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

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.

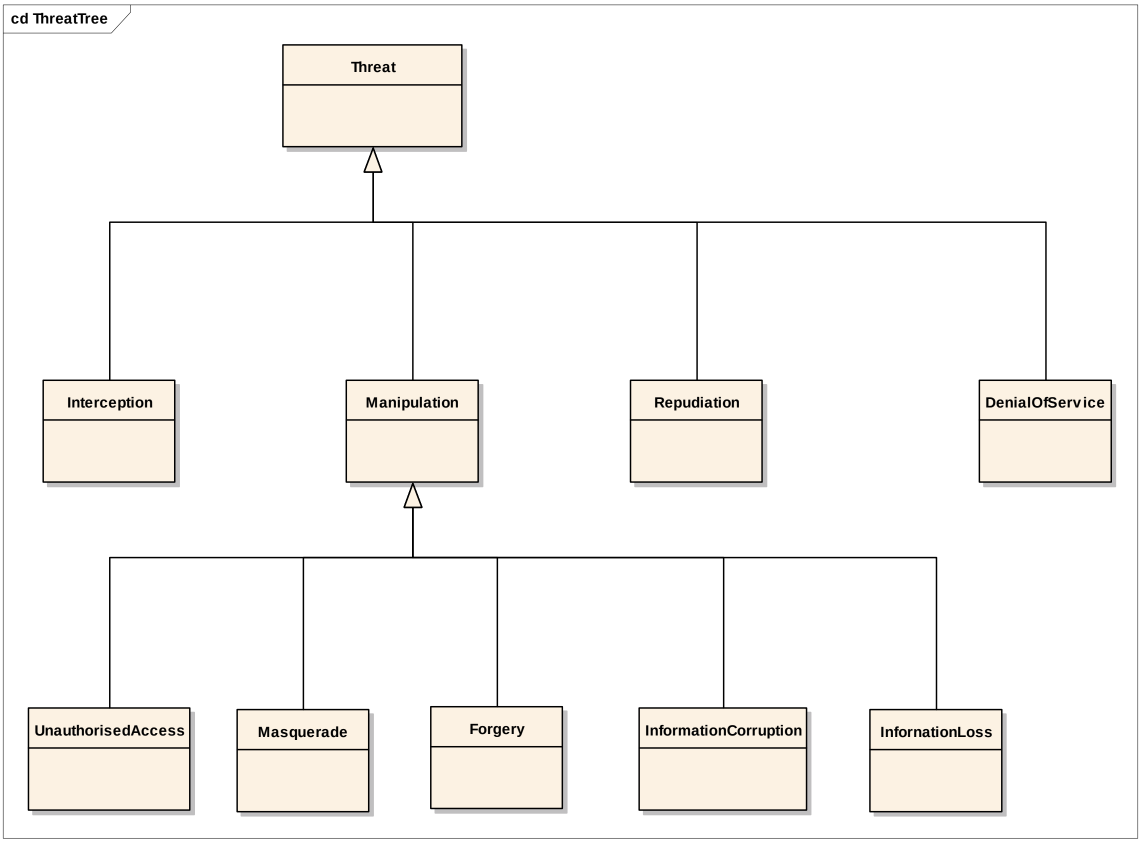

Figure 1: Threat tree (from ETSI TS 102 165-1 [i.5])

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.

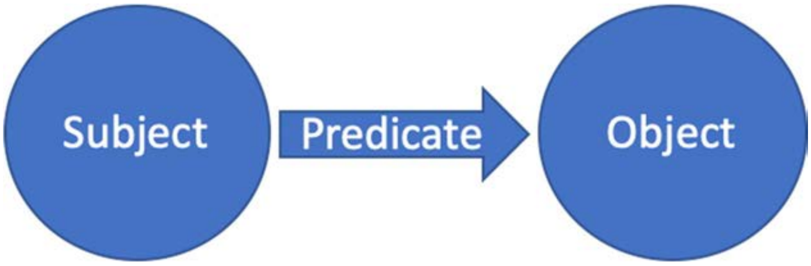

Figure 2: Simplified model of grammar underpinning Ontology

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:

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

$$\beta x \ \neq \ 0$$

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.

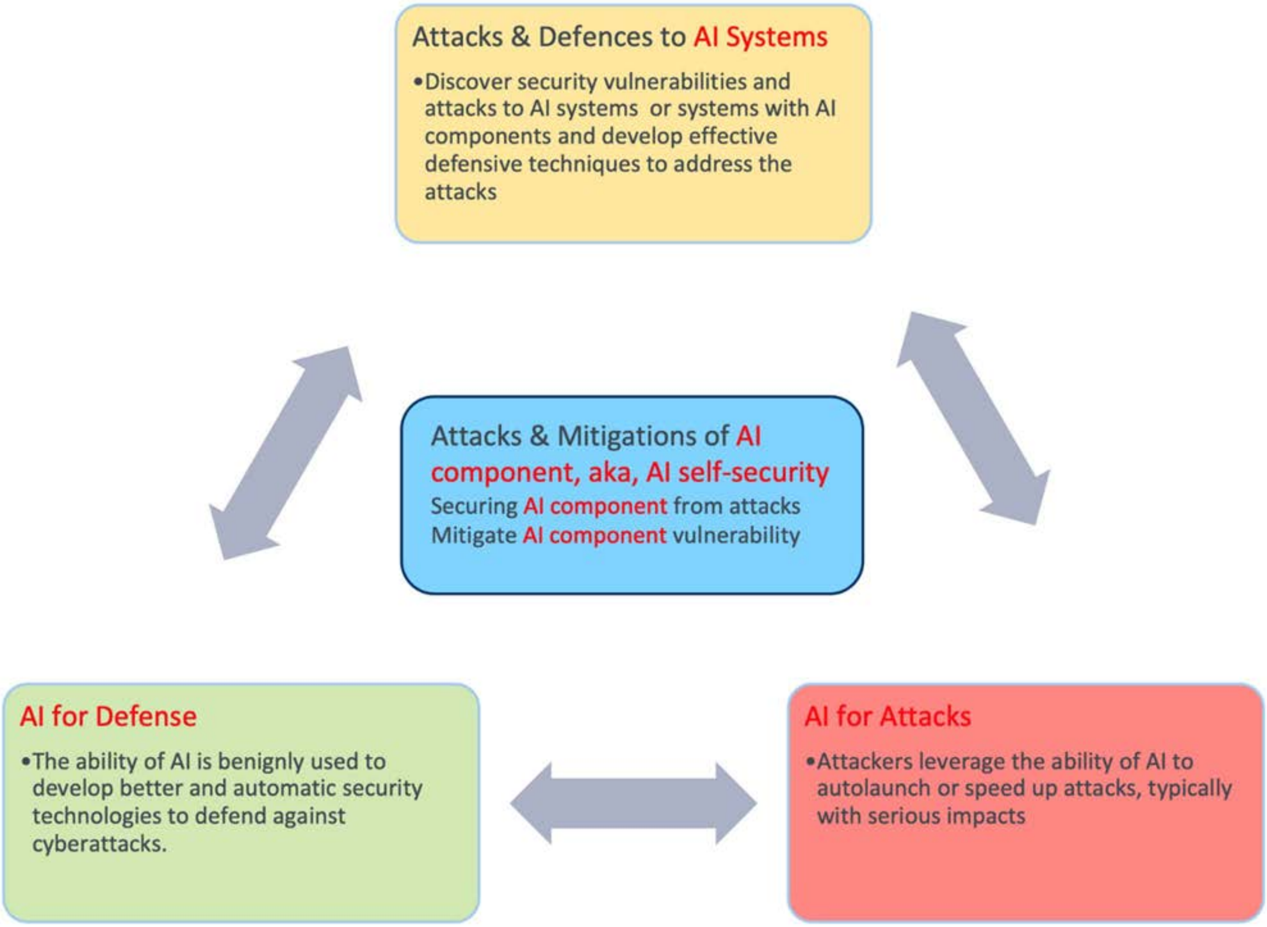
Attacks & Defences to AI Systems
•Discover security vulnerabilities and attacks to AI systems or systems with AI components and develop effective defensive techniques to address the attacks
Attacks & Mitigations of AI component, aka, AI self-security
Securing AI component from attacks
Mitigate AI component vulnerability
AI for Defense
•The ability of AI is benignly used to develop better and automatic security technologies to defend against cyberattacks.
AI for Attacks
•Attackers leverage the ability of AI to autolaunch or speed up attacks, typically with serious impacts

