ETSI GR SAI 009 V1.1.1 (2023-02)

Securing Artificial Intelligence (SAI); Artificial Intelligence Computing Platform Security Framework

Disclaimer

The present document has been produced and approved by the Securing Artificial Intelligence (SAI) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/SAI-009

Keywords
artificial intelligence, security

ETSI

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

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

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

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

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

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Securing Artificial Intelligence (SAI).

Modal verbs terminology

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

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

Introduction

Artificial Intelligence is evolving very fast. It has been deployed everywhere trying to improve the quality of daily life. The compromise of AI systems will directly affect a vast number of people and can cause severe results. In an AI system, AI computing platform acts as the infrastructure for AI applications that provides resource and executing environments. Therefore, it is very important to study the security of AI computing platform. The reasons are two-folds:

- The security of AI computing platform is the baseline guarantee for AI system. If the platform is compromised, all of the applications running on it will be controlled by malicious attackers.
- Common security requirements from different AI applications are best resolved by AI computing platform that will apparently improve the protection levels for AI application and bring convenience for relevant stakeholders during development, use and maintenance of AI applications.

The present document 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.

The study uses ETSI GR SAI 006 [i.3] as a starting point for hardware aspects and avoids overlap.

1 Scope

The present document describes a security framework of AI computing platform containing hardware and basic software to protect valuable assets like models and data deployed on AI computing platform when they are used in runtime or stored at rest. The security framework consists of security components in AI computing platform and security mechanisms executed by security components in the platform. By specifying the security framework, an AI computing platform can be consolidated against the relevant attack and can provide security capabilities to facilitate the stakeholders in AI systems to better protect the valuable assets (model/data) on an AI computing platform.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

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

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

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

[i.1] ETSI GR SAI 004: "Securing Artificial Intelligence (SAI); Problem Statement".

[i.2] ETSI GR SAI 005: "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".

[i.3] ETSI GR SAI 006: "Securing Artificial Intelligence (SAI); The role of hardware in security of AI".

[i.4] Pingchuan Ma, Stavros Petridis, Maja Pantic: "Detecting adversarial attack on audiovisual speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, Toronto, ON, Canada.

NOTE: Available at https://arxiv.org/pdf/1912.08639.pdf.

[i.5] Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan and Edward Mcfowland: "Detecting Adversarial Attacks via Subset Scanning of Autoencoder Activations and Reconstruction Error", Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence.

NOTE: Available at https://www.ijcai.org/proceedings/2020/0122.pdf.

[i.6] Mika Juuti, Sebastian Szyller, Samuel Marchal, N. Asokan: "PRADA: Protecting Against DNN Model Stealing Attacks", 2019 IEEE European Symposium on Security and Privacy (EuroS&P).

[i.7] Ren Pang, Xinyang Zhang, shouling Ji, Xiapu Luo, Ting Wang: "AdvMind: Inferring Adversary Intent of Black-Box Attacks", Proceedings of the 26th ACM SIGKDD, 2020.

NOTE: Available at https://arxiv.org/pdf/2006.09539.pdf.

[i.8] ETSI GS NFV-SEC 009: "Network Functions Virtualisation (NFV); NFV Security; Report on use cases and technical approaches for multi-layer host administration".

[i.9] ETSI GS NFV-SEC 012: "Network Functions Virtualisation (NFV) Release 3; Security; System architecture specification for execution of sensitive NFV components".

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

AI asset: anything that has value to the organization, its business operation and its continuity relevant to AI

NOTE: Model, dataset and training script belongs to AI asset category.

3.2 Symbols

Void.

3.3 Abbreviations

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

AE	AutoEncoder
AI	Artificial Intelligence
API	Application Programming Interface
BIOS	Basic Input Output System
BMC	Baseboard management controller
CPU	Central Processing Unit
CVE	Common Vulnerabilities & Exposures
DDoS	Distributed Denial of Service
DNS	Domain Name System
DoS	Denial of Service
GPU	Graphics Processing Unit
HBRT	Hardware Based Root of Trust
HDD	Hard Disk Drive
HMEE	Hardware-Mediated Execution Enclave
HSM	Hardware Security Module
HTTPS	Hypertext Transfer Protocol Secure
HUK	Hardware Unique Key
IO	Input Output
IOT	Internet Of Things
JTAG	Joint Test Action Group
NIC	Network interface card
NPU	Neural network Processing Unit
OS	Operating System
PCI	Peripheral Component Interconnect
PCIe	Peripheral Component Interconnect express
REE	Rich Execution Environment
RoT	Root of Trust
RTS	Root of Trust for Storage
RTR	Root of Trust for Report
RPO	Recovery Point Objective
RTO	Recovery Time Objective
SDK	Software Development Kit
SE	Secure Element
SIM	Subscriber Identity Module
SMTP	Simple Mail Transfer Protocol
SoC	System on Chip
SSD	Solid State Disk
TEE	Trusted Execution Environment
TLS	Transport Layer Security
TPM	Trusted Platform Module

USB	Universal Serial Bus
VM	Virtual Machine

4 Convention Description

4.1 Notation

For the purpose of the present document, the following notations apply in clause 9:

- [Security component] stands for the "security components" described in clause 7 that are involved in the interactive procedures.
- <Service-based interface> stands for the "Service-based interface" described in clause 8.2 that is utilized to deliver relevant information or data between an AI computing platform and platform users.

