ETSI TS 104 050 V1.1.1 (2025-03)

Securing Artificial Intelligence (SAI);
AI Threat Ontology and definitions

Reference
RTS/SAI-005

Keywords
artificial intelligence

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 IPR online database.

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

Trademarks

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

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

Foreword

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

NOTE: The present document updates and extends ETSI GR SAI 001 [i.20] prepared by ISG SAI.

Modal verbs terminology

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

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

1 Scope

The present document defines what an Artificial Intelligence (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 and defines those terms (see also [1]). The ontology in the present document extends from the base taxonomy of threats and threat agents described in ETSI TS 102 165-1 [2] and addresses the overall problem statement for SAI presented in ETSI TR 104 221 [i.21] and the mitigation strategies described in ETSI TR 104 222 [i.22]. Note that, although both technical reports are listed in clause 2.2, they are indeed essential for understanding the scope of the present document.

NOTE 1: The ontology described in the present document applies to AI both as a threat agent and as an attack target.

NOTE 2: The present document extends the content of ETSI GR SAI 001 [i.20], and retains significant elements of its content where relevant for clarity.

2 References

2.1 Normative references

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

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

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

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

[1] ISO/IEC 22989:2022: "Information technology - Artificial intelligence - Artificial intelligence concepts and terminology".

NOTE: Many of the terms defined in the cited document above are also visible on the ISO Online Browsing Platform: https://www.iso.org/obp/.

[2] ETSI TS 102 165-1: "Cyber Security (CYBER); Methods and protocols; Part 1: Method and pro forma for Threat, Vulnerability, Risk Analysis (TVRA)".

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] Alan Turing: "On computable numbers, with an application to the Entscheidungsproblem".

[i.2] Alan Turing: "Computing Machinery and Intelligence".

[i.3] Philip K. Dick: "Do androids dream of electric sheep?" (ISBN-13: 978-0575094185).

[i.4] Isaac Asimov: "I, robot" (ISBN-13: 978-0008279554).

[i.5] W3C® Recommendation 11 December 2012: "OWL: OWL 2 Web Ontology Language Document Overview (Second Edition)".

[i.6] RDF: RDF 1.1 Primer; W3C® Working Group Note; 24 June 2014.

[i.7] Cohen, Jacob (1960): "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37-46. doi:10.1177/001316446002000104. hdl:1942/28116. S2CID 15926286.

[i.8] W3C® Recommendation 16 July 2020: "JSON-LD 1.1: A JSON-based Serialization for Linked Data".

[i.9] ETSI GS CIM 009 (V1.2.2): "Context Information Management (CIM); NGSI-LD API".

[i.10] "The Emergence Of Offensive AI".

[i.11] "Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter".

[i.12] Li Chen, Chih-Yuan Yang, Anindya Paul, Ravi Sahita: "Towards resilient machine learning for ransomware detection".

[i.13] Alejandro Correa Bahnsen, Ivan Torroledo, Luis David Camacho and Sergio Villegas: "DeepPhish: Simulating Malicious AI".

[i.14] Common Weakness Enumeration Project.

[i.15] ETSI TS 118 112: "oneM2M; Base Ontology".

[i.16] The Smart Appliances REFerence (SAREF) ontology.

[i.17] ETSI TS 102 165-2: "CYBER; Methods and protocols; Part 2: Protocol Framework Definition; Security Counter Measures".

[i.18] ETSI TR 104 048: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

[i.19] Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar: "Threat Modeling AI/ML Systems and Dependencies".

[i.20] ETSI GR SAI 001: "Securing Artificial Intelligence (SAI); AI Threat Ontology".

[i.21] ETSI TR 104 221: "Securing Artificial Intelligence (SAI); Problem Statement".

[i.22] ETSI TR 104 222: "Securing Artificial Intelligence; Mitigation Strategy Report".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in ETSI TR 104 221 [i.21], ISO/IEC 22989 [1] and the following apply:

Artificial General Intelligence (AGI): applying intelligence to any intellectual task, at a level equivalent to a human

NOTE: AGI is also termed Strong AI.

Artificial Intelligence (AI): ability of a system to handle representations, both explicit and implicit, and procedures to perform tasks that would be considered intelligent if performed by a human

NOTE: From ETSI TR 104 221 [i.21].