Figure 3: Modes of application of AI in networks and services

AI can be deployed in attack mode against components or systems, defence mode in countering attacks on components or systems, and proactively in understanding attacks on components and systems. Underpinning both attack and defence modes is the goal of understanding the problem associated to AI and the risk to the system of AI.

In ETSI GR SAI 004 [i.6], "SAI Problem Statement", the following definition of AI is offered:

> ***Artificial intelligence*** *is the 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.*

In ETSI GR SAI 004 [i.6], the intent is to describe the challenges of securing AI-based systems and solutions, in an environment where data, algorithms and models (in both training and implementation environments) are portrayed as challenges to the overall systems analysis and understanding in respect of AI as a system threat. In ETSI GR SAI 004 [i.6] and in ETSI GR SAI 005 [i.7] there is a subtle change of emphasis with regards to data that is input to a system where the dominant AI mechanism is machine learning. In the specific subset of AI that is Machine Learning (ML), there are further modes of learning that can be defined. As noted above, the role of ontologies is implied in most ML systems as a means of structuring the input. In practice most ontologies are incomplete: they tend to be domain specific, whereas in practical systems an entity can exist in more than one domain.

EXAMPLE: A potato will naturally exist in an ontology describing tubers and variants of the nightshades, but will also naturally exist in an ontology describing foodstuffs and diet. Domain specific knowledge, knowing that a potato is related to a tomato in the family of nightshades, would not necessarily reveal the existence of poutine. Or in other words there may not be an obvious link between domains.

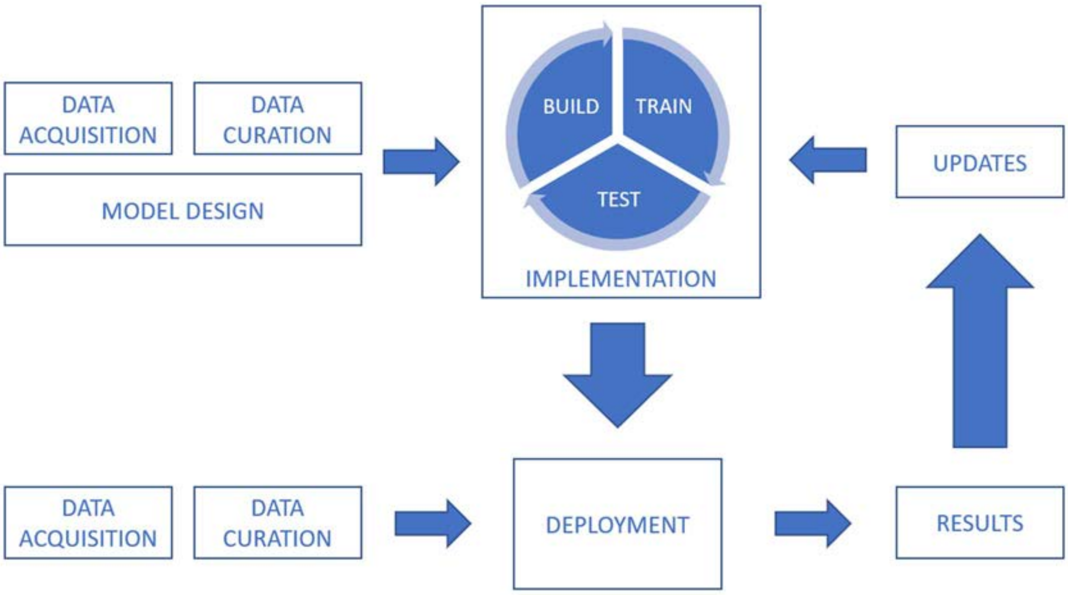

Figure 4: Typical machine learning lifecycle from ETSI GR SAI 004 [i.6]

The ML lifecycle considered in ETSI GR SAI 004 [i.6] identifies a number of ML strategies:

- ***Supervised learning*** - where all the training data is labelled, and the model can be trained to predict the output based on a new set of inputs.
- ***Semi-supervised learning*** - where the data set is partially labelled. In this case, even the unlabelled data can be used to improve the quality of the model.
- ***Unsupervised learning*** - where the data set is unlabelled, and the model looks for structure in the data, including grouping and clustering.
- ***Reinforcement learning*** - where a policy defining how to act is learned by agents through experience to maximize their reward; and agents gain experience by interacting in an environment through state transitions.

In each case ML can be used for both classification problems, and for prediction problems.

In applying annotations or labels to data, it is common practice to first define the domain data using an ontology, in its simplest form of a semantic data set. As an ontological or semantic data definition is unlikely to be complete, one role often assigned to AI/ML is to further develop the data description, by identifying additional patterns through finding new correlations and asserting new causations.

Attack strategies against the ML workflow (illustrated in figure 5) may apply to each stage or process flow, and the risk associated to each is assessed independently. The role of risk management is covered in more detail in clause 5 of the present document.