5 Overview

5.1 AI computing platform description

An AI computing platform is defined as a fundamental platform to provide an execution environment and related resources in an AI system as shown in Figure 1. In particular, the platform provides required resource for AI model and AI application including computing resource, storage resource and networking resource; meanwhile it also provides an execution environment for AI models and AI applications to be able to operate properly and stably including a fundamental software framework, various kinds of libraries and different hardware drivers. On the other hand, for important assets like training dataset, testing dataset, inference data and training script, an AI computing platform also provides storage resource to guarantee the secure storage of these assets.

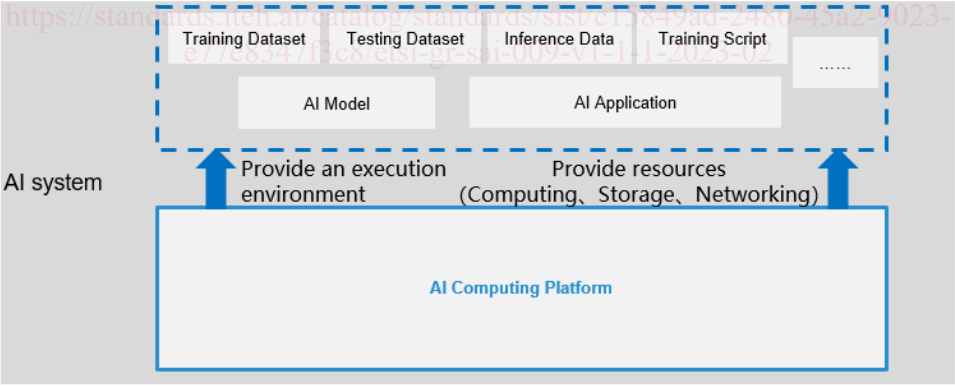

Figure 1: Role of AI computing platform in an AI system

The structure of an AI computing platform is constructed in three layers and illustrated in Figure 2. The hardware layer and basic software layer form the core/basic part of an AI computing platform while the application enabling layer is the extended/optional part as the components in the application enabling layer vary depending on different application scenarios and most components in this layer could be independent of underlying layers.

To be specific, the first layer contains different kinds of hardware to provide computing resource, storage resource and networking resource. Computing hardware includes general computing device (CPU) and AI-specific computing device like GPU and NPU. Storage hardware includes memory and non-volatile storage like hard disk drive and solid state disk. Networking hardware includes network interface card like smart IO card or network device connected by PCIe. The second layer is the basic software layer consisting of operating system, chip-level SDK, virtualization component for VM or containers, and some other software which directly manage the hardware devices and act as an interface to open up the hardware layer capabilities to AI application providers and users. The software in this layer are strongly related with hardware like OS, GPU/NPU accelerating libraries and GPU/NPU operating framework. The top layer in AI computing platform is the application enabling layer which contains different types of deep learning framework, application-level SDK, management module and so on. Herein, application-level SDK is a set of software components which has limited or no relation with hardware to enable the development, deployment and management of AI application; the management module provides management function for AI computing platform and deals with the cooperation in some application scenarios if needed like cluster computing, cloud computing, cloud-edge collaboration, etc.

EXAMPLE 1: Hypervisor and container engine belong to the category of virtualization software in basic software layer.

EXAMPLE 2: Tensorflow, Pytorch and Mindspore belong to the category of deep learning framework.

