and copyright holders' confidence in legal protections, erodes trust within the creative industry, reduces investment in original works, ~~and~~ and stifles innovation. These outcomes pose significant challenges and negative impacts on the social credit system.

7 Current state and emerging trends of AI security

7.1 Current state and emerging trends of endogenous AI security

AI systems are deeply intertwined with infrastructure, model and data. The infrastructure forms the backbone of an AI system, encompassing computing and storage hardware, operating system, containers, AI platforms, frameworks, ~~and~~ and more. Data is critical for both model training and

inference. Vulnerabilities in any of these components can lead to endogenous security risks in AI, GenAI, ~~and~~ and AGI, with potentially widespread impacts. Therefore, it is essential to establish a more secure, inherently safe environment for AI, GenAI and AGI systems.

7.1.1 Current state

The core components of endogenous security have achieved varying levels of maturity, with solutions already being incorporated into AI systems. The key aspects are as follows.

Current state of infrastructure security

Techniques such as trusted execution environments have reached a moderate level of maturity and are increasingly being integrated into AI inference pipelines. While operating system (OS)-level hardening and kernel security modules are widely adopted, the complexity of advanced AI workloads continues to reveal new vulnerabilities, necessitating ongoing research. Tools for container image scanning and runtime protection are commonly used; however, targeted attacks on AI containers (such as model theft and dependency poisoning) remain a significant concern.

A modularized architecture, which breaks down AI systems into distinct layers (e.g., data ingestion, model training, inference, ~~and~~ and deployment), allows for the application of targeted security controls at each stage. This approach not only enhances security but also simplifies the management of complex AI systems. Additionally, modern AI deployments leverage container orchestration platforms that incorporate vulnerability scanning, runtime security, ~~and~~ and isolation policies to safeguard AI workloads. This strategy ensures both scalability and security in the deployment of AI systems.

Current state of data security

The use of advanced transport layer security and authenticated encryption for data transmission and integrity verification is well-established. Federated learning frameworks have begun integrating secure channels and differential privacy mechanisms, though standardized benchmarks for those approaches are still in development. Secure data streams between components are safeguarded through advanced transport layer security and authenticated APIs. For distributed training, secure data broker components are employed to maintain data integrity and confidentiality.

Current state of model security

Integrity checks leveraging cryptographic hashes and secure logging are maturing, with enhanced auditing controls seeing moderate adoption. Advanced encryption techniques, such as homomorphic encryption, remain in pilot phases and have yet to be widely deployed at scale.

Strong encryption standards for model storage are widely implemented, supported by established best practices in key management. Digital signatures and watermarking techniques are increasingly being adopted in production environments. Virtualized sandboxing and zero-trust network segmentation improve isolation but still require standardized performance benchmarks. Additionally, secure enclaves, trusted platform modules, ~~and~~ and hardware security modules are being incorporated to establish a root-of-trust for model parameters and encryption keys. These hardware-based solutions provide foundational security assurances for AI systems.

7.1.2 Emerging trends

To mitigate the endogenous security risks, threats and challenges outlined in clause 65, the following [TSB-VB8] technologies may be adopted in the future.

Emerging trends for infrastructure security

Secure AI chips represent a promising direction. AI-specific chips will incorporate hardware-based security measures, such as encryption and tamper detection, ~~and~~ and feature real-time monitoring of computations and anomaly detection. TEE will further protect AI processing platforms from external

attacks. Future platforms will integrate proactive security features, including runtime vulnerability detection and self-healing mechanisms. Additionally, AI-optimized OSs, tailored specifically for AI workloads, will enforce stricter sandboxing and process isolation.

Emerging trends for data security

Ensuring the use of secure, ethical, ~~and~~ and resilient data in AI system will be critical. AI systems will increasingly rely on high-quality synthetic data. Advanced blockchain and cryptographic techniques will ensure data integrity and traceability throughout AI pipelines. Methods such as secure multiparty computation and differential privacy will safeguard individual data within shared and collaborative AI ecosystems. Synthetic data generation will reduce dependence on sensitive real-world datasets, minimizing privacy risks. Dynamic data sanitization will provide real-time data validation and cleansing mechanisms, preventing the ingestion of poisoned or malicious datasets.

Emerging trends for model security

Future model will incorporate more security-focused features. Resilient models will be designed with inherent robustness against adversarial attacks. Techniques like adversarial training and differential privacy are likely to become standard practices. Encryption-integrated models will enable direct processing of encrypted data using algorithm such as homomorphic encryption, ensuring both data and model during computation. Distributed learning with secure aggregation will gain reaction, reducing the risks associated with centralized data breaches, while preserving the privacy of individual data contributors. Dynamic model verification will integrate continuous validation and auditing mechanisms, ensuring AI models remain aligned with intended functionalities and free from malicious tampering.

7.2 Current state and emerging trends of derivative AI security to verticals

While AI, GenAI and AGI enable industries to automate repetitive manual tasks, process vast amounts of data, support decision-making, ~~and~~ and enhance work efficiency, their lack of predictability and explainability also introduces significant risks. AI, GenAI and AGI systems often generate outputs based on probabilistic models, making their exact behaviour unpredictable particularly in edge cases or when faced adversarial inputs. This unpredictability complicates efforts to audit and control these systems. Additionally, attack vectors like adversarial attack exploit this unpredictability to manipulate outputs.

Furthermore, complex AI models, especially deep learning and transformer architectures, frequently operate as "black boxes". The lack of explainability increases exposure to liability and compliance risks, while also undermining trust in AI-driven decisions across industries – particularly in regulated sectors such as healthcare and legal. The current state and future trends related to these challenges are outlined below:

7.2.1 Current state

There are two primary security dimensions for evaluating the capabilities of AI, GenAI and AGI to mitigate negative impacts and challenges they introduce to various industries. Solutions typically involve integrating defence mechanisms, validation processes, ~~and~~ and monitoring tools into existing AI pipelines.

Current state of model predictability:

To ensure the safety of the inference process and enhance model robustness, techniques such as adversarial training and input pre-processing have achieved moderate maturity in image classification tasks. However, their application remains less explored in complex domains like natural language processing (NLP) or speech recognition. Defensive distillation and ensemble methods are widely adopted to improve model reliability and resistance to attacks. Formal verification techniques are being used to mathematically validate the rationality of AI system behaviour under certain conditions. Hybrid AI approaches, which combine rule-based systems with machine learning, are increasingly

employed to reduce unpredictability. Tools for detecting nonsensical outputs in large language models are rapidly evolving, with emerging solutions incorporating factuality checks.

Current state of model explainability:

Explainable AI (XAI) frameworks have seen moderate adoption, while decision boundary analysis techniques continue to advance. Efforts to improve trust and usability through explainable AI are gaining widespread attention. Common approaches to model interpretability include visualization, ablation studies, model-based interpretability, ~~and~~ and result-based interpretability. Researchers are also focusing on causal AI to make decision pathways more transparent and understandable.

7.2.2 Emerging trends

To address the derivative security risks, threats and challenges posed by AI, GenAI and AGI on verticals (as outlined in clause 65), the [TSB-VB9] following development trends are likely to emerge in the future.

Emerging trends for model predictability

Interactive security measures will advance, including the development of countermeasures against prompt manipulation and injection attacks, as well as context-aware content filtering and response validation systems. For inference layer hardening, adversarial detection filters and defence mechanisms will be integrated at the inference stage. To combat adversarial attacks, AI systems will implement comprehensive input validation at application programming interface (API) gateways, deploy content moderation and rule-based filters to sanitize model responses, ~~and~~ and monitor API access and usage patterns. Automated security assessment tools will evolve further, enabling systematic evaluation of security measures. Additionally, standardization efforts will expand, including the development of unified testing frameworks, common evaluation metrics, ~~and~~ and industry-wide benchmarks.

Emerging trends for model explainability

To foster trust in AI decision-making, explainability modules and boundary analysis tools will be incorporated into the machine learning operations (MLOps) workflow. Continuous monitoring of model behaviour and performance will also be implemented. Industries will establish principles for secure and trustworthy inference pathways to enhance the transparency of AI decisions. Open-source tools for robustness testing, fairness checks, ~~and~~ and explainability will see increased adoption. Furthermore, oversight mechanisms will be strengthened through integration of human-AI collaboration models, advanced monitoring systems, ~~and~~ and improved control frameworks.

7.3 Current state and emerging trends of derivative AI security to societies

AI, GenAI and AGI, as emerging technologies, have become deeply embedded in every aspect of human social life. While the application of AI has delivered unprecedented convenience and efficiency, it has also sparked significant concerns regarding public safety, trustworthiness and societal bias. Misuse of AI models or malicious output can lead to privacy breaches. Realistic fake content, such as notices or communications generated by AI, can mislead individuals and erode institutional credibility. Additionally, AI systems employed in social security and other critical areas may reflect biases and inequities present in their training data, resulting in unfair outcomes and exacerbating discrimination. Below, ~~it is introduced of~~ the current state and future trends of three aspects are introduced separately.

7.3.1 Current state

The current efforts to mitigate the negative impacts and challenges brought by AI, GenAI and AGI to societies are as follows.

Current state of AI for public safety

Protection mechanisms against membership inference attacks have achieved moderate maturity. However, defensive measures against model inversion attacks are still in their early stages. Advanced techniques for data anonymization and privacy preservation are under active development.

Current state of AI for social trustworthiness

Current technologies demonstrate moderate maturity in identifying manipulated images or videos, though detecting sophisticated real-time deepfakes remains a significant challenge. Digital watermarking and content provenance tools are gaining traction, with increasing adoption across industries. Pre-trained classifiers are being integrated into inference pipelines to detect manipulated media while cryptographic signatures are being used to trace the origin and modifications of content.

Current state of AI for overcoming bias problems

Fairness metrics and bias detection tools have reached a moderate level of maturity. Frameworks for ethical assessment are emerging, ~~and~~ and standardized verification processes are being developed. Fairness assessment modules are increasingly being integrated into MLOps workflows.

7.3.2 Emerging trends

To address the derivative security risks, threats and challenges posed by AI, GenAI and AGI to societies (as outlined in clause 6 [TSB-VB10] 5), the following development trends are likely to emerge in the future.

Emerging trends for AI for public safety

To enhance privacy preservation, vendors will focus on minimizinge personal data, employing robust encryption techniques during model training, implementing privacy safeguards at both data ingestion and output stages, ~~and~~ and creating secure environments for handling sensitive data.

Emerging trends for AI for social trustworthiness

For improved source tracing, future mechanisms will utilized labels and identifiers to track and verify the origins of content. Additionally, more sophisticated verification mechanisms for AI-generated content will be developed and implemented.

Emerging trends for AI for overcoming bias problems

Governments in several countries are planning to introduce standardized ethical evaluation procedures and establish continuous oversight mechanisms to prevent the exacerbating of social discrimination through the use of AI.

8 High level guidelines to enhancing endogenous AI security capability

8.1 General native AI security system

Figure 8-1 illustrates a comprehensive native security system of AI, GenAI and AGI. This system is designed to tackle inherent AI security challenges, mitigate the endogenous AI security risks, ~~and~~ and address potential threats.