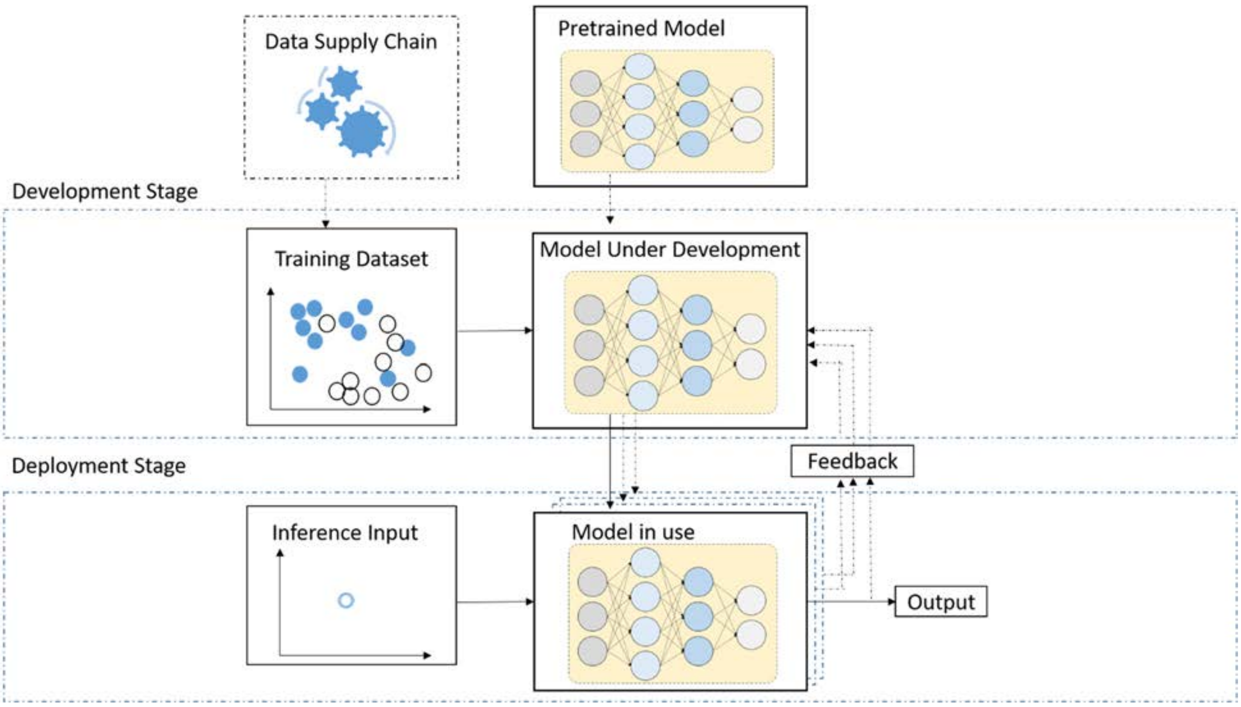

Figure 5: Machine learning model workflow from ETSI GR SAI 005 [i.7]

4.5 Hierarchies and relationships in AI

In many analyses of AI there are a number of intermediate definitions:

- Artificial Narrow Intelligence (ANI), also termed Weak AI:
    - Applying intelligence to only one context, for example, autonomous driving, speech recognition.
- Artificial General Intelligence (AGI), also termed Strong AI and:
    - Applying intelligence to any intellectual task, at a level equivalent to a human.
- 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.

The goal of AGI and of ASI is that all of the characteristics associated to intelligence are met. Those characteristics include:

- reasoning: The application of learned strategies in order to solve puzzles, and make judgments where there is uncertainty in either the input or the expected outcome;
- representing knowledge, including common-sense knowledge, to an independent third party;
- planning;
- learning;
- communicating in natural language (to human third parties); and
- integrating all these skills towards common goals (of the third party).

Other important capabilities include the ability to sense (e.g. to see) and the ability to act (e.g. to move and manipulate objects) in the world where intelligent behaviour is to be observed. The increasingly connected world characterized by the Internet of Things (IoT) suggests that any connected sensor, or actuator, may be available to a motivated AI.

In terms of examples presented in present document, having knowledge of the existence of a potato as a tuber but not being able to link that knowledge of the role of a potato in the diet represents ANI, whereas having knowledge of both roles of the potato, and to gain knowledge of other uses such as in printing, may suggest AGI, whilst ASI may suggest an ability to develop new knowledge relating to application of the potato.

As mentioned in clause 4.1, there is no practical realization of AGI and ASI. ANI, however, is rapidly becoming adopted in a huge range of sectors and applications. As such, the present document focusses primarily on ANI and the opportunities and threats it presents.

5 AI and risk assessment

5.1 Core understanding of risk assessment

5.1.1 Stages in AI

In all models of intelligence there are 3 broad steps:

1) Data gathering.

2) Data processing.

3) Applying insights gained from the data processing.

As outlined in ETSI GR SAI 004 [i.6] and in ETSI GR SAI 005 [i.7] AI presents challenges to each of the 3 steps.

Data is required to inform and instruct both human and machine decision making. With very large volumes of data to evaluate, often at high speeds, the need for programmatic assistance in data processing and decision making becomes key, and thus offers the rationale for development of AI in general. Many applications of ML (as an instantiation of AI) are intended to enable classification, prediction, filtering or annotation of data, while others are used to take action in situations or conditions where humans cannot.

EXAMPLE 1: Classification, a subclass of ML, is often applied in image processing, where based on learned patterns of what (say) each of cats and dogs look like, an application, when presented with a random image can analyse it and classify it as containing a cat, a dog or neither.

NOTE: All ML relies on sufficient volumes of representative training data. In the classification case mentioned above, for example, if all the images for dogs were long-snouted breeds such as Setters, Labradors and Terriers, then a dog breed without such a characteristic (e.g. the flat-faced breeds such as Bulldogs) may not be identified as a dog.

EXAMPLE 2: Regression or Predictive ML takes a set of data and makes a prediction for a future state or value of a data point in the future. This is commonly seen in areas such as weather forecasting where historic data records, and incoming real time sensor readings and images are processed to give a prediction of (say) the temperature likely at a particular geo-location at some point in the future.

EXAMPLE 3: Reinforcement Learning (RL) is a subset of AI where a software agent learns the optimum actions to take in response to environmental conditions, for example financial trading agents choosing when to buy or sell stocks at speeds that humans cannot reach.

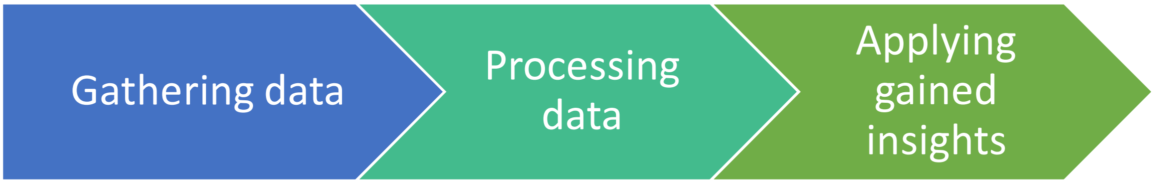

Figure 6: Steps in gaining intelligence from data

From both security and risk perspectives, each stage, and how they are impacted by various forms of threat, how the threat agent is instantiated, and the resultant form of attack (see also ETSI GR SAI 004 [i.6] and ETSI GR SAI 005 [i.7]) are factors that need consideration both in isolation and in concert to determine the resultant risk the AI presents. In more straightforward terms, a risk analysis is required to consider the interaction between stages as well as each stage in isolation. That interaction can be linear (the feed of data from the data gathering phase to the processing phase, for example) or involve feedback loops (the analysis phase can be used to modify the processing phase which in turn can modify the data gathering phase, for example). The succeeding subclauses consider the impact of AI on risk.