Artificial Narrow Intelligence (ANI): applying intelligence to only one context

EXAMPLE: Autonomous driving, speech recognition.

NOTE: ANI is also termed Weak AI.

Artificial Super Intelligence (ASI): extending beyond AGI to apply intelligence to a level significantly beyond those of humans across a comprehensive range of categories and fields of endeavour

cognition: mental action or process of acquiring knowledge and understanding through thought, experience, and the senses

predicate: part of a sentence or clause containing a verb and stating something about the subject

NOTE: In the context of the present document as applied to RDF statements the predicate illustrates the nature of the relationship between two objects or concepts.

reasoning: application of learned strategies in order to solve puzzles, and make judgments where there is uncertainty in either the input or the expected outcome

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in ETSI TR 104 221 [i.21] and the following apply:

AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
CAV	Connected and Autonomous Vehicles
CIA	Confidentiality Integrity Availability
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
CWE	Common Weakness Enumeration
GAN	Generative Adversarial Networks
ICT	Information Communications Technology
IQ	Intelligence Quotient
IT	Information Technology
ITS	Intelligent Transport Systems
JSON	JavaScript Object Notation
LD	Linked Data
ML	Machine Learning
NGSI	Next Generation
NGSI-LD	Next Generation Service Interface - Linked Data
OWL	Ontology Web Language
RDF	Resource Description Framework
RL	Reinforcement Learning
SAI	Securing Artificial Intelligence
TVRA	Threat Vulnerability Risk Analysis
UML	Unified Modelling Language
XSS	Cross Site Scripting

4 From taxonomy to an ontology for secure AI

4.1 Overview

An ontology in information science identifies a set of concepts and categories within a particular field of knowledge that shows the properties of the concepts and categories and the relations between them.

This overview illustrates and demonstrates how the various concepts that are taken for granted in the security standards space are implicit as taxonomies. The overview extends to illustrate that by adopting a broader understanding of these implicit taxonomies in the form of an ontology, in which concepts are related, will help in making systems more resilient against AI attackers, or which make better use of AI in defence.

NOTE 1: The model of ontology from philosophy is the study of being, and addresses concepts such as becoming, existence and reality. For many, the ultimate aim of AI is general intelligence i.e. the ability of a single machine agent able to learn or understand any task, covering the range of human cognition. If and when AI moves closer to any concept of independent sentience, there will be increasing overlap between the worlds of information science and philosophy. However, this is likely to be decades away at least, and so the present document focusses on so-called weak AI: the use of software to perform specific, pre-defined, reasoning tasks. Also, in the philosophical domain there is a degree of crossover in the role of intelligence and the role of ethics. The present document does not attempt to define the role of ethics other than to reflect that in an ontology of intelligence that there are various schools of ethics that apply. So, an intelligence framework is influenced by its ethical framework, where the impact of the ethical framework can be realized in various ways.

In many domains that apply some form of AI, the core data model is presented in an ontological form and from that it is possible to apply more sophisticated search algorithms to allow for semantic reasoning. The technical presentation of an ontology is therefore significant of itself as it can pre-determine the way in which the programming logic is able to express intelligence. Ontologies, in the context of a semantic web, are often designed for re-use. In addition to conventional ontologies and the use of Resource Description Framework (RDF) [i.6] notations, there is growth in the use of Linked Data extensions to data passing mechanisms used widely on the internet.

EXAMPLE 1: JSON-LD [i.8] has been designed around the concept of a "context" to provide additional mappings from JSON to an RDF model. The context links object properties in a JSON document to concepts in an ontology.

EXAMPLE 2: NGSI-LD [i.9]. The term NGSI (Next Generation Service Interfaces) was first developed in work by the Open Mobile Alliance and has been extended using concepts of Linked Data to allow for wider adoption of ontologies and semantic as well as contextual information in data-driven systems.

As a pre-cursor to the development of a threat ontology for AI based threats, there are a number of threat taxonomies, some found in ETSI TS 102 165-1 [2] and in ETSI TS 102 165-2 [i.17]. These can serve as a starting point for the definition of a threat ontology, and more specifically of an AI threat ontology.