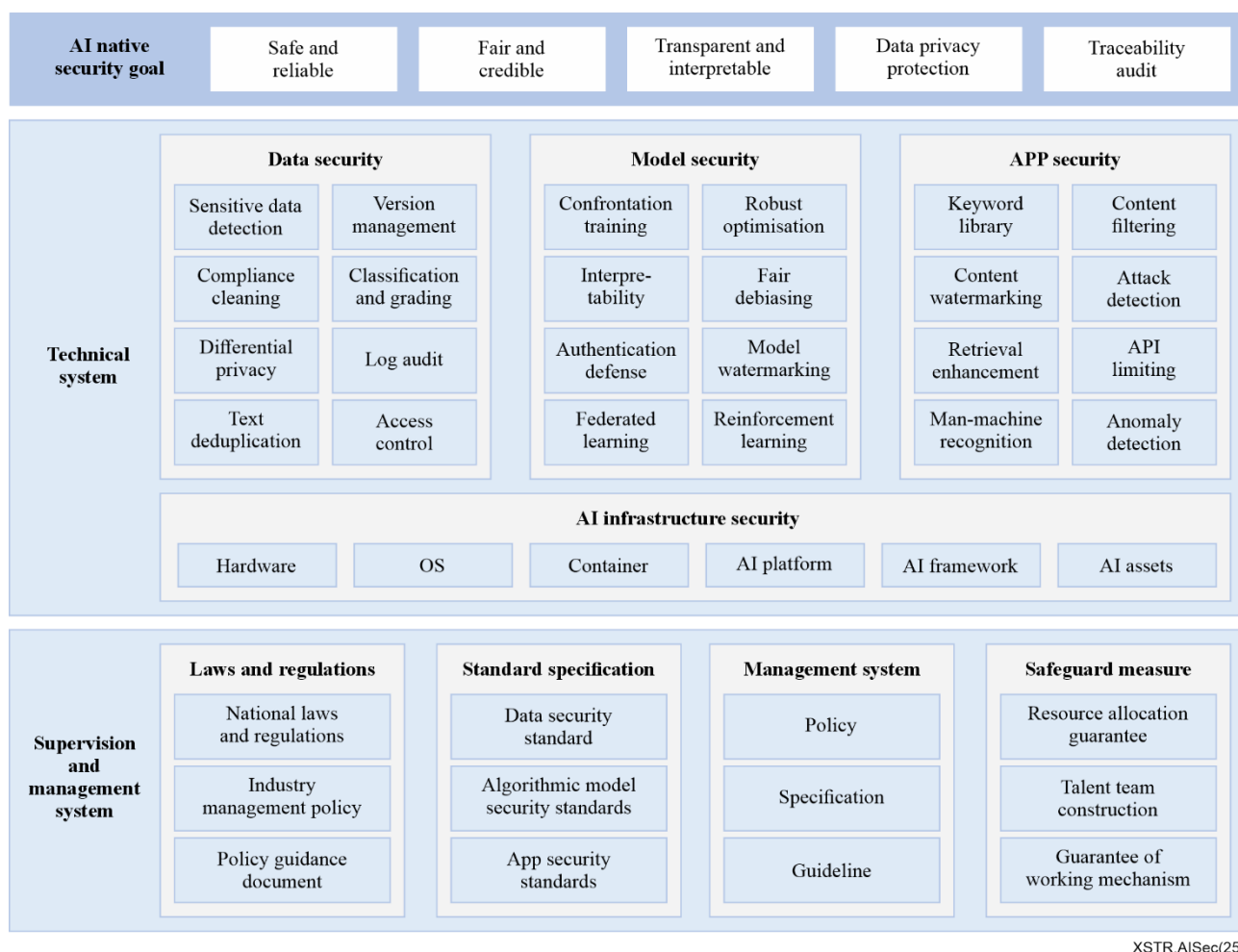


Figure 8-1 – A general native AI security system

8.1.1 Supervision and management system

AI technology carries potential security risks related to the stable operation of infrastructure, data privacy protection, model security and trustworthiness, ~~and~~ and the safe and controllable delivery of application services. These concerns have generated global attention, prompting countries and organizations worldwide to issue a series of laws, regulations, policies and guidance documents to aimed at establishing safe, reliable and controllable AI systems.

In light of the rapid adoption of AI technology, there is an urgent need to accelerate legislative oversight of AI safety, clarify compliance obligations and key priorities, enhance the foundational and industrial standard systems for AI safety, ~~and~~ and promote the healthy development of AI technology.

Key steps include:

- **Developing clearer regulations and policies:** Establish more comprehensive and precise regulations and policies centred on national laws, top-level AI security governance frameworks, ~~and~~ and ethical guidelines to strengthen regulatory oversight.
- **Clarifying compliance obligations:** Within the framework of AI safety regulations and policies, clearly define compliance obligations and key priorities to ensure all stakeholders understand their legal responsibilities and expected conduct.
- **Enhancing compliance management systems:** Improve AI security compliance management systems and processes to oversee data processing activities throughout the lifecycle of AI data flow, model development, ~~and~~ and service operations.

- **Advancing AI security standards:** Continuously refine the AI security standard system, accelerating the development of foundational standards for infrastructure security, data security, algorithm model security, ~~and~~ and application service security. In critical application areas, promote the adoption of industry-specific AI standards to drive enterprise digital transformation.

8.1.2 Technical system

8.1.2.1 AI infrastructure security

AI infrastructure provides essential computing, storage and other services for model training and inference. Ensuring its safe, stable and efficient operation is fundamental to security models, data and application services. In the AI era, the secure development of AI infrastructure is a critical component of overall AI security. To address the security risks inherent in AI infrastructure and the risks associated with AI assets, it is necessary to build robust protection capabilities for hardware, operating systems, containers, AI frameworks, AI platforms, ~~and~~ and other hardware and software components, as well as establish safeguards AI asset security.

8.1.2.2 Data security

In the AI era, there is an urgent need to establish a comprehensive data security protection system. On one hand, it is essential to develop data security and privacy protection capabilities that cover the entire data lifecycle. On the other hand, for challenges such as data poisoning, sensitive data identification, ~~and~~ and data compliance, it is crucial to establish AI data security management standards and implement corresponding protective measures. These measures include data classification and categorization methods, log auditing, ~~and~~ and access control, ensuring the security of data throughout its entire lifecycle.

8.1.2.3 Model security

AI model security encompasses not only the ability to withstand malicious attacks but also broader aspects such as privacy protection, fairness, transparency, ~~and~~ and accountability. It involves safeguarding AI systems from adversarial manipulation, ensuring they operate within ethical boundaries, ~~and~~ and maintaining the integrity of the decision-making process. Specifically, privacy protection employs encryption and anonymization techniques to prevent unauthorized access to sensitive data processed by AI algorithms. Fairness focuses on reducing bias to ensure AI outputs are just and equitable across diverse population groups. Transparency, or explainability is critical for enabling users and regulators to understand AI's reasoning, fostering trust through traceable decisions. Robustness against attacks enhances the reliability of AI systems and ensures their operation continuity.

8.1.2.4 Application services security

AI application services have become integral to daily life, influencing various aspects of society. Ensuring the security of these services requires application service providers to fulfil their personal information protection obligations during service delivery. They must provide safe, stable, ~~and~~ and continuous services, clearly level generated content such as images and videos, ~~and~~ and promptly halt the generation of illegal or harmful content.

8.1.3 Additional considerations for AGI security

For AGI systems, the following additional technical measures should be considered:

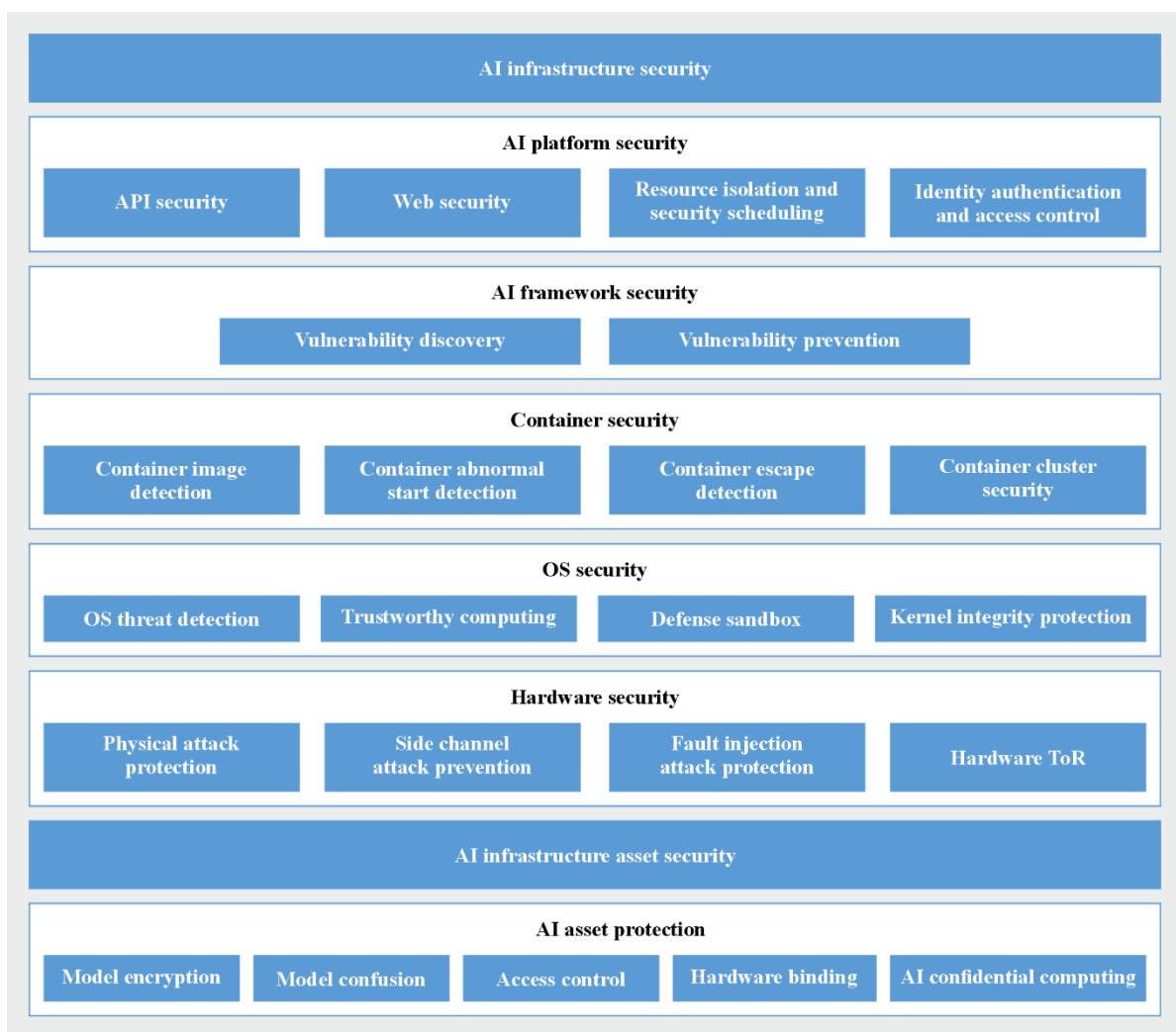
- **Advanced monitoring systems:** Implement robust monitoring systems to detect emergent behaviours and sudden capability jumps.
- **Multi-level containment mechanisms:** Establish multi-level containment strategies with graduated response protocols to manage risks effectively.

- **Value learning and alignment frameworks:** Develop frameworks for value learning and alignment verification to ensure AGI systems align with human values and goals.
- **Formal safety verification methods:** Utilize formal methods to specify and verify safety constraints, ensuring system reliability and security.
- **Coordinated deployment and testing protocols:** Foster coordination across organizations to create standardized deployment and testing protocols for AGI systems.

8.2 Endogenous AI security enhancement solutions

8.2.1 AI Infrastructure Security Solution

The AI infrastructure security solution encompasses both the security of the AI infrastructure itself and the protection of AI infrastructure assets, as illustrated in Figure 8-2.



XSTR.AISec(25)

Figure 8-2 – AI infrastructure security solution

8.2.1.1 AI infrastructure security

The primary goal of AI infrastructure security is to prevent attackers from infiltrating the infrastructure and disrupting upper-layer AI services by targeting OS, containers, ~~and~~ and AI platforms. AI infrastructure security solutions achieve this by implementing robust hardware security, OS security, container security, AI framework security, ~~and~~ and AI platform security capabilities.