AI/ML tools, as they currently exist, are inherently statistical. They can identify and extract correlation but cannot apply the critical reasoning required to derive analytical causation: a model cannot know what is sensible, only what patterns align with the data it has previously been supplied with. An attacker can take advantage of this to manipulate data, outputs and actions in all phases, depending on the ML paradigm being used.

5.1.2 Business modelling of risk

Business-centric assessments of risk often overlap with the ICT modelling of risk, although the overall view of risk is focussed on aspects of business continuity. Thus, the dimensions offered in table 1 are addressed with respect to the role of AI on the business.

Table 1: Business assessment indicators for risk modelling

Indicator	Risk assessment questions
Magnitude	What risks will failure of AI create? Classifications of where risk lies include: Monetary loss (i.e. direct financial impact) Compliance (i.e. will existing compliance procedures be maintained) Legal (e.g. will existing legal safeguards apply) Reputation (e.g. how will readiness or failure to be ready impact the reputation of the organization either in absolute terms or by comparison to peer and competitor organizations)
Duration	How long has operation to be maintained for each asset or class of assets impacted by AI?
Scope	How far down the supply chain is the impact of AI?
Severity	Can damage due to degradation or interruption of each services that uses AI be quantified?
Response	Is there a plan to migrate compromised AI components to alternative modes of operation?

Where an AI implementation is working on behalf of a business, it is expected to inherit the aims, mores, and ethics of that business. As indicated in table 1, the AI component is treated in terms that are similar to assessments of a staff member or other business asset that is subject to some form of attack. The greater the dependence of the organization on AI, the greater the impact of any attack on the assets of the AI (the data, the processing algorithms, and so forth).

5.1.3 ICT modelling of risk

The model of risk analysis given in ETSI TS 102 165-1 [i.5] calculates risk as the product of impact (of an attack) and likelihood (of an attack). Likelihood is calculated using several metrics that assess the level of knowledge and access of the attacker to attack the system. Figure 7, taken from ETSI TS 102 165-1 [i.5] considers a system design derived from analysis of system objectives and requirements, security objectives and requirements, and associated assurance objectives and requirements. The system itself is modelled as an aggregation of assets. An AI entity can be instantiated as one or more of the system assets.

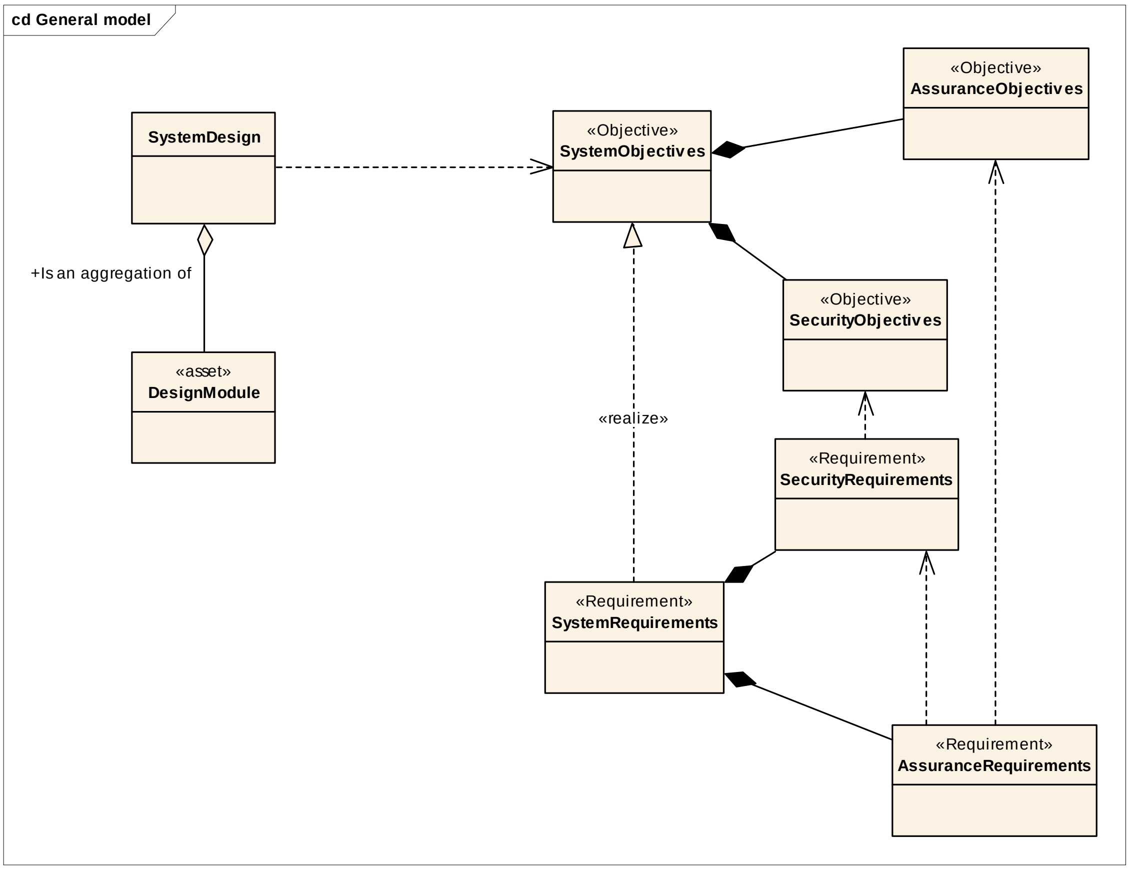

Figure 7: Relationship between system design, objectives and requirements

The risk analysis model from [i.5] identifies the asset-threat-weakness-vulnerability-countermeasure relationship to system design, and is illustrated in figure 8. The consideration of AI in this model is somewhat dependent on the nature of the AI component, as different paradigms will have different associated assets and dependencies. There can be a dependency on specific forms of hardware, e.g. GPUs, or on other assets in a less direct fashion than has been considered in conventional system design. The role of data is also very different in an AI system as opposed to conventional software (see below).

EXAMPLE: In conventional software coding the software designer makes specific decisions regarding how the software reacts to input, thus the knowledge and experience is captured by the code but does not of itself form part of the code. In contrast a machine learning algorithm carries its learning knowledge as an explicit system asset.