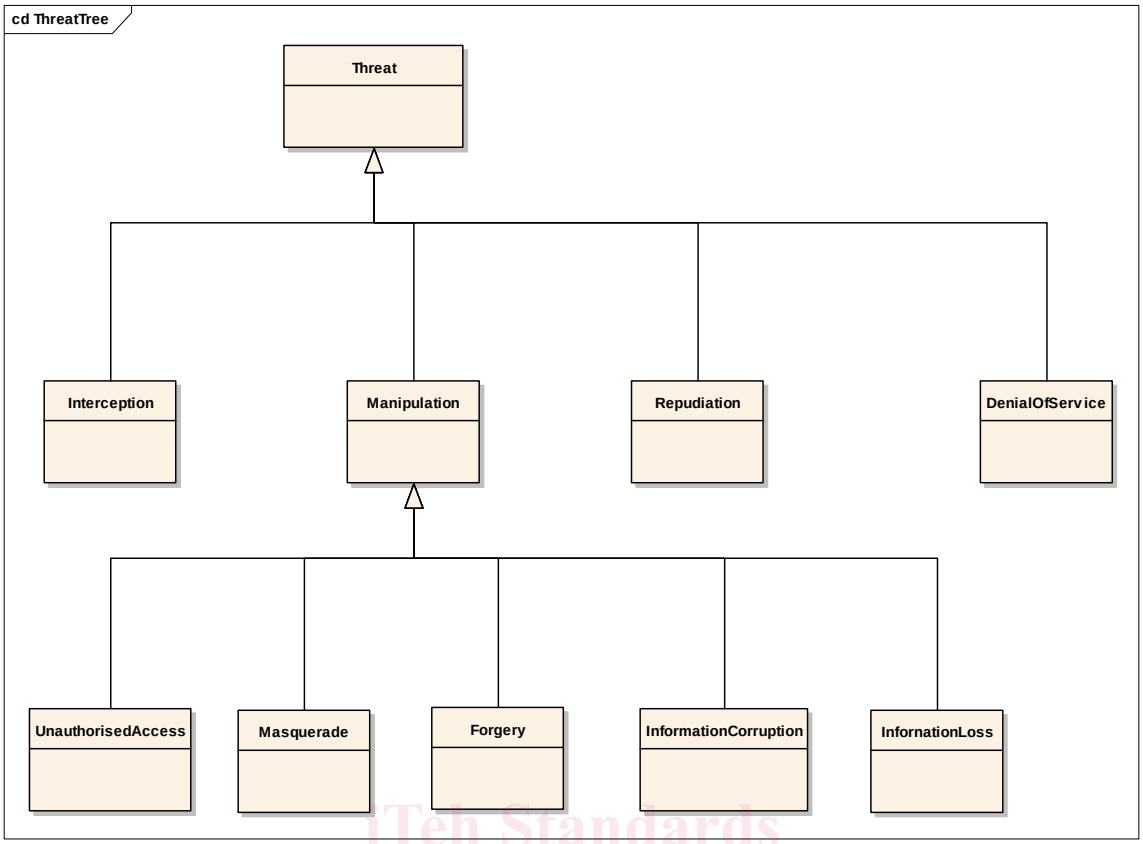

Figure 1: Threat tree (from ETSI TS 102 165-1 [2]) as a taxonomy

In the conventional taxonomy, as in Figure 1 for threats, the core relationship between entities is of type "*is a*", thus Forgery "*is a*" Manipulation, "*is a*" Threat. The relationships in a conventional taxonomy are often unidirectional, whereas in an ontology the normal expectation is that relationships are bidirectional and asymmetric.

EXAMPLE 3: Trust is asymmetric, a pupil is expected to trust a teacher, whereas the teacher is not expected to trust the child.

A simple taxonomy such as in Figure 1 does not easily express side channel attacks, or composite attacks, nor does it capture the asymmetric relationships of things like trust.

EXAMPLE 4: In order to perform a masquerade attack it is often necessary to first have intercepted data, or in order to corrupt data it can be necessary to first have masqueraded as an authorized entity.

Many of the forms of attack on AI that are described in the SAI Problem Statement (ETSI TR 104 221 [i.21]) are in the manipulation tree: data poisoning is a form of information corruption; incomplete data is a form of information loss. The relationship in these cases are "modifies", and "is modified by". Similarly, the terms "threat" and "vulnerability" as defined in ETSI TS 102 165-1 [2] are loosely expressed in the form of ontological relationships. Thus, threat is defined as the potential cause of an incident that may result in harm to a system or organization, where a threat consists of an asset, a threat agent and an adverse action of that threat agent on that asset, and a threat is enacted by a threat agent, and may lead to an unwanted incident breaking certain pre-defined security objectives.