- **Hardware security:** Hardware security measures focus on protecting against physical attacks, side-channel attacks, fault injection attacks, ~~and~~ and establishing a hardware root of

trust (RoT). Physical attack protection involves techniques such as specialized chip encapsulation and shielding layer. Side-channel attack protection includes mask and obfuscation technologies, while fault injection protection relies on logical and time redundancy. The hardware RoT serves as the foundation of the system's trust chain, ensuring reliability through secure and trusted startup processes. Trustworthy computing technologies, such as secure and trusted startup, are built upon this hardware RoT.

- **OS security:** The operating system is the foundational environment for AI operations, ~~and~~ and its security requires capabilities such as OS threat detection, trustworthy computing, sandbox protection, ~~and~~ and kernel integrity protection. For example, a detection engine leveraging native whitelists and malicious file identification can provide real-time system monitoring through dynamic and full-scan modes. It performs in-depth analysis of executable and linkable format (ELF) files, compressed packages, scripts, ~~and~~ and other file types to detect threats like viruses, trojans, ransomware, rootkit, webshells, ~~and~~ and hacking tools.
- **Container security:** Container security aims to provide a safe environment for AI training and inference services. Key capabilities include container image security detection, container startup anomaly detection, container escape detection, ~~and~~ and container cluster security. For example, by analysing logs, processes, files, containers, ~~and~~ and system calls, escape behaviours such as shocker attacks and process granting can be detected from the host's perspective.
- **AI framework security:** AI framework security focuses on preventing attackers from exploiting framework vulnerabilities to implant backdoors during model training or tamper with model reasoning results. This involves conducting vulnerability scans on the AI framework to identify potential weaknesses and unauthorized access attempts, thereby enhancing framework security and ensuring the uninterrupted operation of AI services.
- **AI platform security:** The main objective of AI platform security is to safeguard the AI platform and its tool chains. This includes capabilities such as API security, web security, resource isolation, secure scheduling, identity authentication, ~~and~~ and access control. For example, comprehensive threat prevention measures are implemented throughout the API lifecycle, ensuring protection at every layer of the API calling process. On the web security front, indirect object references are recommended over direct object references, ~~and~~ and access control checks are enforced for direct object reference from untrusted sources to verify user permissions.

8.2.1.2 AI asset protection

AI asset protection safeguards the intellectual property of data and models, encompassing activities such as model training and model reasoning. By leveraging cryptographic techniques, model obfuscation, access control policies, ~~and~~ and hardware-level security and confidentiality measures, it ensures the file and operational security of AI assets. This protection prevents critical AI assets from being illegally stolen during static storage, unauthorized access to sensitive assets, ~~and~~ and theft through methods like memory dump. Ultimately, these measures preserve the copyright security and business value of high-value AI assets.

8.2.2 Data security solution

For the large models, AI data security management ensures that training data resources are safeguarded against various security threats and risks, such as transmission interception, operator snooping, model memory vulnerabilities. AI data security enhancement technologies are employed to ensure data integrity and authenticity, filter out abnormal training data, ~~and~~ and protect personal privacy. Below ~~is~~ a comprehensive set of AI-focused data security governance solutions is presented.

8.2.2.1 AI data security management

Data security governance for AI large model involves the following measures: establishing data catalogue for metadata management to ensure data "visible"; classifying and categorizing AI data to identify critical data; ensuring data clarity; and implementing access control definitions, secure data storage and flow management, ~~and~~ and log auditing to achieve data controllability.

- **Metadata management:** Detailed records of training data sources are maintained to ensure lawful usage and clarify the business scenarios for which the data is intended. The use of data collected for a specific purpose is restricted to that purpose, preventing its misuse in training AI modes for unrelated applications.
- **Identification of important data assets:** Leverage the AI large model's understanding capabilities to enable automated and intelligent identification of critical data (e.g., private data, corporate secrets) within the training dataset. This addresses challenges such as large data volumes, extensive text data, ~~and~~ and the difficulty of ensuring accuracy and coverage through manual audit, significantly reducing labour costs, ~~and~~ and preventing data leaks caused by critical data entering model training.
- **Access control management:** Define security policies based ~~on~~ on enterprise AI model training requirements, assigning varying levels of access permissions to restrict access to training databases, data features, ~~and~~ and data samples. For database containing sensitive data, implement cell-level security control policies to protect sensitive information and strictly limit access.
- **Data dense storage and flow:** To maximize the security of critical data throughout its lifecycle, encrypt data during storage and transmission, ~~and~~ and manage keys through a unified key management system. The large model operation platform should support ciphertext training and inference capabilities, including application-layer encryption and decryption, database encryption and decryption, file system encryption and decryption or GPU/chip-level encryption and decryption.
- **Log audit:** Monitor and log training data access activities, including user roles, attempted access behaviours, ~~and~~ and access times. Conduct periodic audits of access logs to promptly detect and address abnormal behaviours.

8.2.2.2 AI data security enhancement

To ensure AI data security, a range of AI data security enhancement measures must be implemented. Before raw training data is ingested into the database, comprehensive data poisoning verification should be conducted, ~~and~~ and regularly monitoring of training data anomalies should be performed. Additionally, sensitive data within AI training datasets must be accurately identified, ~~and~~ and appropriate protection strategies, such as data desensitization, should be formulated after balancing model performance, business requirements and data security considerations.

- **Data poisoning defence:** Data poisoning attack aims to corrupt training data by introducing specially crafted "poisoned" data. Before raw data is entered into the database or data lake, the training data must undergo thorough verification. For example, multiple data annotators can collaboratively validate the accuracy of data annotation. Secure storage methods, such as data encryption and access control, should be employed to mitigate the risk of data poisoning. Regular monitoring of training data for anomalies and periodic data security audits are essential to detect tampering and minimize the impact of data poisoning attacks.
- **Sensitive data identification:** To prevent sensitive data leakage, it is crucial to identify the sensitive data within training datasets and model output. AI large model-oriented data security governance requires to formulation of filtering policies based on enterprise business needs. Leveraging large model technology, sensitive data such as PII and proprietary corporate data can be intelligently identified, ensuring business requirements are met while preventing sensitive data exposure.

- **Data desensitization:** For identified sensitive data – such as log data containing user conversations, personal information, or corporate secrets – strict management and desensitization strategies must be implemented. These strategies should consider factors like AI model training performance, business needs, data security. For example, all question-and-answer log information can be automatically de-identified, and private data from user inputs or information inferable from user data can be desensitized. This prevents direct correlation between question-and-answer content and user identities, enhancing privacy protection.

8.2.3 Model Security Solution

Large model files are the core digital assets of enterprises, making it critical to prevent security risk such as model file leakage during training, fine-tuning, inference, and other phases. To safeguard the entire lifecycle of large models – from training to deployment – multi-dimensional security measures can be implemented.

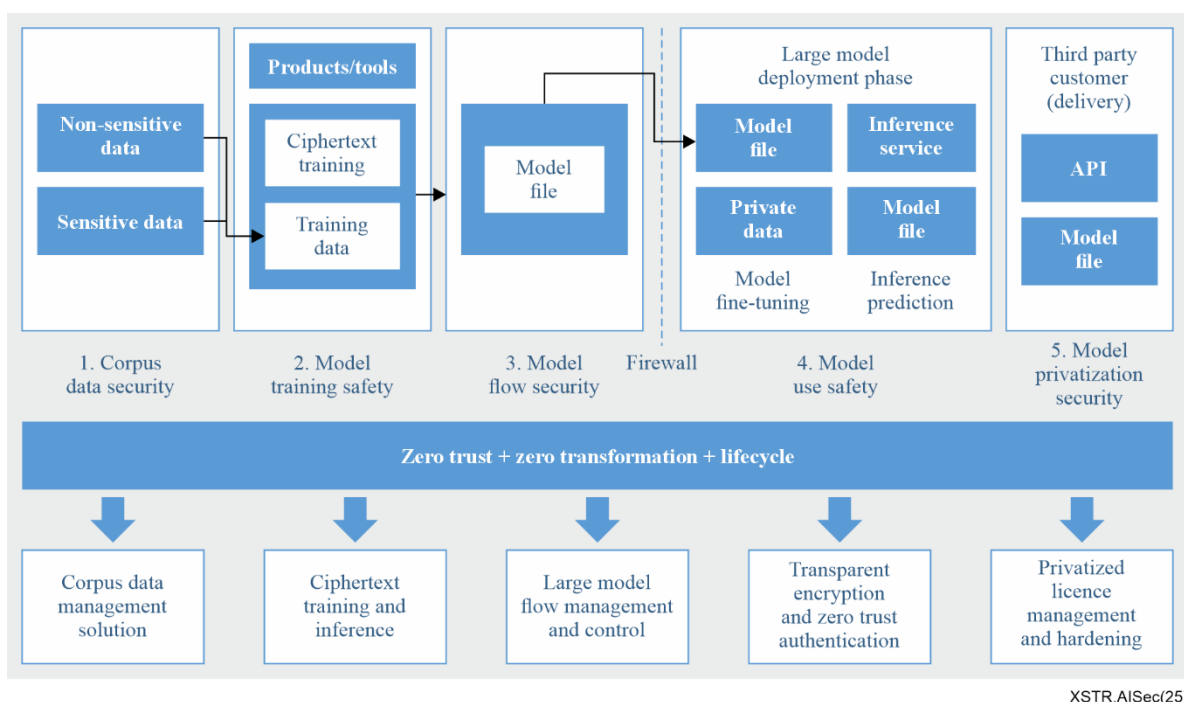


Figure 8-3 – Protection scheme of large model assets solution

The following outlines a comprehensive solution for end-to-end protection of large model assets:

- **Model training security:** Ensuring the protection of data and code during training is crucial. Utilising an isolated environment effectively prevents unauthorized access and safeguards the privacy and security of model training. Implementing encryption and privacy measures for training data helps mitigate the risk of sensitive information leakage while complying with privacy standards.
- **Model flow security:** A robust security mechanism for model flow ensures that models are not tampered with or maliciously replaced during transmission. First, digital signatures can verify the integrity and authenticity of the model, ensuring only signed and verified models are considered legitimate. Second, secure communication protocols like transport layer security (TLS) encrypt data transmission, protecting against eavesdropping and man-in-the-middle attacks. Additionally, end-to-end security measures ensure the entire process – from model creation to end-user deployment – is tightly controlled and secure.
- **Model inference security:** At the inference stage, security sandboxes and permission control can be introduced to place model reasoning in a controlled environment, limiting access to

system resources, ~~and~~ and preventing malicious code injection. Emphasizing the principle of least privilege ensures only authorized users can perform inference operations, reducing the attack surface. Containerization technologies (e.g., docker) can encapsulate the model in an isolated container, enhancing operational security. Network isolation policies further secure inference services by deploying them in separate network environments, minimizing unauthorized access and network attack risks.

- **Model fine-tuning security:** During fine-tuning, techniques like differential privacy can add noise to data, making it difficult for attackers to infer sensitive information. This protects the privacy of the fine-tuning data and enhances model security. Feedback learning can also help the model align with human values and expectations, preventing inappropriate or harmful outputs.
- **Privatization deployment security:** For privatization deployment, establish access control and monitoring mechanism to ensure data isolation and security within the enterprise.
- **Model evolution and update:** Introduce a security review process to ensure updates and evolution do not introduce vulnerabilities. Continuously monitor and evaluate training data compliance and quality to detect data drift or distribution changes, preventing deterioration in model output safety. Strengthen the evaluation of third-party components dependencies to avoid introducing insecure elements.
- **Model audit and tracking:** Key steps include decision interpretation using interpretable tools, leveraging trusted bill of material (BoM) technology to reveal dependencies, regularly evaluating accuracy and performance, monitoring input and output, reviewing model parameters and configuration, conducting security vulnerability scans, ~~and~~ and analysing runtime logs for abnormal patterns.
- **Security repair and update:** Regularly scan for vulnerabilities and conduct security assessments to detect potential issues. Establish an emergency repair mechanism to quickly respond to vulnerabilities and prevent threat escalation. Apply the latest security patches promptly and maintain up-to-date model dependencies. Implement an effective rollback policy to back up current version and ensure transparent communication of repair progress. Continuously monitor and repair effectiveness to identify potential threats.