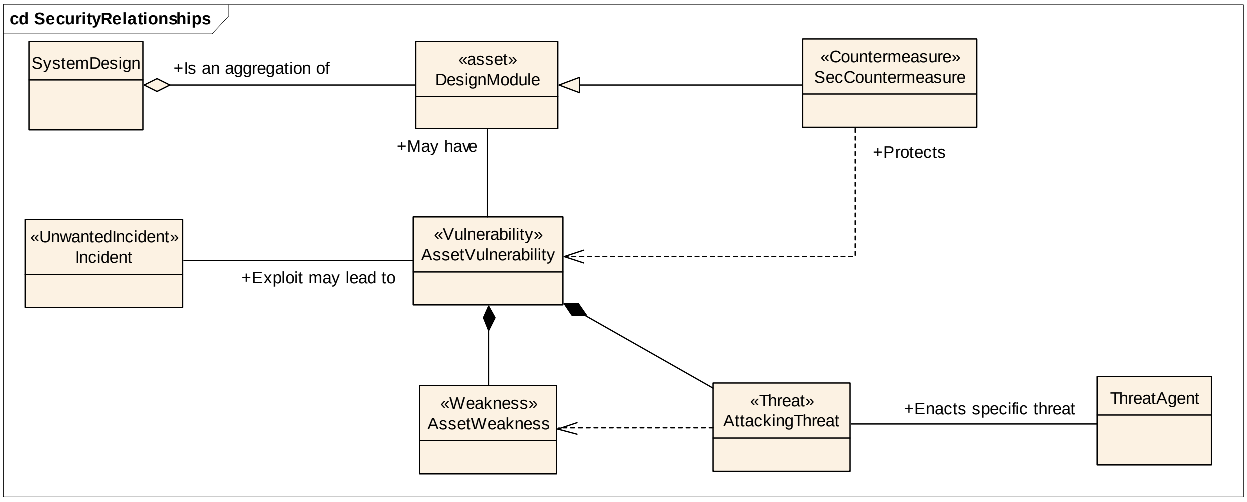

Figure 8: Generic security TVRA model

Thus in the specific system model identified in figure 8 the role of AI can be seen at least as follows:

- AI acting in lieu of a (human) threat agent;
- AI as a specific asset of the system (i.e. as a designed in element) and integral to the system design;
- AI as an instance of a countermeasure to protect specific assets in the system;
- AI as an instance of a specific threat against known weaknesses in the system.

As mentioned above, a key way in which ML/AI algorithms differ from traditional software systems is that they allow inference from data to determine outputs, meaning that programs have a dependence on the training data that they are constructed from which can be exploited in several ways. Datasets in an AI system hence represent significant assets with a different type of relationship to other components. It is also noted that an AI asset is mutable in that how it behaves is dependent on the logic of the AI/ML component and on the nature of the inputs.

This data dependence also means that for any given input, one cannot expect a single, testable expected output, as one might from a traditional software system. This means that small variations in the system outputs could be hiding a deliberate alteration made by an attacker, and adds an additional level of complexity to assessing risks and assuring the security of systems.

5.2 Extension of ETSI TVRA method in presence of AI

5.2.1 Effect on risk impact

Using the weighted model in ETSI TS 102 165-1 [i.5], if a system's purpose with or without AI is identical, then the impact of system failure is identical irrespective of the role of the AI. The impact of AI on risk impact should therefore be zero. If a system is modified by the addition of an AI, then the impact of system failure is very likely to be modified as the purpose is modified.

EXAMPLE: If the AI is the cause of failure and the business is dependent on the AI to provide functionality then impact will be high, whereas if the functionality is not dependent on the AI (i.e. similar functionality can be offered without an AI) the impact can remain at either low or medium if that was the original assessment.

Table 2: Asset impact (from [i.5])

Impact	Explanation	Value
Low	The concerned party is not harmed very strongly; the possible damage is low.	1
Medium	The threat addresses the interests of providers/subscribers and cannot be neglected.	2
High	A basis of business is threatened and severe damage might occur in this context.	3

However, introducing AI can change working practices, or can make available new capabilities. The way in which AI impacts the business may not be immediately apparent but is likely to have been introduced for reasons which will affect the business, for example speeding up a process, modifying a process or initiating a new process.

5.2.2 Effect on risk likelihood

Risk likelihood is determined, in the context of ETSI TS 102 165-1 [i.5] as the weighted summation of the following factors:

- System knowledge.
- Time.
- Expertise.
- Opportunity.
- Equipment.

An embedded AI can detrimentally affect each of these factors.

To enumerate the influence of AI, it is essential to define and understand what AI is and how it acts. If an ontology of threat is developed, and the threats are impacted by AI then it suggests that an ontology of AI is also required.

For simple analysis the present document addresses a number of ways to use AI and how the likelihood is affected:

- AI embedded in business systems and operating some or part of the business system:
    - In this case the AI is a core asset of the system and cannot be removed from the system without destroying the system. In such an instance the "intelligence" of the system is an integral part of the system software (in conventional software systems the system "intelligence" is external to the system and can be seen in decisions made by the operators of the system).
- AI associated to a business system and operating to protect the business system from attack:
    - In this case the AI is acting as a countermeasure. Whilst not directly impacting the active business assets of the system it is in a position to monitor the system, take data from the system and manage other elements of the countermeasure fabric of the system. Simplified examples of AI, or AI-like, application in this case include monitoring of the system behaviour in order to determine if an attack, represented by unexpected changes in behaviour, is underway. Similar examples include scanning of system data to determine the presence, or likelihood of the presence, of malicious content or behaviour.
- AI acting as the adversary, or as an orchestrator of multiple adversaries to break the business system:
    - In this instance the AI is acting as the threat agent or as an orchestrator of multiple threat agents. In like manner to the case of intelligence being embedded in the business system here the AI is embedded into the adversarial system. Such systems may be able to modify their attack behaviour based on feedback and accumulated knowledge of the systems.

As has been indicated above the ETSI TVRA method can be extended to address the role of AI components in system threat modelling. Additional guidance on AI-specific threat modelling can be found in the bibliography.

6 Threat landscape

6.1 Threat dimensions

The TVRA model in figure 8 above states that "A ***threat agent*** enacts a specific ***attack*** against a system ***weakness*** to exploit a ***vulnerability***". It is assumed that AI should be considered in the context of each of the items in with **bold text**, in both offensive and defensive contexts. The SAI problem statement in ETSI GR SAI 004 [i.6] identifies ways in which a threat agent can invoke particular forms of attack, while in ETSI GR SAI 005 [i.7] a number of specific mitigations are identified. In each of ETSI GR SAI 004 [i.6] and ETSI GR SAI 005 [i.7] the focus is on attack and defence of AI-inspired attacks on AI systems, whereas in the present document the focus is on the understanding of what AI means and of defining the AI domain itself.