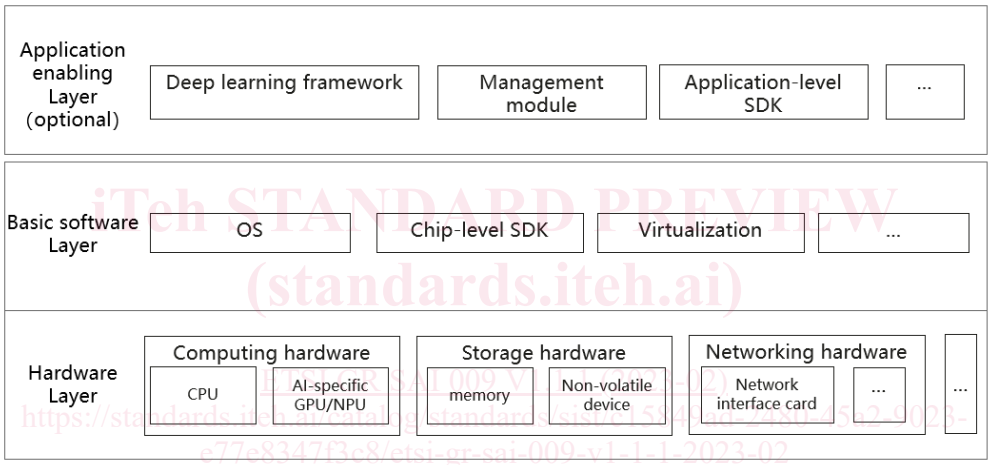

Figure 2: Structure of AI computing platform

A typical implementation of AI computing platform in common scenarios is an AI server equipped with GPU/NPU, server-grade OS and GPU/NPU related chip-level SDK.

EXAMPLE 3: GPU/NPU and the related chip-level SDKs are available on the market.

5.2 Implementation recommendations

The present document focuses on AI computing platforms that are deployed in data centres or edge computing environments. AI computing platforms in these environments have more resources (computing power and energy) for computing tasks, and for constructing security capabilities as well. In addition, these environments can involve more complex scenarios like ones supporting multi-tenant settings and remote access usage which introduce specific threats and may need corresponding mitigations. Other types of platforms like mobile phones or embedded devices capable of executing AI functionality may also refer to this security framework according to their specific conditions and security requirements.

5.3 Threats analysis of AI computing platform

As introduced in clause 5.1, an AI computing platform provides an execution environment and related resources to AI applications and AI models. Therefore, to comprehensively summarize the security requirements for an AI computing platform, a threats analysis is conducted in two aspects:

- The threats against the AI computing platform itself.
- The threats against valuable assets on the platform.

NOTE: It is very important to emphasize that the threats in Tables 1 and 2 are not exhaustive . New threats can emerge based on specific scenarios and technology development.

The first aspect of threats can be categorized and analysed as follows in Table 1. They are partitioned into several high level categories concentrated on the AI computing platform as the affected asset. Then, a detailed description is presented in each category. On the other hand, threats listed in Table 1 will directly or indirectly affect AI assets after the AI computing platform is being attacked. Direct affect can be caused by the threat directly affecting specific AI-related assets after the AI computing platform being attacked. Indirect affect is produced by the threats indirectly affecting specific AI-related assets after the AI computing platform being attacked. However, this kind of threats can create the available condition or execution environment for threat agent to implement further attacks directly on AI-related assets on the platform. The fourth column marks this information.

Table 1: Threats analysis for the AI computing platform itself

Category	No.	Threat description	Risk and negative consequence	Threat to AI assets
Physical destruction	1	The threat agent utilizes physical method to destroy the AI computing platform. Physical method can include fire, water, physical damage, etc.	AI computing platform can be destroyed and stop working. Valuable assets like business data/model/data set can be corrupted.	Direct on availability
Outsider manipulation	2	The threat agent utilizes a physical interface of the AI computing platform like USB and JTAG to implement unauthorized operation and access to the platform.	Configuration data and sensitive information can be leaked and the control permission of the platform can be gained by the threat agent to implement further attack.	Indirect
	3	The threat agent utilizes a remote interface of the AI computing platform like service interface and remote login to implement unauthorized operation and access to the platform.	Configuration data and sensitive information can be leaked and the control permission of the platform can be obtained by the threat agent to implement further attack.	Indirect
	4	The threat agent exploits known vulnerabilities of hardware and software components in the AI computing platform.	Sensitive information can be leaked. Or the control permission of the AI computing platform can be obtained by the threat agent for further attack.	Indirect
Insider manipulation	5	The threat agent utilizes a legitimate account or a legitimate asset like normal OS user account, virtual machine instance on AI platform to conduct privilege escalation to acquire more permissions than allowed.	Sensitive information can be leaked and the control permission of the platform can be obtained by the threat agent to implement further attack.	Indirect
	6	The threat agent utilizes legitimate resource/object like virtual machine, container and even PCI hardware device which it has operation permission to conduct lateral attack to other resource which belongs to other users.	Sensitive information and valuable assets of users who share resource of the AI computing platform can be leaked, tampered or stolen by malicious users on the same platform.	Indirect