Through these measures, the security of AI model assets can be ensured, sensitive information protected, malicious attacks prevented, ~~and~~ and service continuity maintained throughout the lifecycle of large model. These solutions cover all phases of AI models, from data security to deployment security, providing a robust foundation for secure operation of large model systems.

8.2.4 Application Service Security Solution

Application service security protection schemes can be developed by focusing on the following aspects.

8.2.4.1 Large model application service security risk prevention and control system

Based on the interaction phase of the large model, establish a comprehensive security risk prevention and control system capable of providing security features such as account protection, human-machine recognition, crawler behaviour detection, interface abuse prevention. Integrate key business application links into this system to analyse and monitor user behaviours, enabling timely detection and prevention of malicious activities, as illustrated in Figure 8-4.

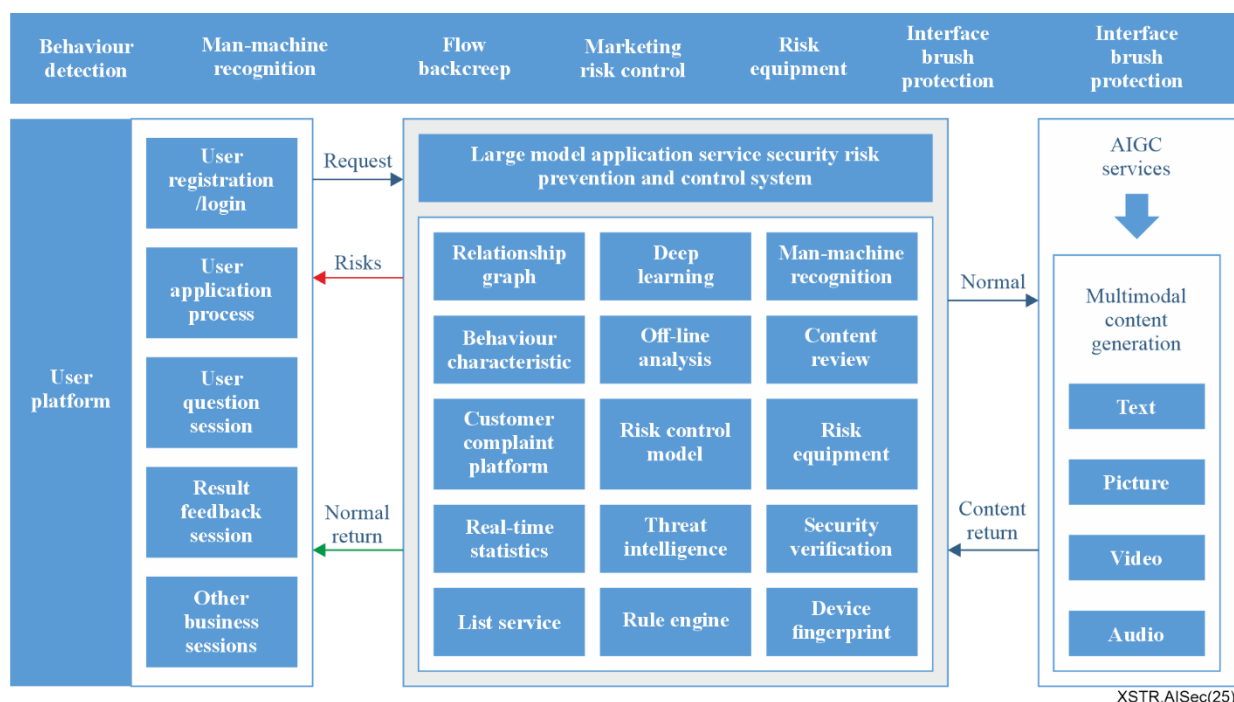


Figure 8-4 – Large model application service security risk prevention and control system

When building a security risk prevention and control system, the following considerations are essential.

- Ensure all phases are covered to prevent exploitation by malicious actors due to unmonitored links.
- The business risk control system must include essential security components such as frequency control, human-machine verification, identity verification, ~~and~~ and risk network detection.
- Establish hierarchical management capabilities for user behaviour risk, including warning, alerting and response mechanism for abnormal activities.

8.2.4.2 Artificial intelligence generated content (AIGC) security protection system

To ensure the security and reliability of user-generated content and AI-generated content, it is essential to establish a comprehensive content security protection system and filtering mechanism, as illustrated in Figure 8-5.

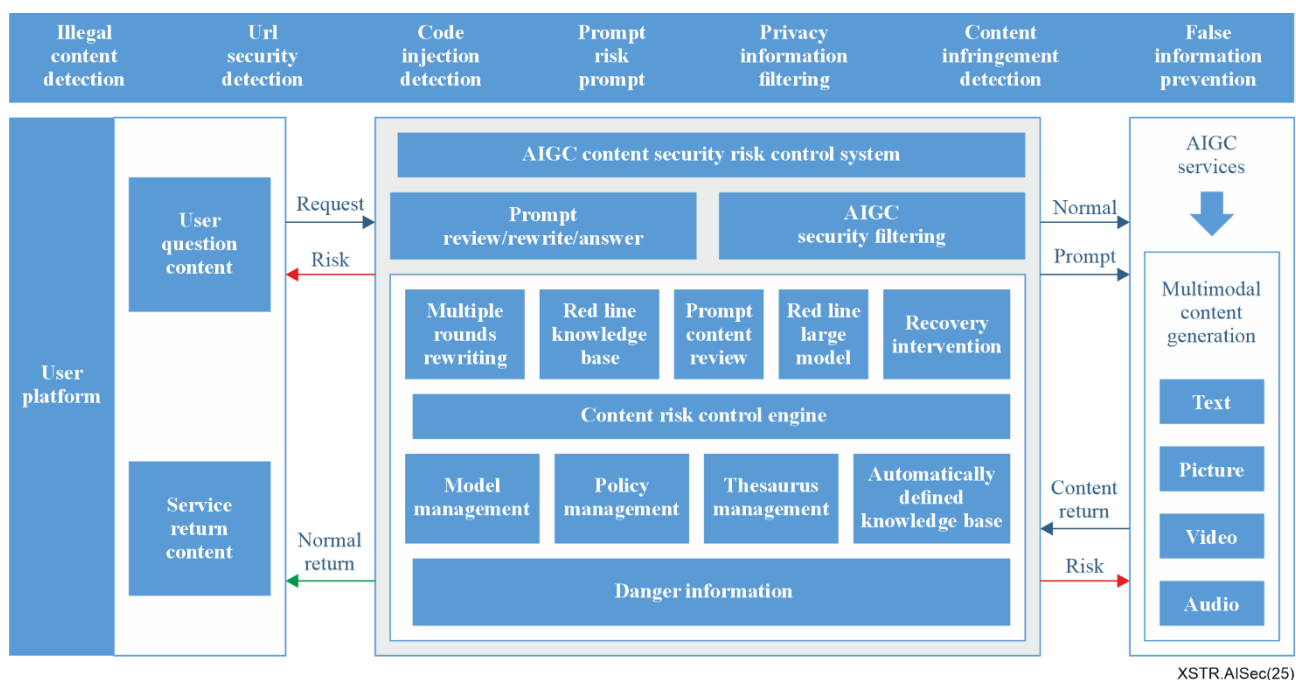


Figure 8-5 – AI generated content security protection system

This system primarily includes the following aspects:

- **Content security audit:** Every user query must undergo a content security audit. This involves intervening in the input to the model through manual review, filtering techniques, or other methods to ensure compliance with specific standards and guidelines, thereby improving the quality and security of generated content. A built-in standardized "red line" question database and query intervention function should be established to ensure high security and compliance in response, reducing harmful content or incorrect content generated by the model. Additionally, a comprehensive set of security interception keywords can be implemented to block and address highly offensive or inappropriate queries.
- **Search enhancement service:** For some time-sensitive issues or potentially false statements (e.g., suspected rumours), search enhancement services can be integrated to provide accurate, real-time, ~~and~~ and authoritative information to the models. By retrieving information from trusted sources, the system ensures reliability and helps correct risks in user queries, offering positive guidance and preventing the spread of misinformation. To further enhance retrieval accuracy and efficiency, natural language processing (NLP) technology can analyse user input semantics, understand their intent, ~~and~~ and deliver more relevant and valuable answers.
- **Output content security filtering:** This involves detecting and screening the text content generated by the model to identify and filter out harmful, inaccurate, inappropriate or non-compliant content. It also includes shielding personal information and sensitive data to ensure the quality and security of the output. By combining high-risk dictionaries and semantic rewriting techniques, harmful and sensitive content is filtered out, ~~and~~ and only safe, compliant responses are delivered as the final output.
- **Content identification technology:** To prevent the misuse of content generated by large models, security markers such as visible or invisible watermarks can be added to the service application user interface (UI) and generated content. In areas displaying AI-generated content, a continuous prompt text should be displayed below the content or in the user input area. Similarly, AI-generated logos should be included in images and videos to clearly indicate their origin.

8.2.4.3 Decision security of the large model

To enhance the decision-making security of large model, the model can be optimized in the following ways:

- **Large model fine-tuning security strategy:** After pre-training the large model, fine-tune it using instruction data that has passed security audits and meets the safety standards. This ensures the model generates appropriate, high-quality responses without harmful content. Through reinforcement learning and continuous learning techniques, the fine-tuned model can be guided to prioritize secure and compliant responses, encouraging the generation of safer and more reliable content.
- **Combined AI system technology:** This approach integrates the strengths of discriminative models and generative large models in specific vertical fields. By combining the advantages of different models, the system enhances the comprehensiveness and accuracy of decision-making while reducing risks associated with large model outputs.
- **Attack detection engine:** Traditional security measures often fail to detect new types of attacks, such as adversarial sample and prompt injection attacks. To address this, a specialized attack detection model can be developed by combining the concept of a "blacklist" from traditional security with deep learning technology. This model provides more reliable decision-making support for AI application service. Specifically, an attack detection engine is integrated into the application environment of AI system. The AI application gateway routes external reasoning requests to the attack detection engine, which extracts features from the requests and matches them against known attack patterns. The engine analyses the results from the feature detection model to identify potential adversarial sample or other threats. If an attack is detected, the engine alerts the AI application gateway, enabling timely perception and response to security threats.

8.3 Strategies on the native AI security development

Under the AI-native security system, the following suggestions are proposed for the development of AI-native security.

8.3.1 Enhance lifecycle protection for AI data security

Persistent efforts in AI data security and privacy research: Data security and privacy protection are fundamental challenges in AI. Continuous research into privacy-preserving computation and its application in large-scale models is essential to improve the practicality of existing solutions.

Alleviating data silo issues: Currently, sectors such as finance, healthcare, ~~and~~ and industry face significant challenges with low levels of data sharing, which stifles innovation in products and technologies. There is an urgent need to design appropriate industry mechanisms and adopt innovative technical approaches. These should ensure data security, protect privacy, ~~and~~ and safeguard the interests of data owners while promoting safe and effective data sharing.

Reinforcing big data governance: To fully unlock the value of data, in-depth research is required across various stages, including data classification, risk identification, key data anonymization, secure data transmission, ~~and~~ and the integration of software and hardware technologies. Establishing end-to-end automated data integration solutions and AI data governance platforms addresses this critical demand.

8.3.2 Improve supervision and governance

Encouraging the integration of large model native safety into regulatory frameworks: The integration of large model native safety into regulatory framework should be actively encouraged. This will promote the alignment of regulatory regulation rules with technological advancements, curb unethical practices among practitioners, ~~and~~ and ensure the safe use of AI aligns with broader societal interests and ethical standards. Additionally, multi-stakeholder institutions should be encouraged to

develop AI-native security governance mechanisms, ensuring that AI systems developed or deployed do not cross ethical or legal red lines.