Acquired intelligence, i.e. intelligence from learning, requires knowledge of data semantics (i.e. what data elements mean) and data context (i.e. how data elements are related), and conventional domain ontologies offer this form of data labelling. The richer the ontology of the input data, i.e. the more that data is labelled, the closer the ontology is to the world model required by the AI to represent the world view for the intelligence in the machine. In other words, semantic labelling is a major step forward in gaining an understanding of data.

EXAMPLE: The numerical value 42 can be syntactically represented as a signed integer which in computing terms means certain functions can be applied to it (arithmetic functions say) and a compiler will be able to warn if functions will fail based on knowledge of the syntax. However, of itself the value of the integer does not confer knowledge whereas adding a semantic label to it allows reasoning to be applied. Simple semantic labels can be seen in the names given by programmers to constants and variables, but in the wider context semantics have to be transferred with the data in order that the receiver has knowledge of what the value means, or is associated to.

6.2 Attacks as instance of threat agent

According to a 2019 report by Forrester, 86 % of cybersecurity decision makers are concerned about the offensive use of AI by threat actors [i.13]. As with many other organizations, adversaries are increasingly looking to AI to automate, scale and speed up activities which are currently conducted manually. This is particularly where malicious actors target indiscriminately, where the lower likelihood of a successful single attack is offset by the opportunity to attack many more targets. As such, although the impact of individual threats is unlikely to change, the scale of attacks, the likelihood of attacks being successfully carried out, and the difficulty in responding or remediating in a timely manner may all increase dramatically.

Opportunities range across the phases of the attack chain, from reconnaissance to exfiltration and impact. Public data, particularly social media, presents a significant opportunity for AI-based exploitation. For example, supervised and unsupervised learning can be used as tools to mine data for large-scale target discovery and spearphishing email generation, and reinforcement learning for conducting automated phishing [i.14].

AI can be used to evade defence mechanisms, including those defences based themselves on AI techniques. Examples include using Generative Adversarial Networks (GANs) to evade ransomware detection [i.15], and using deep neural networks to evade phishing detection [i.16].

AI tools can also be used to enable other kinds of attack. ML approaches are increasingly being demonstrated in side-channel analysis as an alternative to traditional statistical tools. The increasing sophistication of AI-based techniques for generating fake biometric data, and creating falsified images (so-called DeepFakes, can also pose a threat to, and undermine, trust in the integrity and authenticity of data in all aspects of a business.

EXAMPLE 1: In one case, AI-based voice-mimicking software was used to defraud an energy company of $240 000, by convincing an employee to make a fraudulent bank transfer.

EXAMPLE 2: Deepfakes, a term applied to synthetic media in which a person in an existing image or video is replaced with someone else's likeness, has been used in a number of situations both positive and negative, including replacement of images to commit fraud or misdirection, and in more positive environments to place a dead actor in a film.

In all these examples, the use of AI is unlikely to change the impact of a successful exploitation. However, it can increase the likelihood of an organization being targeted and/or of attack attempts being successful, and hence can increase the overall risk.

6.3 Adversarial Goals

6.3.1 Violation of Confidentiality

As mentioned above, data is a crucial asset in an AI-based system. By definition, information about training data is encoded in a model itself: a model can be considered the aggregated understanding of a scenario or task derived from analysis of many examples of that scenario. Techniques exist whereby an adversary can infer aspects of this information, for example reconstructing training data examples (so-called *model inversion*) which represents a violation of confidentiality, and potentially privacy where a model has been trained on personal data. Similarly, interrogation of a model can leak information about the model itself, which can represent a leak of proprietary information. This can be particularly damaging where a business model is based around a well-trained model.

6.3.2 Violation of Integrity and Availability

As described in clause 5, the role of AI in a system is usually to make inferences about data to enable downstream decision making (human or machine), or to carry out actions based on input data. Compromise of the AI component can hence lead to a violation of integrity of the system, in that inferences and decisions will be inaccurate, and overall system performance will be degraded. If significant enough, this degradation can constitute a loss of availability, either of the component itself or of the whole system, depending on the AI's role in that system. The various modes of compromise are described in ETSI GR SAI 004 [i.6].

One variant of availability compromise that particularly applies in an AI context is "reputational compromise". AI is not well understood in many domains, and one of the purposes of the present document is to offer a wider understanding of AI. Users may be wary of trusting model outputs, especially when model reasoning is difficult to explain or is sufficiently different from human reasoning. An attacker can target an AI system in an attempt to damage trust in AI itself, and disrupt or damage an organization by preventing or reversing the adoption of AI technologies.

A key aspect of the integrity definition of the CIA model is the ability to reverse the effects of attacks. The relationship between data and a model, as well as the probabilistic nature of models themselves, make understanding and reversing adversarial action more challenging for AI systems than for traditional software. Once a model is trained, it is extremely difficult to undo the effect of malicious datapoints without significant retraining, which requires large amounts of reliable training data and compute. Techniques exist to harden some types of AI against adversarial input, either as a preventative or remediation measure; these are areas of active research.

6.4 Threat modelling

6.4.1 Attacker objectives

As with any cyberattack, an adversary will ultimately be aiming to extract information from a system or affect its operation in some way. The adversary can choose to do so using AI, or by attacking AI components, but ultimately the objective will be to affect a system or the information within it. Where an AI component of a system is used to provide context for decision making within a system, or make decisions itself, then compromise of the AI component will affect those decisions. The effect of compromising decisions will depend on the design and purpose of the system.

On a more granular level, attacks on AI systems can be thought of as aiming to force a model to do, learn or reveal the wrong things:

- **Do** - the actor aims to engineer an input to a model such that the output will be incorrect. The actor has control over the input but not the model itself. This class of attack is known as *evasion*. An example would be a malware author manipulating an executable binary so that an ML-based security product classifies that binary as benign software.

- **Learn** - the actor wishes to *poison* a model such that it will fail to operate as the intended, in a targeted or indiscriminate way. The actor has control over the data and/or model. The actor may be looking to degrade the overall performance of a model (functionally a denial-of-service attack), or to introduce a backdoor or trojan. In the former case, the degradation or disruption can be the actor's ultimate aim, or they wish to use the reliably poor performance of a model to achieve a downstream effect. In the backdoor case, while overall model performance will remain consistent, the actor will be able to reliably conduct an evasion attack as above.

- **Reveal** - the actor aims to uncover information about the model and/or the data used to train it. This can be for espionage or theft purposes: actors wish to steal the model itself, reveal sensitive training data or learn if particular examples have been used for training. Many of the evasion and poisoning attacks above are enabled by access to the model or an approximation of it: as such, an actor may wish to steal a model as a preparatory step in conducting another type of attack.

As already discussed, an adversary's objective in using AI for offensive purposes is likely to be similar to any organization's: increased efficiency, speed and scale; automation; and easier exploitation of large amounts of data. If attacks using AI fail, a sufficiently motivated attacker may still be able to perform the attack manually.