Category	No.	Threat description	Risk and negative consequence	Threat to AI assets
Tampering	7	The threat agent maliciously tampers with the system clock which is maintained in the AI computing platform.	Some key subsystem running on the platform can be disrupted if the time clock is disordered. Log consequence and detail information can be disrupted so that the threat agent can utilize this disorder to avoid security audit or hide any malicious action on the platform.	Indirect
	8	The threat agent maliciously tampers with the components in the AI computing platform like firmware of devices and software in the platform.	The functionality of firmware can be modified so that the system fails. Or some malicious programs can be inserted into firmware, based on which threat agent can further control the system or steal the valuable information in the platform.	Indirect
	9	The threat agent deploys a malicious software onto the platform.	The performance of the platform can be affected like resource exhaustion.	Indirect
Data attack	10	The threat agent steals the storage device like HDD or SSD containing valuable assets if he has the permission to physically contact with AI computing platform.	Valuable assets like business data/model/data set or system configuration information can be leaked or stolen.	Direct on confidentiality
	11	The threat agent corrupts the storage of the platform by tampering or deleting the data on it.	Valuable assets like business data/model/data set or system configuration information can be corrupted or destroyed.	Direct on availability or integrity
Communication attack	12	The threat agent conducts DoS or DDoS attack to the exposed interface of the AI computing platform.	Communication availability and service accessibility of AI computing platform and the corresponding AI system it supports can be degraded or even destroyed.	Direct on availability
	13	The threat agent utilizes exposed interface to conduct an attack like injection attack and buffer overflow attack.	Sensitive information can be leaked and the control permission of the platform can be obtained by the threat agent to implement further attack.	Indirect
	14	The threat agent conducts an eavesdropping attack on the communication to and from the AI computing platform.	Sensitive information transferred to and from AI computing platform can be leaked.	Direct on confidentiality
	15	The threat agent conducts a spoofing attack in the procedure of communication to and from the AI computing platform.	Sensitive information transferred to and from the AI computing platform can be leaked and tampered by the threat agent.	Direct on confidentiality and integrity
Occupation	16	The threat agent gains complete control of the computing platform through specific physical or social engineering methods.	The platform can be fully controlled, all of the assets stored on the platform can be leaked and tampered. The input and output of the process running on the platform can be arbitrarily changed.	Direct on confidentiality, integrity and availability.

On the other hand, an AI computing platform should provide sufficient functionality to protect the assets running or stored on the platform from known threats. Threat analysis for assets on the AI computing platform is important when summarizing the security requirements on an AI computing platform. The threats against valuable assets on an AI computing platform can be categorized and analysed as follows in Table 2.

Table 2: Threats analysis for valuable assets on AI computing platform

Asset	No.	Threat description	Risk and negative consequence	Security goal affected
Model	1	The threat agent accesses the storage where the model is stored and steals the key structure information and parameter value of model in the model file.	The intellectual property of the model provider is stolen which can be a huge loss as a well-performed model costs massive resources including computing resource, dataset resource, time resource and labour resource.	Confidentiality
Model	2	The threat agent accesses the memory where the model is loaded and steals the key structure information and parameter value of the model in running.	The intellectual property of model provider is stolen which can be a huge loss as a well-performed model costs massive resources including computing resource, dataset resource, time resource and labour resource.	Confidentiality
Model	3	The threat agent accesses the storage where the model is stored and unauthorizedly changes the content of the model file like structure information and parameter value.	The original model can be changed in the aspect of the performance and therefore the changes influence the outcomes of the AI application/system which utilizes this model. Worse results can be caused when such AI applications/systems are used in safety-critical scenarios like automobile and e-health.	Integrity
Model	4	The threat agent accesses the memory where the model is loaded and unauthorizedly changes the parameters of the model file like structure information and parameter value.	The original model can be changed in the aspect of the performance and therefore the changes influence the outcomes of the AI application/system which utilizes this model. Worse results can be caused when such AI applications/systems are used in safety-critical scenarios like automobile and e-health.	Integrity
Model	5	The model is not deleted/ erased from storage correctly and thoroughly. The threat agent recovers the model from storage.	The intellectual property of the model provider is stolen which can be a huge loss as a well-performed model costs massive resources including computing resource, dataset resource, time resource and labour resource.	Confidentiality
Model	6	The threat agent forges a well-designed fake model and inserts it into the AI application/system.	The outcome of the AI application/system can be misled as the threat agent wishes. Worse results can be caused when such AI applications/systems are used in safety-critical scenarios like automobile and e-health.	Integrity
Model	7	The threat agent accesses the storage where the model is stored and deletes the model file.	For model users, the business which relies on the model can be terminated. While for model providers, it can be seen as a huge business loss since resources for the training process of the model are wasted if the model is deleted.	Availability
Model	8	The threat agent destroys the model by physical methods to the storage media of the model file.	For model users, the business which relies on the model can be terminated. While for model providers, it can be seen as a huge business loss since resources for the training process of the model are wasted if the model is deleted or destroyed.	Availability