ETSI GR SAI 001 V1.1.1 (2022-01)

GROUP REPORT

Securing Artificial Intelligence (SAI); AI Threat Ontology

Disclaimer

The present document has been produced and approved by the Secure AI (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-001

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

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Notice of disclaimer & limitation of liability

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

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

Copyright Notification

Contents

Intellectual Property Rights

Essential patents

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

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. 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].

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

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

[i.6] ETSI GR SAI 004: "Securing Artificial Intelligence (SAI); Problem Statement".

[i.7] ETSI GR SAI 005: "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".

[i.8] W3C® Recommendation 11 December 2012: "OWL: OWL 2 Web Ontology Language Document Overview (Second Edition)".

[i.9] RDF: RDF 1.1 Primer; W3C® Working Group Note; 24 June 2014.

[i.10] 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.11] W3C® Recommendation 16 July 2020: "JSON-LD 1.1: A JSON-based Serialization for Linked Data".

[i.12] ETSI GS CIM 009 (V1.2.2): "Context Information Management (CIM); NGSI-LD API" (NGSI-LD).

[i.13] "The Emergence Of Offensive AI".

NOTE: Available at https://www.oixio.ee/sites/default/files/forrester_the_emergence_of_offensive_ai.pdf.

[i.14] "Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter".

NOTE: Available at https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter.pdf.

[i.15] Li Chen, Chih-Yuan Yang, Anindya Paul, Ravi Sahita: "Towards resilient machine learning for ransomware detection".

NOTE: Available at https://arxiv.org/pdf/1812.09400.pdf.

[i.16] Alejandro Correa Bahnsen, Ivan Torroledo, Luis David Camacho and Sergio Villegas: "DeepPhish: Simulating Malicious AI".

NOTE: Available at https://albahnsen.files.wordpress.com/2018/05/deepphish-simulating-malicious-ai_submitted.pdf.

[i.17] Common Weakness Enumeration Project.

NOTE: Available at https://cwe.mitre.org/index.html.

[i.18] ETSI TS 118 112: "oneM2M; Base Ontology".

[i.19] The Smart Appliances REFerence (SAREF) ontology.

NOTE: Available at http://ontology.tno.nl/saref/.

[i.20] ETSI TS 102 165-2: "CYBER; Methods and protocols; Part 2: Protocol Framework Definition; Security Counter Measures".

[i.21] ETSI GR SAI 002: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

[i.22] Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar: "Threat Modeling AI/ML Systems and Dependencies".

NOTE: Available at https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml.

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in ETSI GR SAI 004 [i.6] apply.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in ETSI GR SAI 004 [i.6] and the following apply:

AI	Artificial Intelligence
CAV	Connected and Autonomous Vehicles
ICT	Information Communications Technology
ITS	Intelligent Transport Systems
JSON	JavaScript Object Notation
NGSI-LD	Next Generation Service Interface - Linked Data

OWL	Web Ontology Language
RDF	Resource Description Framework
SAI	Securing Artificial Intelligence

4 From threat 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. There exist several sources of ontologies in security and intelligence, a summary of which are given in the bibliography of the present document. The present document complements much of the existing work, but with a focus on understanding and identifying the impact of AI on risk, particularly where mitigations use AI techniques, or where the adversary uses AI techniques. The role of AI in risk assessment is addressed in more detail in clause 5, while clause 6 extends this analysis to consider the roles of AI when building and assessing the threat landscape. The understanding in clauses 4, 5 and 6 inform the design and discussion of an ontology for AI Threats (and mitigations) given in clause 7.

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: 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.9] notations, there is growth in the use of Linked Data extensions to data passing mechanisms used widely in the internet.

EXAMPLE 1: JSON-LD [i.11] 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.12]. 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 [i.5] and in ETSI TS 102 165-2 [i.20]. These can serve as a starting point for the definition of a threat ontology, and more specifically of an AI threat ontology.