6.4.2 Attack surface

6.4.2.1 AI effect on impact and likelihood

As discussed in clause 6.2, the use of AI in an offensive context is not likely to increase the attack surface, or if it does, only because AI is being used to attack vulnerabilities in AI components. It can, however, increase the likelihood of attacks being prosecuted at all and/or successfully. This restates the assertion that the impact of an attack is immutable irrespective of the mode of attack, but the likelihood of an attack will change over time. AI is a vector to modify, mostly by increasing, the likelihood of an attack.

EXAMPLE: The well-known existence of deepfakes, e.g. a video, can affect trust models in that the video as a form of validation may not be as trustworthy as in the past.

The typical machine learning lifecycle is given in figure 4. All elements of the pipeline and their dependencies should be considered in threat modelling. Some of these dependencies will be specific to AI. However, traditional dependencies, for example the code in popular ML libraries and model serving frameworks, should also be considered.

6.4.2.2 Data acquisition and curation

As previously described, compromise of training data represents a compromise of the system itself as the training data represents a core asset of the AI's performance. This is the case regardless of whether by "training data" it is meant the data used to create the initial model, or any additional data used to fine-tune or retrain a model after deployment.

Compromise of data can be achieved via manipulation of data points themselves and/or their labels (and potentially their ordering, see ETSI GR SAI 002 [i.21]), or by injection of extra data samples, and can be carried out in a targeted or indiscriminate way. Dataset modification in any form can be carried out at any part of the data lifecycle: as data is collected (for sensor input can be manipulated), stored, processed or transferred. Further discussion of data supply chains, their vulnerabilities and methods to detect and mitigate attacks are given in ETSI GR SAI 002 [i.21].

6.4.2.3 Implementation

The training phase is where any offline manipulation of training data becomes encoded into a model i.e. where dataset compromise becomes model compromise. In reinforcement learning, an adversary can modify the environment in which the RL agent is learning, in order to cause it to learn an incorrect or suboptimal policy.

Logic corruption attacks target the code used to generate or train a model, maliciously editing it to change the way that the algorithm learns and hence behaves. This mode of behaviour is not unique to AI: any software can be targeted to change the way a downstream system behaves, and the threat can be mitigated by following standard software supply chain security practises. However, it should be emphasized that the effect of the attack will not be felt just within the system where the code is run, rather wherever the poisoned artifacts of that code (i.e. the model) will be deployed.

Similarly, models are increasingly trained in one environment and deployed in another, either as is or with some modifications. This approach, known as transfer learning, is particularly common in applications requiring vast amounts of data and/or compute to generate, for example natural language processing. If an upstream or pre-trained model is poisoned, the subsequent effect can be transferred into different environments, and is not necessarily removed by fine tuning with local data. As such, the model supply chain should be considered and assured alongside the supply chains for data and other software.

Once a model is deployed, it can be refined or retrained using new data points, labels, or other feedback from the environment, including user interactions. This data should be handled with as much care and caution as the original training data.

6.4.2.4 Deployment

Ultimately, compromise of a model cannot cause harm until that model, or one based upon it, is deployed. If an actor can trigger an incorrect response from a model (e.g. misclassification by inputting adversarial examples or by taking advantage of a backdoor), then they may be able to achieve an effect on the system downstream. In assessing threats to a system, one identifies how changing model outputs cause changes in behaviour within a system, and what effects those different behaviours ultimately have on users and the physical world.

As described in clause 6.4.1, a malicious actor can infer information about training data or the model itself by interacting with that model once deployed. Malicious user interactions can also be incorporated into a model, as described above.

6.4.2.5 Humans

The human interaction with, and understanding of, AI is a factor that is worth emphasizing in this context. As previously mentioned, AI is a growing field, and understanding of AI is likely to be less widespread than for other software paradigms. This lack of understanding can affect threats and potential outcomes.

As previously noted, an adversary can attempt to damage an AI component not because of any downstream effect per se, but to damage trust in the overall system or technology.

EXAMPLE: An attacker wishes to flood a human security analyst with false positive results, damaging their trust in AI analytics.

Conversely, another potential threat is that of overconfidence in predictions generated by any model. As previously described, a weak AI model is effective only within the bounds of the problem for which it has been trained, and its outputs are based on probabilities. For example, an AI-based antivirus only classes as malicious malware from families appearing in its training set. A human user, not realizing this limitation, can then assume that novel malware will also be caught.

6.4.3 Trust model

Trust is at the root of security in ICT systems. In symmetric relationships for example, there is trust that Alice does not share with Eve any secrets at the heart of her relationship with Bob.

Several classes of actors are relevant to a deployed ML-based system as shown in table 3.

Table 3: Classes of actor in deployed ML systems

Actor	Role	Role in an example system (face recognition)
Data owners	Owners or trustees of the environment in which a system is deployed and the data which that system stores and use	An IT organization deploying a face recognition authentication service
System/model providers	Constructors of the system, models and/or algorithms that are used in an AI system	Authentication service software vendors
Training data providers	Provider of any labelled datasets used to (re)train the model at any stage of its development or deployment	Face dataset labelling service
Consumers	Consumers of the service the system provides	Enterprise users in the IT organization
Outsiders	Anyone capable of influencing system inputs at any stage, whether by explicit or incidental access	Insider threats, organization guests, external threat actors

It should be noted that there can be multiple examples of each class of user involved in a given deployment, and that individuals can fulfil multiple roles (a consumer can provide training data, for example). The roles of the data and model providers also should be considered explicitly, given the novel role of data and models within an AI system compared to a traditional software system. As previously noted, interactions with the system may (depending on its function and architecture) significantly impact the system itself. Further discussion of trust boundaries in AI systems is given in [i.22].

6.5 Statistics in AI and ML

In many ML systems, such as those offered in packaged programming suites and covered in ML practitioner education, the underlying approach is based on forms of statistical analysis of large data sets. In ontological terms the role of statistical analysis in ML can be addressed by defining the relationship "is enabled by" from ML to the statistics group. It is possible to dismiss all ML as effectively being statistical analysis, although that misses the intent of ML as a waypoint on the continuum from ANI to AGI and thence to ASI.

As mentioned in clause 6.3, it should be noted that, unlike traditional programming, AI systems are inherently probabilistic as opposed to deterministic. The same inputs will not result in a single, testable output; for example, two models trained on the same data may not be identical, in part this is because randomness factors are introduced to the training or learning algorithms. This makes detecting deliberate misuse more difficult. Similarly, human users may not correctly interpret the results of an ML model if they do not understand the reliance on probability.