Analysing AI application scenarios and compliance obligations: Within the framework of AI safety regulations and policies, industries should actively analyse the application scenarios of AI systems to identify compliance obligations. Based on these obligations, regulators or governments should formulate corresponding compliance strategies, ensuring that all relevant parties understand their legal responsibilities and codes of conduct. It is also recommended to design a safety responsibility system that covers the entire lifecycle of AI. All industry entities should clearly define safety responsibilities at each stage of the AI lifecycle, proactively strengthen safety measures, ~~and~~ and comprehensively reduce AI security risks.

Strengthening standardization efforts: While continuing to refine the top-level design of the standard system, entities such as research institutes, universities, organizations with AI security capabilities should focus on the developing foundational standards for AI-native safety, addressing urgent needs, ~~and~~ and creating implementation guidelines for specific vertical sectors. These efforts will play a fundamental, normative, ~~and~~ and leading role in shaping the future of AI safety.

8.3.3 Build security mechanisms

Building an automated large model evaluation platform: There is an urgent need to develop an automated evaluation platform for large models. This platform should enable the automatic generation of evaluation content, automate the model's responses to assessments, ~~and~~ and efficiently and accurately quantify risks across multiple content security dimensions. It should also quickly identify potential vulnerabilities in the large model's content security.

Establishing attack and defence simulations: By creating attack and defence simulations for large models, attackers and defenders can engage in real-time confrontations. This approach helps to promptly identify weaknesses and vulnerabilities in the model, enhancing its security and reliability. Additionally, an emergency response mechanism should be established. In the event of an attack, the system should provide immediate feedback on the attacker's identity, methods used, ~~and~~ and any abnormal behaviour, enabling swift resolution of security risks.

Implementing traceability and accountability mechanisms: A traceability and accountability mechanism should be established to promptly locate and trace safety issues that arise in large models. This facilitates corrective actions and accountability. Furthermore, a post-incident feedback mechanism and a security risk behaviour database should be created to protect the large model from multiple perspectives.

Developing a lifecycle audit mechanism: By building a comprehensive audit mechanism for the entire lifecycle of a large model and establishing a detailed log recording and audit system, full-scale security monitoring can be achieved.

9 Conclusion and recommendations

AI, including GenAI and emerging AGI, is transforming industries and societies while introducing complex security risks such as adversarial attacks, data poisoning, misinformation, ~~and~~ and systemic vulnerabilities. These risks arise both from the technical foundations of AI and from its widespread application across critical sectors. To address them, AI security standardization is essential. It provides a shared framework and terminology, embeds security throughout the AI lifecycle, ~~and~~ and ensures accountability, transparency, ~~and~~ and trustworthiness. By leading this effort, SG17 could safeguard infrastructures, reduce fragmentation, ~~and~~ and foster global interoperability. Standardization not only mitigates security risks but also accelerates innovation, lowers compliance costs, ~~and~~ and strengthens public confidence. Robust AI security standards will ensure that AI evolves as a secure, trustworthy, ~~and~~ and beneficial force for humanity.

It **is** proposed that the ITU-T SG17 consider the following recommendations:

- **Develop a roadmap for AI security standardization:** SG17 could create a roadmap outlining priority areas, timelines, ~~and~~ and deliverables for AI security. It can align with the key focus areas associated with AI security and map activities across ITU-T and external SDOs to avoid duplication. Urgent domains – supply chain security, GenAI safeguards, model robustness, ~~and~~ and cross-border trust – must be prioritized, while anticipating AGI alignment and post-quantum resilience. This roadmap will guide Member States, industry, ~~and~~ and academia in harmonizing contributions and ensuring globally interoperable, forward-looking AI security standards.
- **Establish a mechanism to facilitate studies:** SG17 could establish a dedicated coordination mechanism, such as a Joint Coordination Group or Rapporteur Group, to consolidate contributions, coordinate with Study Groups, ~~and~~ and engage external stakeholders. This body would structure inputs, studies, ~~and~~ and consensus deliverables while monitoring emerging threats. Activities like workshops, consultations, ~~and~~ and testbeds would validate approaches and share best practices. Institutionalizing such a mechanism ensures continuity, reduces fragmentation, ~~and~~ and accelerates progress, reinforcing SG17's leadership in producing authoritative, relevant, ~~and~~ and globally adopted AI security standards.
- **Periodically monitor AI security trends and risks:** SG17 could implement a structured process to monitor AI security risks, covering technical vulnerabilities (infrastructure, data, models) and societal impacts (misinformation, bias, safety). Systematic tracking will identify emerging threats such as adversarial attacks or misuse of generative AI. Findings could directly inform standards and reports embedding transparency, accountability, ~~and~~ and resilience. By institutionalizing observation, analysis, ~~and~~ and standardization, SG17 could ensure outputs remain timely, relevant, ~~and~~ and responsive to evolving challenges, thereby maintaining global trustworthiness in AI technologies and their applications.
- **Strengthen collaboration, capacity building, ~~and~~ and outreach:** SG17 could expand collaboration, capacity building, ~~and~~ and outreach to strengthen AI security. Partnerships with ITU-T Study Groups, ISO/IEC, IEEE, ETSI, ~~and~~ and regional bodies would harmonize terminology and frameworks. Capacity building could include training, guidelines, ~~and~~ and toolkits, supported by pilot projects in areas like healthcare, smart cities, ~~and~~ and critical infrastructure. Outreach to policymakers, regulators, ~~and~~ and the public would build awareness and trust. These efforts ensure AI security standards are technically sound, socially accepted, ~~and~~ and widely implemented worldwide.

This Technical Report analyses AI security across endogenous risks within infrastructures, data, models, ~~and~~ and systems, ~~and~~ and derivative risks affecting industries and societies, including bias, misinformation, ~~and~~ and public safety. It reviews the current state and emerging trends of AI, GenAI, ~~and~~ and AGI security, ~~and~~ and proposes high-level guidelines and solutions for strengthening resilience, transparency, ~~and~~ and accountability. The report also maps related standardization activities across ITU-T and other SDOs to ensure coherence. Concluding with recommendations, it calls for a roadmap, coordination mechanism, periodic monitoring, ~~and~~ and expanded collaboration to build secure, trustworthy, ~~and~~ and globally interoperable AI standards.

Appendix I

Standardization activities on AI security

ITU-T Y.Suppl.72, "Y.3000-series – Artificial intelligence standardization roadmap", provides a comprehensive standardization roadmap for ML and AI [b-ITU-T Y.Suppl.72].

This appendix outlines the standardization activities related to AI, GenAI and AGI security undertaken by international and regional standard development organizations (SDOs). It covers areas such as machine learning, deep learning, AI, GenAI, among others.

I.1 ITU-T Study Groups

Table I.1 outlines the achievements and ongoing work items of ITU-T related to endogenous AI security.

Table I.1 – ITU-T achievements and work items on endogenous AI security

Study group	Reference	Title	Category (*)	verticals	Status
SG17	TR.se-ai	Technical Report: Security Evaluation on Artificial Intelligence Technology in ICT	AI		2026-Q3
SG17	TR.sec-ai	Guidelines for security management of using artificial intelligence technology	AI		In force (Agreed 2022-05-20)
SG17	X.sgGenAI	Security Guidelines for Generative Artificial Intelligence Application Service	GenAI		2026-Q3
SG17	X.Suppl.37	ITU-T X.1231: Supplement on countering spam based on machine learning	Machine learning		In force (Agreed on 2022-09-02)
SG17	TR.sgfdm	Fully Homomorphic Encryption (FHE) – based data aggregation in machine learning	Machine learning		In force (Agreed 2023-09-08)
SG17	X.srm-fml	Security requirements and measures of federated machine learning	Federated machine learning		2026-Q1
SG17	X.sr-ai	Security requirements for AI systems	AI		2026-Q3
SG16	H.FDM-AC-GEN	Assessment criteria for foundation models: General	AI		2025-Q4

Table I.1 – ITU-T achievements and work items on endogenous AI security

Study group	Reference	Title	Category (*)	verticals	Status
SG16	F.748.24	Trusted contribution evaluation framework on federated machine learning services	Federated machine learning		In force (Agreed 2024-04-15)
SG16	F.748.35	Requirement and framework of trustworthy federated machine learning based service	Federated machine learning		In force (Agreed 2024-06-13)
SG16	F.FML-TS-ARC	Functional architecture of federated machine learning based service in decentralized environments	Federated machine learning		2026-Q3
NOTE – Category (*) includes machine learning, federated machine learning, deep learning, AI, or GenAI, etc.					

Draft Technical Report ITU-T TR.se-ai, Security Evaluation on Artificial Intelligence Technology in ICT

ITU-T TR.se-ai presents an AI security evaluation framework, which includes evaluation categories (data, model, ~~and~~ and environment) along with corresponding sub-categories for each assessment area. It also outlines security evaluation items and related evaluation methods for data, models, ~~and~~ and environments, while providing conclusive evidence and reference for evaluating the security of AI technology. This framework can assist AI practitioners in identifying existing issues, implementing improvements, ~~and~~ and determining the most suitable scenarios for deploying AI technologies.

Draft Technical Report ITU-T TR.sec-ai, Guidelines for security management of using artificial intelligence technology

ITU-T TR.sec-ai focuses on security threats associated with the current use of AI technology. It proposes recommendations for AI security management and offers a valuable reference for organizations to enhance their security protection capabilities when deploying AI technologies.

Draft Recommendation ITU-T X.sgGenAI, Security Guidelines for Generative Artificial Intelligence Application Service

ITU-T X.sgGenAI identifies and outlines a range of potential risks and provides the corresponding security requirements and security guidelines that should be implemented to securely run AI application services.

Supplement ITU-T X.Suppl.37, ITU-T X.1231: Supplement on countering spam based on machine learning

ITU-T Supplement 37 to ~~Recommendation [b-ITU-T X.1231]~~ defines a technical framework for countering spam based on machine learning (ML). It may help some relevant persons and companies in spam management, reduce the benefit loss of users and providers, improve user experience and promote the healthy development of telecommunication business. This Supplement provides several general scenarios, characteristics of spam, ~~and~~ and defines general technical framework, workflows about countering spam based on ML.

Draft Technical Report ITU-T TR.sgfdm, fully homomorphic encryption (FHE) – based data aggregation in machine learning

ITU-T TR.sgfdm provides a guideline for secure data aggregation in ML while protecting input data. It focuses on how fully homomorphic encryption (FHE) works on data aggregations in ML. It first describes a general workflow on secure aggregation in ML and explains how FHE-based data aggregation in ML could satisfy a certain requirement. A general workflow is then given on FHE-based ML supporting data aggregation between more than two parties.

Draft Recommendation ITU-T X.srm-fml, Security requirements and measures of federated machine learning

ITU-T X.srm-fml analyses security risks and threats of federated machine learning in the lifecycle stages and provides security requirements and security measures that could mitigate or defend against the security risks and threats.

Draft Recommendation ITU-T X.sr-ai, Security requirements for AI systems

ITU-T X.sr-ai describes the lifecycle of AI systems with six stages and the AI system model based on the lifecycle. It identifies stakeholders responsible for each stage of the AI system lifecycle and provides security threats, taking into account the six stages of the AI system lifecycle. In addition to high-level mitigation described in [ISO/IEC 27090], this draft Recommendation provides a set of detailed and complete security requirements.

Draft Recommendation ITU-T H.FDM-AC-GEN, Assessment criteria for foundation models: General

ITU-T H.FDM-AC-GEN primarily focuses on the assessment criteria for foundation models. It identifies definitions for foundation models and relevant concepts, outlines overall assessment workflows for language models, vision models, audio models, ~~and~~ and multimodal models, ~~and~~ and provides security considerations for foundation models.