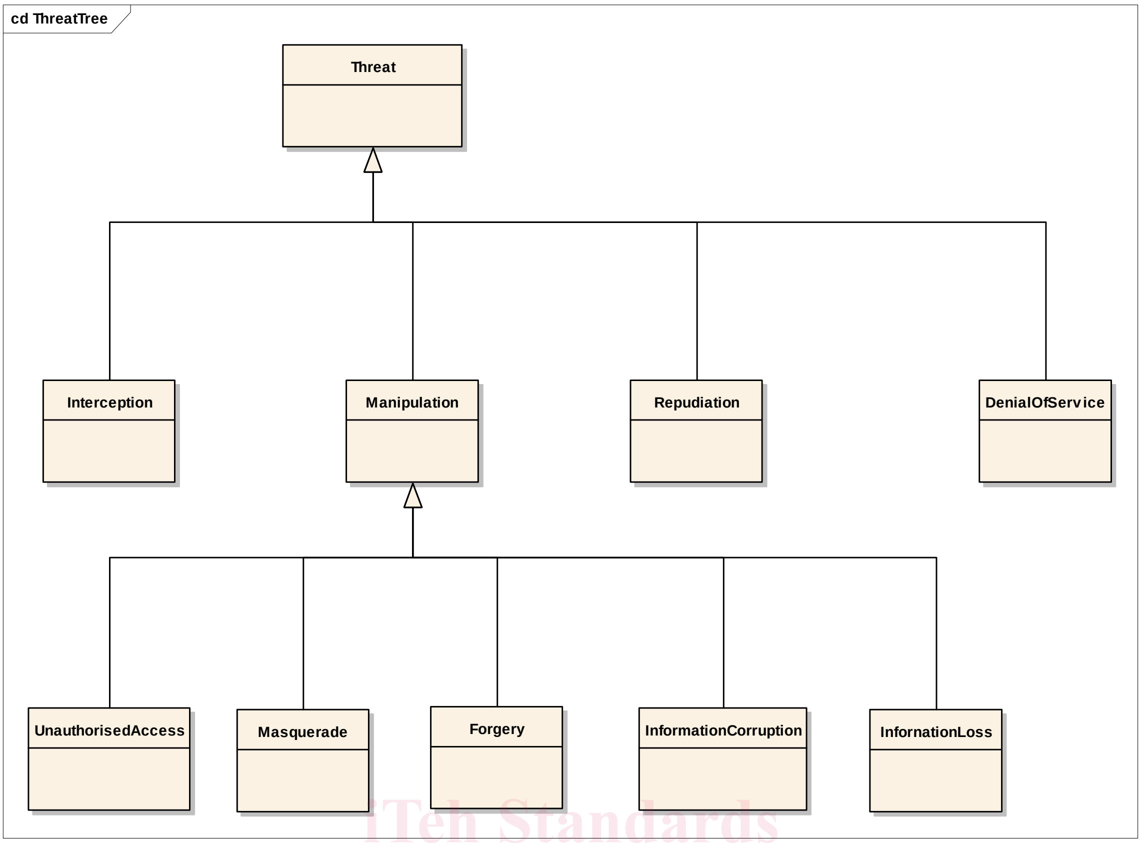

Figure 1: Threat tree (from ETSI TS 102 165-1 [i.5])

In the conventional taxonomy, as in figure 1, 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 1: 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 2: In order to perform a masquerade it can be 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 GR SAI 004 [i.6]) 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 [i.5] 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 it is noted that a threat consists of an asset, a threat agent and an adverse action of that threat agent on that asset, and further that a threat is enacted by a threat agent, and may lead to an unwanted incident breaking certain pre-defined security objectives.

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.6], and in the SAI report on mitigation strategies [i.7].

4.2 Formal expression of an ontology

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

- OWL - Web Ontology Language [i.8]
- RDF - Resource Description Framework [i.9]

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.

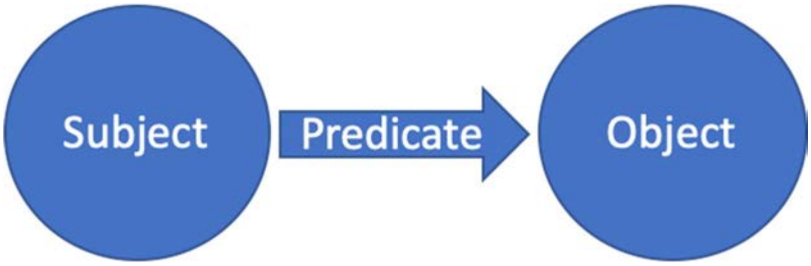

Figure 2: Simplified model of grammar underpinning Ontology

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

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, it can be useful to classify concepts around intelligence as nouns, and relationships as verbs, adverbs, adjectives. However, it should be understood that 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). This difficulty in understanding what intelligence is, how AI differs from human intelligence, and the philosophical nature of intelligence is one of the purposes of the present document to highlight although not attempt to resolve.

An initial problem with AI, and security aspects of AI, is that the domain does not appear to be well bounded, and the level of uncertainty is high. With respect to the Rumsfeld statement quoted, below the domain of AI has many unknown unknowns (*the ones we don't know we don't know*), the most pressing of which is a definitive view of intelligence.

QUOTE: *"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones. Attributed to Donald Rumsfeld on 12-February-2002."*

However, in addressing the AI problem in the present document, whilst there is a degree of uncertainty regarding Artificial General Intelligence (AGI), it has, for the purposes of the present document, a low likelihood of actualisation and therefore the focus of the present document is on Artificial Narrow Intelligence (ANI). See also the discussion in clause 4.5.

Intelligence and intellect are two additional terms that are often confused. An ANI will not be considered as intellectual although a AGI will be, where the definition of intellect is given as having the faculty of reasoning and understanding objectively, especially with regard to abstract matters, and a machine having this faculty can be described as intellectual. An ANI will in many cases be designed in such a way that it cannot be intellectual - rather it is designed to be good at a single function. In contrast a AGI will be able to apply learning or knowledge from one field to another and to abstract knowledge to multiple fields.

4.3 Alternative mathematical approach

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 semi-formal structure of a standard document therefore, the ontology defines classes (concepts) for sets of the different 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.

As indicated in clause 4.2, information in an ontology is conventionally encoded, in languages such as RDF and in representations such as OWL, as a list of triplets (the "subject-relationship-object" concept), where the subject is the domain under analysis, the objects are all relevant concepts affecting the subject and the relationships are the indicators for the involving level of each concept (concepts) in the problem (subject) and the interdependency relationships between concepts and between the concepts and the subject.

For illustrative purposes, this can be expressed as a mathematical representation of a linear system:

$$Y = \beta 0 + \beta 1 X1 + \beta 2\, X2 + \beta 3\, X3 + \dots + \beta n\, Xn + \mu$$

$$\beta x \neq 0$$

where Y is the domain variable to be explained by the ontology, X are the concepts as explicative variables, and coefficient β represents the relationship of the explicative variables over the variable Y, and μ is the error factor for the "unknown" concepts in Y.

In regard to standardization, ontologies, when formally modelled, provide explicit knowledge models for particular domains that can assist in both structuring the problem and in identifying where standards can assist in specifying the nature of the domain in such a way that it becomes known.

4.4 Relationship to other work

In the scope of the present document an ontology is also developed to assist in the development of strategies in securing AI. This addresses the modes in which AI can exist in a system, shown figuratively in figure 3 below.