NOTE 2: The nature of data poisoning is complex to clearly identify. The consequence of data poisoning are to limit the ability of the reasoning element of AI to reason towards the "right" solution, but rather to lead the reasoning algorithms to an invalid answer, often in favour of the attacker.

NOTE 3: The deliberate introduction of poisoned data may lead to the AI system exhibiting bias, where the original (non-poisoned) data may have zero biases.

NOTE 4: Whilst it is suggested that incomplete data is a form of information loss the attack vector in an AI system may be quite different than that from non-AI systems.

The structure of the term vulnerability has a similar ontological grouping of relationships, being modelled as the combination of a weakness that can be exploited by one or more threats. A more in-depth examination of the problems of and from AI is found in the SAI Problem Statement [i.21], and in the SAI report on mitigation strategies [i.22].

4.2 Formal expression of an ontology

There are many ways to express an ontology in information science. The most common are:

- OWL - Ontology Web Language [i.5]
- RDF - Resource Description Framework [i.6]

It should be noted, however, that OWL and RDF, whilst common when referring to ontologies, are not equivalent but are mutually supportive.

A simple model that underpins both OWL and RDF is the subject-predicate-object grammar structure (see Figure 2). However, there is also a more complex set of data structures that also look like the object-oriented design concepts (e.g. inheritance, overloading) underpinning design languages such as UML, and coding languages such as C++, Swift and Java. Such taxonomical classifications are also common in science, particularly in the biological sciences.

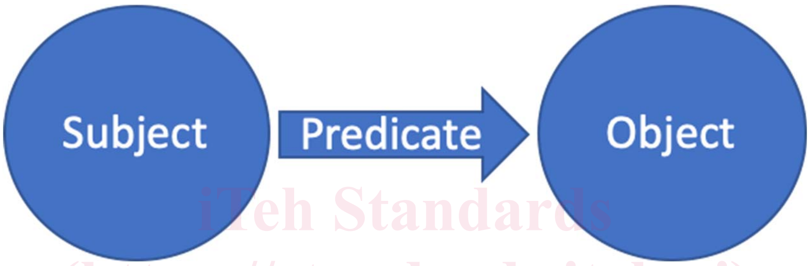

Figure 2: Simplified model of grammar underpinning Ontology

An ontology is expected to consist of the following elements:

- Classes, also known as type, sort or category.

- Attributes, which describe object instances, such as "has name", "has colour", "by definition has a".

EXAMPLE 1: A *protected object* belongs to class *network object*, of sub-type *router*, with name "Router-1" and, by definition, has 1 or more Ethernet ports.

EXAMPLE 2: *Ransomware* belongs to class *threat*, of subclass *denial-of-service*, with attribute *file-encryption*.

- Relationships, as outlined in 4.1 identifies how one class is associated to another.

Expanding from the taxonomy in [i.5], *threat* is modelled as one class, with *threat agent* modelled as another. This is then consistent with the definitions given for the terms "threat" and "vulnerability", and for the relationship to assets as the subject or object in the simplified grammar of ontology.

In the gap between an ontology and natural language, the present document classifies concepts around intelligence as nouns, and relationships as verbs, adverbs, adjectives.

NOTE: Whilst there is a risk in trying to explain AI only by mapping to programming constructs (e.g. objects and classes), or only from data modelling (e.g. tables, lists, numbers, strings and the relationships or type constraints a data model can impose) it is not addressed by the present document but is considered in ETSI TS 102 165-1 [2].

As stated above, an ontology is often described as a specification of a conceptualization of a domain. The result of such an approach to an ontology is to provide standardized definitions for the concepts of a specific domain. In the structure of a technical standard the ontology defines classes (concepts) for sets of objects in the domain that have common characteristics. The objects include specific events, actions, procedures, ideas, and so forth in addition to physical objects. In addition to the concepts, the ontology describes their characteristics or attributes, and defines typed relationships that may hold between actual objects that belong to one or more concepts.