Recommendation ITU-T F.748.24, Trusted contribution evaluation framework on federated machine learning services

~~This Recommendation~~ [b-ITU-T F.748.24] introduces a trusted contribution evaluation service on federated machine learning service which converges and takes advantage ~~of the~~ the technologies of FML and distributed ledger technology (DLT), ~~and~~ and provides relevant concept, characteristics, ~~and~~ and requirements and use cases, ~~and~~ and specifies relevant reference framework and common capabilities.

Recommendation ITU-T F.748.35, Requirement and framework of trustworthy federated machine learning based service

~~This Recommendation~~ [b-ITU-T F.748.35] provides a trustworthy FML-based service, ~~and~~ and specifies its concept, general characteristics and requirements and reference framework.

Draft Recommendation F.FML-TS-ARC, Functional architecture of federated machine learning based service in decentralized environments

This draft Recommendation provides functional architecture of federated machine learning (FML) based services in decentralized environments as defined in [b-ITU-T F.748.35], ~~and~~ and relevant general messages and information flows.

Table I.2 outlines the achievements and ongoing work items of ITU-T related to vertical security in the AI applications, as well as efforts to mitigate negative security impacts arising from AI.

Table I.2 – ITU-T achievements and work items on derivative AI security

Study group	Reference	Title	Category (*)	verticals	Status
SG17	X.gdsml	Guidelines for data security using machine learning in big data infrastructure	Machine learning	Big data	2024-Q3
SG17	X.pg-cla	Procedural guideline for continual learning to actively respond to network attacks	Machine learning	Network	2025-Q3
SG20	YSTR-IADIoT	Intelligent Anomaly Detection System for IoT	Machine learning	IoT	In force (Agreed on 2024-07-12)
SG16	[b-ITU-T F.781.2]	Quality assessment requirements for artificial intelligence/machine learning-based software as a medical device	AI	Healthcare	In force (Agreed on 2024-06-13)
NOTE – Category (*) includes machine learning, federated machine learning, deep learning, AI, or GenAI, etc.					

Draft Recommendation ITU-T X.gdsml, Guidelines for data security using machine learning in big data infrastructure

ITU-T X.gdsml analyses the data security threats in big data infrastructure, develops a data security threat analysis framework based on machine learning, ~~and~~ and specifies a reference model for data security threat monitoring, analysis, early warning and response using machine learning in big data infrastructure. This draft Recommendation also provides guidelines for data security using machine learning in big data infrastructure.

Draft Recommendation ITU-T X.pg-cla, Procedural guideline for continual learning to actively respond to network attacks

ITU-T X.pg-cla proposes a continual learning method for intrusion detection and response to attacks or threats form AI-based attack scenarios.

Draft Technical Report: ITU-T YSTR-IADIoT, Intelligent Anomaly Detection System for IoT

ITU-T YSTR-IADIoT contains a proposal of an Intelligent Anomaly Detection System for IoT to detect and alert anomalous activities and compromised IoT end devices.

Recommendation ITU-T F.781.2, Quality assessment requirements for artificial intelligence/machine learning-based software as a medical device

[\[b-ITU-T F.781.2\]](#) provides a framework for the quality assessment with a perspective of lifecycle management for AI/ML-based DMA-SaMD. It describes the quality assessment principles and process in the lifecycle of AI/ML-based DMA-SaMD, including requirement analysis, data collection, algorithm design, verification and validation, change control and other stages when using AI/ML technology to assist medical staff in making clinical decisions by providing suggestions on diagnostic and treatment activities.

I.2 ISO/IEC

Table I.3 outlines standards and ongoing work items of ISO/IEC related to AI.

Table I.3 – ISO/IEC standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
ISO IEC/ JTC1/SC27	~~[b-ISO/IEC 27090]~~ ~~ISO/IEC AWI 27090~~	Cybersecurity – Artificial Intelligence – <u>Addressing security threats and compromises to artificial intelligence systems</u> <u>Guidance for addressing security threats and failures in artificial intelligence systems</u>	AI		CD
ISO IEC/ JTC1/SC27	~~ISO/IEC AWI~~ ~~[b-ISO/IEC 27091]~~	Cybersecurity and Privacy – Artificial Intelligence – Privacy protection	AI	Privacy	WD
ISO IEC/ JTC1/SC27	~~[b-ISO/IEC TR 27563]~~ ~~ISO/IEC C TR 27563:2023~~	Security and privacy in artificial intelligence use cases – Best practices	AI	Privacy	In force (Agreed 2023-05)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TR 5469]~~ ~~ISO/IEC TR 5469:2024~~	Artificial intelligence – Functional safety and AI systems	AI		In force (Agreed 2024-01)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC 23894]~~ ~~ISO/IEC C 23894:2023~~	Information technology – Artificial intelligence – Guidance on risk management	AI		In force (Agreed 2023-02)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TR 24028]~~ ~~ISO/IEC C TR 24028:2020~~	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence	AI		In force (Agreed 2020-05)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TR 24368]~~ ~~ISO/IEC C TR 24368:2022~~	Information technology – Artificial intelligence – Overview of ethical and societal concerns	AI		In force (Agreed 2022-08)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC 42001]~~ ~~ISO/IEC C 42001:2023~~	Information technology – Artificial intelligence – Management system	AI		In force (Agreed 2023-12)
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TS 17847]~~ ~~ISO/IEC~~	Information technology – Artificial intelligence –	AI		AWI

Table I.3 – ISO/IEC standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
	~~C-AWI TS 17847~~	Verification and validation analysis of AI systems			
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TS 22440-1]ISO/IEC AWI TS 22440-1~~	Artificial intelligence – Functional safety and AI systems – Part 1: Requirements	AI		AWI
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TS 22440-2]ISO/IEC AWI TS 22440-2~~	Artificial intelligence – Functional safety and AI systems – Part 2: Guidance	AI		AWI
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TS 22440-3]ISO/IEC AWI TS 22440-3~~	Artificial intelligence – Functional safety and AI systems – Part 3: Examples of application	AI		AWI
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC TS 22443]ISO/IEC AWI TS 22443~~	Information technology – Artificial intelligence – Guidance on addressing societal concerns and ethical considerations	AI		AWI
ISO IEC/ JTC1/SC42	~~[b-ISO/IEC 42105]ISO/IEC AWI 42105~~	Information technology – Artificial intelligence – Guidance for human oversight of AI systems	AI		AWI
ISO/IEC JTC 1/SC 37	~~[b-ISO/IEC 30137]ISO/IEC C 30137-1:2024~~	Use of biometrics in video surveillance systems		video surveillance	In force (Agreed 2024-03)
ISO/IEC JTC1/SC 42	~~[b-ISO/IEC 12792]ISO/IEC C-DIS 12792~~	Information technology – Artificial intelligence – Transparency taxonomy of AI systems	AI		DIS
NOTE – Category (*) includes machine learning, federated machine learning, deep learning, AI, or GenAI, etc.					

~~[b-ISO/IEC 30137] ISO/IEC 30137-1:2024~~ Use of biometrics in video surveillance systems (VSS)

It highlights the need for robust governance processes to provide assurance that the implemented security, privacy and personal data protection measures specific to the use of biometric technologies with a VSS (e.g., internationally recognizable signage) are fit for purpose, ~~and~~ and that societal considerations are reflected in the deployed system.

[b-ISO/IEC 27090] [ISO/IEC CD 27090] Addressing security threats and compromises to artificial intelligence systems~~Guidance for addressing security threats and failures in artificial intelligence systems~~

It provides guidance for organizations to address security threats and failures in AI systems and aims to provide information to organizations to help them better understand the consequences of security threats to AI systems, throughout their lifecycle, ~~and~~ and descriptions of how to detect and mitigate such threats.

[ISO/IEC WD 27091] Cybersecurity and privacy – Artificial intelligence – Privacy protection

It provides guidance for organizations to address privacy risks in AI systems and ML models and helps organizations identify privacy risks throughout the AI system lifecycle and establishes mechanisms to evaluate the consequences of and treat such risks.

[b-ISO/IEC TR 27563] [ISO/IEC In force 27563:2023] Security and privacy in artificial intelligence use cases – Best practices

It provides comprehensive best practices for evaluating security and privacy within AI applications, focusing on the use cases detailed in [ISO/IEC TR 24030]. It addresses key areas including a holistic assessment of security and privacy for AI systems, identification of concerns and risks, implementation of controls, assurance measures, ~~and~~ and the formulation of security and privacy plans, while recognizing the distinct nature of security and privacy analyses.

[b-ISO/IEC TR 5469] [ISO/IEC In force 5469:2024] Artificial intelligence – Functional safety and AI systems

It delineates the characteristics, associated risks, ~~and~~ and methodologies pertinent to the integration of AI within safety-critical functions, the employment of traditional safety functions to secure AI-operated machinery, ~~and~~ and the application of AI systems in the design and development of safety-relevant operations.

[b-ISO/IEC 23894] [ISO/IEC In force 23894:2023] Information technology – Artificial intelligence – Guidance on risk management

It offers tailored guidance for organizations to effectively manage AI-specific risks across development, production, deployment, ~~and~~ and utilization of AI-driven products, systems, ~~and~~ and services. It facilitates the integration of risk management strategies into various AI-related activities and functions, outlining processes for successful implementation and contextual customization to suit the unique needs of any organization.

[b-ISO/IEC TR 24028] [ISO/IEC In force 24028:2020] Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

It offers an in-depth survey of trustworthiness in AI systems, exploring methods to foster trust through transparency, explainability, ~~and~~ and controllability. It identifies common engineering challenges, associated threats, ~~and~~ and risks to AI systems, ~~and~~ and suggests mitigation strategies. Additionally, it discusses approaches for evaluating and ensuring various trustworthiness attributes, such as availability, resiliency, reliability, accuracy, safety, security, ~~and~~ and privacy, without defining specific trustworthiness levels for AI systems.

[b-ISO/IEC TR 24368] [ISO/IEC In force 24368:2022] Information technology – Artificial intelligence – Overview of ethical and societal concerns

It offers a comprehensive high-level perspective on the ethical and societal implications of AI, furnishing relevant information on principles, processes, ~~and~~ and methodologies for a diverse audience including technologists, regulators, ~~and~~ and the broader society. It maintains neutrality by not endorsing any particular set of values and incorporates a review of international standards that tackle the array of concerns emerging from AI's ethical and societal impacts.

~~[b-ISO/IEC 42001][ISO/IEC In force 42001:2023]~~ Information technology – Artificial intelligence – Management system

It provides a framework for organizations to systematically approach the governance of artificial intelligence, including the development, deployment, ~~and~~ and operation of AI systems, with a focus on ethical, transparent, ~~and~~ and continuous improvement practices.

~~[b-ISO/IEC TS 17847][ISO/IEC AWI 17847]~~ Information technology – Artificial intelligence – Verification and validation analysis of AI systems

It provides guidance on the verification and validation processes for AI systems, covering both the system components and their interactions with non-AI components, ~~and~~ and includes the use of formal methods, simulations, ~~and~~ and evaluations. It is suitable for organizations involved in the development, deployment, ~~and~~ and use of AI systems throughout the AI system life cycle as outlined in [ISO/IEC 22989].

~~[b-ISO/IEC TS 22440-1][ISO/IEC AWI TS 22440-1]~~ Artificial intelligence – Functional safety and AI systems – Part 1: Requirements

~~[b-ISO/IEC TS 22440-2][ISO/IEC AWI TS 22440-2]~~ Artificial intelligence – Functional safety and AI systems – Part 2: Guidance

~~[b-ISO/IEC TS 22440-3][ISO/IEC AWI TS 22440-3]~~ Artificial intelligence – Functional safety and AI systems – Part 3: Examples of application

They provide requirements and guidance on the terminology, properties, risk factors, processes, methods, techniques and architectures relating to use of AI technology within a safety-related function; use of safety-related function based on conventional technology to ensure safety of a system using AI technology; and use of AI technology to design, develop and verify safety-related functions. They include general considerations on how security threats can affect safety of an AI system.

~~[b-ISO/IEC TS 22443][ISO/IEC AWI TS 22443]~~ Information technology – Artificial intelligence – Guidance on addressing societal concerns and ethical considerations

It provides guidance on how an organization can identify and address societal concerns and ethical considerations during the life cycle of AI systems that can potentially harm individuals and society. It expands existing AI system governance, management system and impact assessment standards.

~~[b-ISO/IEC 42105][ISO/IEC AWI 42105]~~ Information technology – Artificial intelligence – Guidance for human oversight of AI systems

It provides guidance on human control and monitoring of AI systems, which is referred to as human oversight.

~~[b-ISO/IEC 12792][ISO/IEC DIS 12792]~~ Information technology – Artificial intelligence – Transparency taxonomy of AI systems

It defines a taxonomy of information elements to assist AI stakeholders with identifying and addressing the needs for transparency of AI systems. It describes the semantics of the information elements and their relevance to the various objectives of different AI stakeholders.

I.3 IEEE

Table I.4 outlines standards and ongoing work items of IEEE related to AI security.

Table I.4 – IEEE standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
C/AISC	~~[b-IEEE 2894]~~ IEEE 2894-2024	Guide for an architectural framework for explainable artificial intelligence (XAI)	AI		Active Standard (Agreed 2024-02-15)
C/AISC	~~[b-IEEE 2986]~~ IEEE 2986-2023	Recommended practice for privacy and security for federated machine learning	Federated machine learning		Active Standard (Agreed 2023-12-06)
C/AISC	~~[b-IEEE 3156]~~ IEEE 3156-2023	Standard for requirements of privacy-preserving computation integrated platform	AI	Privacy	Active Standard (Agreed 2023-12-06)
BOG/CAG	~~[b-IEEE 3169]~~ IEEE P3169	Standard for security requirement of privacy-preserving computation	AI	Privacy	Active PAR
C/S2ESC	~~[b-IEEE 7002]~~ -2022	Standard for data privacy process	AI	Privacy	Active Standard (Agreed 2022-02-09)
SSIT/SC	~~[b-IEEE P7012]~~	Standard for machine readable personal privacy terms	AI	Privacy	Active PAR
	~~[b-IEEE 2945]~~ -2023	Standard for technical requirements for face recognition	AI	Face Recognition	
	~~[b-IEEE 2884]~~ -2023	Standard for performance evaluation of biometric information: facial recognition	AI	Face Recognition	
	~~[b-IEEE 1686]~~ -2022	Standard for intelligent electronic devices cybersecurity capabilities	AI	Device	
	~~[b-IEEE P3152]~~	~~Draft~~ Standard for transparent agency identification of humans and machines			
C/S2ESC	~~[b-IEEE 610.4]~~ -1990	Standard glossary of image processing and pattern recognition terminology	Machine learning	Pattern recognition	Inactive-Withdrawn Standard

Table I.4 – IEEE standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
SRISSHWG	~~[b-IEEE P2899.1]~~	Standard for speech recognition and interaction system for smart home – engineering specifications for a listening room	AI	Speech recognition	Active PAR
RAIBS	~~[b-IEEE 3129-2023]~~	Standard for robustness testing and evaluation of artificial intelligence (AI)-based image recognition service	AI	Image recognition	Active Standard (Agreed 2023-02-15)
BITWG	~~[b-IEEE 2891-2023]~~	Standard for performance evaluation of biometric information: fingerprint recognition	AI	Fingerprint recognition	Active Standard (Agreed 2023-06-05)
RAIBS	~~[b-IEEE 3168-2024]~~	Standard for robustness evaluation test methods for a natural language processing service that uses machine learning	AI	Nature language processing	Active Standard (Agreed 2024-03-21)
ET-WG	~~[b-IEEE 7014]-2024]~~	Standard for ethical considerations in emulated empathy in autonomous and intelligent systems	AI	Ethic	Active Standard (Agreed 2024-05-20)

~~[b-IEEE 2945-2023]~~ IEEE Standard for technical requirements for face recognition

It specifies a general technical architecture and technical requirements of functional, performance, ~~and~~ ~~and~~ security for face recognition systems.

~~[b-IEEE 2884]~~ ~~[IEEE 2884-2023]~~ IEEE standard for performance evaluation of biometric information: facial recognition

It defines requirements and testing methods for a facial recognition testing system, as well as the system and security requirements of facial recognition.

~~[b-IEEE 1686-2022]~~ IEEE Standard for intelligent electronic devices cybersecurity capabilities

It defines the functions and features to be provided in intelligent electronic devices to accommodate cybersecurity programs.

[~~b-IEEE P3152~~] IEEE ~~Draft~~ Standard for transparent agency identification of humans and machines

It describes audio and visual marks to assist with the identification of the agency behind media or communications, such as a machine intelligence, a human being, or a combination of entities.

[~~b-IEEE 2894~~]~~IEEE P2894/D9~~ IEEE ~~Approved Draft~~ Guide for an Architectural Framework for Explainable Artificial Intelligence (XAI)

It provides a technological blueprint for building, deploying and managing machine learning models while meeting the requirements of transparent and trustworthy AI by adopting a variety of XAI methodologies and defines the architectural framework and application guidelines for explainable AI.

[~~b-IEEE P2986/D1.1~~] IEEE ~~Draft Recommended~~ Practice for Privacy and Security for Federated Machine Learning

It provides recommended practices related to privacy and security for FML, including security and privacy principles, defence mechanisms against non-malicious failures and examples of adversarial attacks on a FML system and defines an assessment framework to determine the effectiveness of a given defence mechanism under various settings.

[~~b-IEEE 3156~~]~~IEEE 3156-2023~~ IEEE Standard for Requirements of Privacy-Preserving Computation Integrated Platform

Requirements of privacy computation integrated platforms, including the reference architecture, the functional requirements, the performance requirements, ~~and~~ and the security requirements of privacy-preserving computation integrated platforms, are provided by this standard.

[~~b-IEEE P3169~~] IEEE Standard for security requirement of privacy-preserving computation

It describes a series of security attacks and classifies the requirements of the privacy-preserving computing system to resist security attacks. In addition, the use cases of privacy preserving computation technology schemes are specified for the application and implementation of privacy preserving computation technology.

[~~b-IEEE 7002-2022~~] IEEE Standard for data privacy process

It defines requirements for a systems/software engineering process for privacy-oriented considerations regarding products, services, ~~and~~ and systems utilizing employee, customer, or other external user's personal data.

[~~b-IEEE P7012~~] IEEE Standard for machine readable personal privacy terms

It identifies and addresses the manner in which personal privacy terms are proffered and how they can be read and agreed to by machines.

[~~b-IEEE 610.4-1990~~] IEEE standard glossary of image processing and pattern recognition terminology

It identifies terms currently in use in the field of image processing and pattern recognition.

[~~b-IEEE P2899.1~~] IEEE Standard for speech recognition and interaction system for smart home – engineering specifications for a listening room

It specifies the general technical requirements for a test site with its own background noise playback system used in the performance test of speech recognition and interaction system for smart home. And it specifies the requirements for test sites that simulate the acoustic environment of buildings.

[b-IEEE 3129-2023] IEEE Standard for robustness testing and evaluation of artificial intelligence (AI)-based image recognition service

Test specifications with a set of indicators for common corruption and adversarial attacks, which can be used to evaluate the robustness of artificial intelligence-based image recognition services are provided in this standard. Robustness attack threats and establishes an assessment framework to evaluate the robustness of artificial intelligence-based image recognition service under various settings are also specified in this standard.

[b-IEEE 2891-2023] IEEE Standard for performance evaluation of biometric information: fingerprint recognition

The components of a fingerprint test system for smart devices, such as a mobile phone are specified in this standard. Test system requirements are specified together with criteria for performance evaluation, testing methods and procedures and result reporting. Performance metrics which reflect expected device usage are specified, including false accept rate and false reject rate. Multiple testing levels are defined to support a variety of device types.

[b-IEEE 3168-2024] IEEE Standard for robustness evaluation test methods for a natural language processing (NLP) service that uses machine learning

Adversarial characters, words, or sentence samples are generated by adversarial attacks, causing the models underpinning the NLP services to produce incorrect results. A method for quantitatively evaluating the robustness the NLP services is proposed by this standard. Under the method, different cases the evaluation needs to perform against are specified. Robustness metrics and their calculation are defined. With the standard, understanding of the robustness of the services can be developed by the service stakeholders including the service developer, service providers, ~~and~~ and service users. The evaluation can be performed during various phases in the life cycle of the NLP services, the testing phase, in the validation phase, after deployment, ~~and~~ and so forth.

[b-IEEE 7014-2024] IEEE Standard for ethical considerations in emulated empathy in autonomous and intelligent systems

It provides guidance and actions for the ethical development, deployment, or decommission of autonomous and intelligent systems that attempt to emulate aspects of human empathy.

I.4 ETSI SAI

Table I.5 outlines standards and ongoing work items of ETSI related to AI security.

Table I.5 – ETSI standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
ETSI SAI	DTR/SAI-001	Global AI Security Ecosystem	AI		Early draft
ETSI SAI	DTR/SAI-002	Critical Security Controls for Effective Cyber Defense; Artificial Intelligence Sector	AI		Early draft
ETSI SAI	~~[b-ETSI DTR/SAI-003]~~ ~~DTR/SAI-003~~	~~Securing Artificial Intelligence (SAI); Collaborative Artificial Intelligence. Collaborative~~	AI		Publication (2024-01)

Table I.5 – ETSI standards and work items on AI security

Study group	Reference	Title	Category (*)	verticals	Status
		[TSB-VB11] Artificial Intelligence			
ETSI SAI	[b-ETSI DTR/SAI-004]DTR/SAI-004	Securing Artificial Intelligence (SAI); Traceability of AI Models.Traceability of AI Models	AI		Publication (2024-02)
ETSI SAI	DTS/SAI-005	AI Threat Ontology and definitions	AI		Early draft
ETSI SAI	DTS/SAI-006	Artificial Intelligence Computing Platform Security Framework	AI		Final draft for approval
ETSI SAI	DTS/SAI-008	Problem Statement	AI		Early draft
ETSI SAI	DTR/SAI-009	Mitigation Strategy Report	AI		Stable draft
ETSI SAI	DTR/SAI-0011	Security aspects of using AI/ML techniques in telecom sector	AI, Machine learning		Early draft
ETSI SAI	DTR/SAI-0014	Baseline Cyber Security Requirements for AI Models and Systems	AI		TB adoption of WI
ETSI SAI	DTR/SAI-0015	Security Testing of AI	AI		Early draft
ETSI SAI	DTR/SAI-0016	Explicability and transparency of AI processing	AI		Early draft
ETSI SAI	DTR/SAI-0017	Proofs of Concepts Framework	AI		Publication (2024-04-05)
ETSI SAI	DTR/SAI-0018	Privacy aspects of AI/ML systems	Machine learning		Publication (2024-04-17) [TSB-VB12]

[DTR/SAI-003] Collaborative artificial intelligence

It describes collaborative AI from securing AI perspectives. Collaborative AI could take place among AI agents, between AI agents and human, ~~and~~ ~~and~~ even among people who provide and use AI. As such, the security and performance of collaborative AI may range from AI/ML-specific issues to other system specific issues (e.g., AI-to-AI communications, joint computing and communicating optimization). It investigates collaborative AI use cases and involved technical aspects and analyses potential security and performance issues (e.g., AI-to-AI communications, trustworthy collaboration,

etc.) among those AI-related entities. It also overviews existing approaches to tackle and/or mitigate these issues.

[DTR/SAI-004] Traceability of AI models

It describes the role of traceability in the challenge of securing AI and explores issues related to sharing and re-using models across tasks and industries. The scope includes threats, ~~and~~ and their associated remediation where applicable, to ownership rights of AI creators as well as to verification of model's origin. Mitigation can be non-AI-specific (digital right management applicable to AI) and AI-specific techniques (e.g., ML watermarking) from prevention and detection phases. They can be both model-agnostic and model enhancement techniques. It aligns terminology with existing ETSI ISG SAI documents and studies, ~~and~~ and references/complements previously studied attacks and remediation. It also gathers industrial and academic feedback on traceability and ownership rights protection and model verification in the context of AI.

[DTS/SAI-005] AI threat ontology and definitions

It defines what an AI threat is and defines how it can be distinguished from any non-AI threat. The model of an AI threat is presented in the form of an ontology to give a view of the relationships between actors representing threats, threat agents, assets and so forth. The ontology in the present document extends from the base taxonomy of threats and threat agents described in ETSI TS 102 165-1 [i.5] and addresses the overall problem statement for SAI presented in ETSI GR SAI 004 [i.6] and the mitigation strategies described in ETSI GR SAI 005 [i.7].

[DTS/SAI-006] Artificial intelligence computing platform security framework

It first studies the role of AI computing platform in AI systems, its common structure and security requirements in AI systems. Secondly, the security components and interaction between these components that form the security framework of AI computing platform are studied. Last but not the least, the mechanisms which guarantee the security of the platform itself and provide services for relevant stakeholders in AI systems are studied in detail. The aim of the present document is to set out a reference security framework for AI computing platform developer and users to mitigate the security threats against AI systems in a cooperatively manner.

[DTS/SAI-008] Problem statement

It describes the problem of securing AI-based systems and solutions, with a focus on machine learning, ~~and~~ and the challenges relating to confidentiality, integrity and availability at each stage of the machine learning lifecycle. It also describes some of the broader challenges of AI systems including bias, ethics and explainability. A number of different attack vectors are described, as well as several real-world use cases and attacks.

[DTR/SAI-009] Mitigation strategy report

It summarizes and analyses existing and potential mitigation against threats for AI-based systems as discussed in ETSI GR SAI 004 [i.1]. The goal is to have a technical survey for mitigating against threats introduced by adopting AI into systems. The technical survey shed light on available methods of securing AI-based systems by mitigating against known or potential security threats. It also addresses security capabilities, challenges, ~~and~~ and limitations when adopting mitigation for AI-based systems in certain potential use cases.

[DTR/SAI-0015] Security testing of AI

It identifies methods and techniques that are appropriate for security testing of ML-based components. Security testing of AI does not end at the component level. As for testing of traditional software, the integration with other components of a system needs to be tested as well.

[DTR/SAI-0016] Explicability and transparency of AI processing

It identifies steps to be taken by designers and implementers of AI platforms that give assurance of the explicability and transparency of AI processing. AI processing includes AI decision making and AI data processing. It also identifies its target audience as designers and implementers who are making assurances to a lay person.

[DTR/SAI-0017] Proofs of concepts framework

It provides information about the "lightweight" framework to be used by ETSI ISG SAI to create multi-partner proofs of concepts (PoCs).

[DTR/SAI-0018] Privacy aspects of AI/ML systems

It identifies the role of privacy as one of the components of the security of AI, ~~and~~ and defines measures to protect and preserve privacy in the context of AI that covers both, safeguarding models and protecting data, as well as the role of privacy-sensitive data in AI solutions. It documents and addresses the attacks and their associated remediation where applicable, considering the existence of multiple levels of trust affecting the lifecycle of data.

Bibliography

- [b-ITU-T F.748.24] Recommendation ITU-T F.748.24 (2024), *Trusted contribution evaluation framework on federated machine learning services.*
- [b-ITU-T F.748.35] Recommendation ITU-T F.748.35 (2024), *Requirement and framework of trustworthy federated machine learning based service.*
- [b-ITU-T F.781.2] Recommendation ITU-T F.781.2 (2024), *Quality assessment requirements for artificial intelligence/machine learning-based software as a medical device.*
- [b-ITU-T X.1231] Recommendation ITU-T X.1231 (2008), *Technical strategies for countering spam.*
- [b-ITU-T X.1252] Recommendation ITU-T X.1252 (2021), *Baseline identity management terms and definitions.*
- [b-ITU-T Y.Suppl.72] Supplement ITU-T Y.Suppl.72 (2022), *Y.3000-series – Artificial intelligence standardization roadmap.*
- ~~[b-IEEE P3652.1] IEEE Standard P3652.1 (2020), *Guide for Architectural Framework and Application of Federated Machine Learning.*~~
- ~~[b-ISO 31000] ISO 31000:2018, *Risk management – Guidelines.*~~
- ~~[b-ISO 31073] ISO 31073:2022, *Risk management – Vocabulary.*~~
- ~~[b-ISO/IEC 22989] ISO/IEC 22989:2022, *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology.*~~
- ~~[b-ISO/IEC TR 24028] ISO/IEC TR 24028:2020, *Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence.*~~
- [b-ISO/IEC/IEEE 24641] ISO/IEC/IEEE 24641:2023, *Systems and Software engineering – Methods and tools for model-based systems and software engineering.* [b-ETSI DTR/SAI-001] DTR/SAI-001 (ETSI TR 104 029), *Securing Artificial Intelligence (SAI); Global AI Security Ecosystem, under development (Early draft).*
- [b-ETSI DTR/SAI-002] DTR/SAI-002 (ETSI TR 104 030), *Securing Artificial Intelligence (SAI); Critical Security Controls for Effective Cyber Defense; Artificial Intelligence Sector, under development (Early draft).*
- [b-ETSI DTR/SAI-003] ETSI TR 104 031 (2024), *Securing Artificial Intelligence (SAI); Collaborative Artificial Intelligence.*
- [b-ETSI DTR/SAI-004] ETSI TR 104 032 (2024), *Securing Artificial Intelligence (SAI); Traceability of AI Models.*
- [b-ETSI DTS/SAI-005] ETSI TS 104 050, *Securing Artificial Intelligence (SAI); AI Threat Ontology and definitions, under development (Early draft).*
- [b-ETSI DTS/SAI-006] DTS/SAI-006 (ETSI TS 104 033), *Securing Artificial Intelligence (SAI); Artificial Intelligence Computing Platform Security Framework, under development (Final draft for approval).*
- [b-ETSI DTS/SAI-008] DTS/SAI-008 (ETSI TR 104 221), *Securing Artificial Intelligence (SAI); Problem Statement, under development (Early draft).*

<u>[b-IEEE 2894]</u>	<u>IEEE 2894 (2024), IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence.</u>
<u>[b-IEEE 2986]</u>	<u>IEEE 2986 (2023), IEEE Recommended Practice for Privacy and Security for Federated Machine Learning.</u>
<u>[b-IEEE 3156]</u>	<u>IEEE 3156 (2023), IEEE Standard for Requirements of Privacy-Preserving Computation Integrated Platforms.</u>
<u>[b-IEEE 3169]</u>	<u>IEEE 3169 (2025), Standard for Security Requirement of Privacy-Preserving Computation.</u>
<u>[b-IEEE 3652.1]</u>	<u>IEEE Standard 3652.1 (2020), Guide for Architectural Framework and Application of Federated Machine Learning.</u>
<u>[b-IEEE 7002]</u>	<u>IEEE 7002 (2022), IEEE Standard for Data Privacy Process.</u>
<u>[b-IEEE 7012]</u>	<u>IEEE 7012 (2026), Standard for Machine Readable Personal Privacy Terms.</u>
<u>[b-IEEE 2945]</u>	<u>IEEE 2945 (2023), IEEE Standard for Technical Requirements for Face Recognition.</u>
<u>[b-IEEE 2884]</u>	<u>IEEE 2884 (2023), IEEE Standard for Performance Evaluation of Biometric Information: Facial Recognition.</u>
<u>[b-IEEE 1686]</u>	<u>IEEE 1686 (2022), IEEE Standard for Intelligent Electronic Devices Cybersecurity Capabilities.</u>
<u>[b-IEEE 3152]</u>	<u>IEEE 3152 (2025), Standard for Transparent Human and Machine Agency Identification.</u>
<u>[b-IEEE 610.4]</u>	<u>IEEE 610.4 (1990), IEEE Standard Glossary of Image Processing and Pattern Recognition Terminology.</u>
<u>[b-IEEE 2899.1]</u>	<u>IEEE 2899.1 (2025), Standard for Speech Recognition and Interaction System for Smart Home – Engineering Specifications for a Listening Room.</u>
<u>[b-IEEE 3129]</u>	<u>IEEE 3129 (2023), IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service.</u>
<u>[b-IEEE 2891]</u>	<u>IEEE 2891 (2023), IEEE Standard for Performance Evaluation of Biometric Information: Fingerprint Recognition.</u>
<u>[b-IEEE 3168]</u>	<u>IEEE 3168 (2024), IEEE Standard for Robustness Evaluation Test Methods for a Natural Language Processing Service That Uses Machine Learning.</u>
<u>[b-IEEE 7014]</u>	<u>IEEE 7014 (2024), IEEE Standard for Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems.</u>
<u>[b-ISO 31000]</u>	<u>ISO 31000:2018, Risk management – Guidelines.</u>
<u>[b-ISO 31073]</u>	<u>ISO 31073:2022, Risk management – Vocabulary.</u>
<u>[b-ISO/IEC 22989]</u>	<u>ISO/IEC 22989:2022, Information technology – Artificial intelligence – Artificial intelligence concepts and terminology.</u>
<u>[b-ISO/IEC 27090]</u>	<u>ISO/IEC 27090, Cybersecurity – Artificial intelligence – Addressing security threats and compromises to artificial intelligence systems.</u>

[b-ISO/IEC 27091]	ISO/IEC 27091, <i>Cybersecurity and privacy – Artificial intelligence – Privacy protection.</i>
[b-ISO/IEC TR 27563]	ISO/IEC TR 27563:2023, <i>Security and privacy in artificial intelligence use cases – Best practices.</i>
[b-ISO/IEC TR 5469]	ISO/IEC TR 5469:2024, <i>Artificial intelligence – Functional safety and AI systems.</i>
[b-ISO/IEC 23894]	ISO/IEC 23894:2023, <i>Information technology – Artificial intelligence – Guidance on risk management.</i>
[b-ISO/IEC TR 24028]	ISO/IEC TR 24028:2020, <i>Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence.</i>
[b-ISO/IEC TR 24368]	ISO/IEC TR 24368:2022, <i>Information technology – Artificial intelligence – Overview of ethical and societal concerns.</i>
[b-ISO/IEC 42001]	ISO/IEC 42001:2023, <i>Information technology – Artificial intelligence – Management system.</i>
[b-ISO/IEC TS 17847]	ISO/IEC TS 17847, <i>Information technology – Artificial intelligence – Verification and validation analysis of AI systems.</i>
[b-ISO/IEC TS 22440-1]	ISO/IEC TS 22440-1, <i>Artificial intelligence – Functional safety and AI systems – Part 1: Requirements.</i>
[b-ISO/IEC TS 22440-2]	ISO/IEC TS 22440-2, <i>Artificial intelligence – Functional safety and AI systems – Part 2: Guidance.</i>
[b-ISO/IEC TS 22440-3]	ISO/IEC TS 22440-3, <i>Artificial intelligence – Functional safety and AI systems – Part 3: Examples of application.</i>
[b-ISO/IEC TS 22443]	ISO/IEC TS 22443, <i>Information technology – Artificial intelligence – Guidance on addressing societal concerns and ethical considerations.</i>
[b-ISO/IEC 42105]	ISO/IEC 42105, <i>Information technology – Artificial intelligence – Guidance for human oversight of AI systems.</i>
[b-ISO/IEC 30137]	ISO/IEC 30137-1:2024, <i>Information technology – Use of biometrics in video surveillance systems – Part 1: System design and specification.</i>
[b-ISO/IEC 12792]	ISO/IEC 12792, <i>Information technology – Artificial intelligence – Transparency taxonomy of AI systems.</i>
[b-ISO/IEC/IEEE 24641]	ISO/IEC/IEEE 24641:2023, <i>Systems and Software engineering – Methods and tools for model-based systems and software engineering.</i>