Where statistical models are used in assisting classification modelling (e.g. pattern recognition), primary, secondary or tertiary, judgement schemes can be applied. In any judgement-based system independent observers can diverge on how they classify items. Statistical measures such as Cohen's kappa [i.10] can be used to determine the degree of agreement between two independently developed AI when classifying N items into C mutually exclusive categories. Such statistical models can be used as a tool in identifying forms of attack that lead to AI based uncertainty, however detecting misuse when the model is addressed as a non-clear-box is generally infeasible.

7 AI and SAI ontology

7.1 Nouns, verbs, adverbs and adjectives

An ontology is fundamentally a language to describe a domain. For intelligence, and for the particular domain of AI, and then for the action of securing AI, it is useful to move from natural language to the ontological expression in small steps.

Table 4: Use of natural language terms

Term	Modal form	Definition
Intelligence	Mass noun	The ability to acquire and apply knowledge and skills
Security	Mass noun	The state of being free from danger or threat, i.e. a state with attributes
Learn	Verb	Gain or acquire knowledge of or skill in (something) by study, experience, or being taught
Knowledge	Mass noun	Facts, information, and skills acquired through experience or education
Reason	Verb	think, understand, and form judgements logically
Think	Verb	
Thought	Noun	An idea or opinion produced by thinking, or occurring suddenly in the mind
Logic	Mass noun	reasoning conducted or assessed according to strict principles of validity

The distinction of mass noun is important for intelligence and for AI, as the definition states that a mass noun is a noun denoting something that cannot be counted (e.g. a substance or quality), in English usually a noun which lacks a plural in ordinary usage and is not used with the indefinite article (one can say "the knowledge", but not "a knowledge"). In the context of the present document intelligence is not directly quantified which is consistent with the measure of intelligence used in many human societies, where intelligence is measured relative to the wider population, often referred to as Intelligence Quotient (IQ). The IQ term introduces measurement of intelligence as a measure of reasoning. There is, from this definition, no direct measure of intelligence, rather it is measurable only by its relation to another intelligent entity.

NOTE: For the purposes of the present document IQ is viewed as a number representing a person's reasoning ability (measured using problem-solving tests) as compared to the statistical norm or average for their age, taken as 100.

Learning is one means of achieving intelligence. Intelligence is a pre-requisite of learning. This may appear to be a circular dependency but it is one that underpins machine learning. Learning builds the knowledge base, intelligence refines the knowledge gained. Systems apply the knowledge using rules and mores also achieved through learning. In normal human discourse, there are multiple, shared, knowledge bases. Contamination of any one base has an indirect impact on other bases. If AI systems mimic human intelligence structures, then this creates side-channel attack vectors.

7.2 Taxonomy and ontology

A taxonomy is a means of classification, and an ontology is an extension of a taxonomy by the inclusion of how things are related across classifications. Thus, in the security domain a taxonomy will often identify authentication methods within a taxonomical class, or when reporting vulnerabilities using the CVE or CVS approaches, or when expressing weaknesses using the CWE approach [i.17].

EXAMPLE: The Common Weakness Enumeration project maintains a long list of weaknesses and illustrates their relationships (parents, children, peers and so on). The CWE project is an extended taxonomy with several types of classification.

A more detail summary of the CWE project examining a single weakness, the "CWE-79: Improper Neutralization of Input During Web Page Generation ('Cross-site Scripting')" variant follows (from http://cwe.mitre.org/data/definitions/79.html):

- CWE-Class → An abstract view of the weakness lying between a Pillar Weakness (more abstract) and a Base Weakness (more specific)

- CWE-Variant → Specific to a particular technology, e.g. an XSS weakness that is specific to a particular browser or server
- CWE-Chain → An element that links weaknesses from different classes together in an OR logical arrangement (at least one element in the chain has to be present)
- CWE-Composite → A join of two or more weaknesses in an AND logical arrangement (all weaknesses have to be present)
- CWE-Base → A weakness mostly independent of technology/implementation
- CWE-Category → A container class with a set of other entries showing a common characteristic

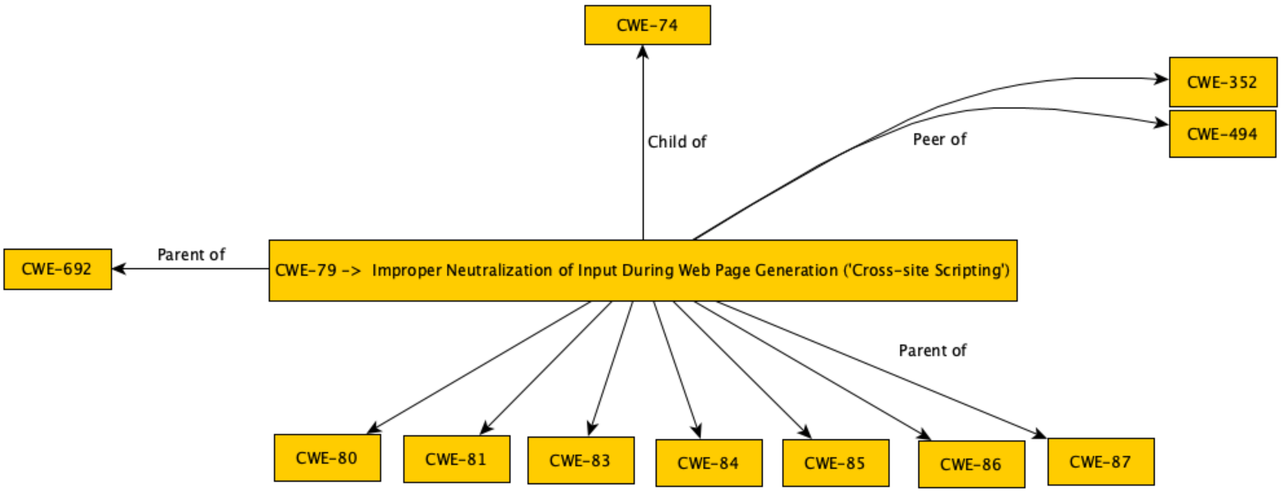

Figure 9: Illustration of CWE-79 in the form of a taxonomy or hierarchy

7.3 Core SAI ontology relationships

Security can be understood within a wider ontology representing the state of relations between objects that symbolize, in broad terms, the Confidentiality Integrity Availability (CIA) paradigm. As such, attacks that undermine the relationships inherent in that paradigm are at the heart of understanding the role of securing AI.

The core concepts of a security ontology and taxonomy already exist in ETSI TS 102 165-1 [i.5] and to a lesser extent in ETSI TS 102 165-2 [i.20]. This has already been captured in figure 8, although not expressed as an ontology. It is also recognized that ETSI TS 102 165-1 [i.5] and ETSI TS 102 165-2 [i.20] identify taxonomies for many forms of countermeasure.

EXAMPLE: In ETSI TS 102 165-2 [i.20] the authentication countermeasure has specializations for cryptographic authentication using a challenge response protocol to prove knowledge of a shared secret, and for digital signature approaches when an asymmetric system is used.

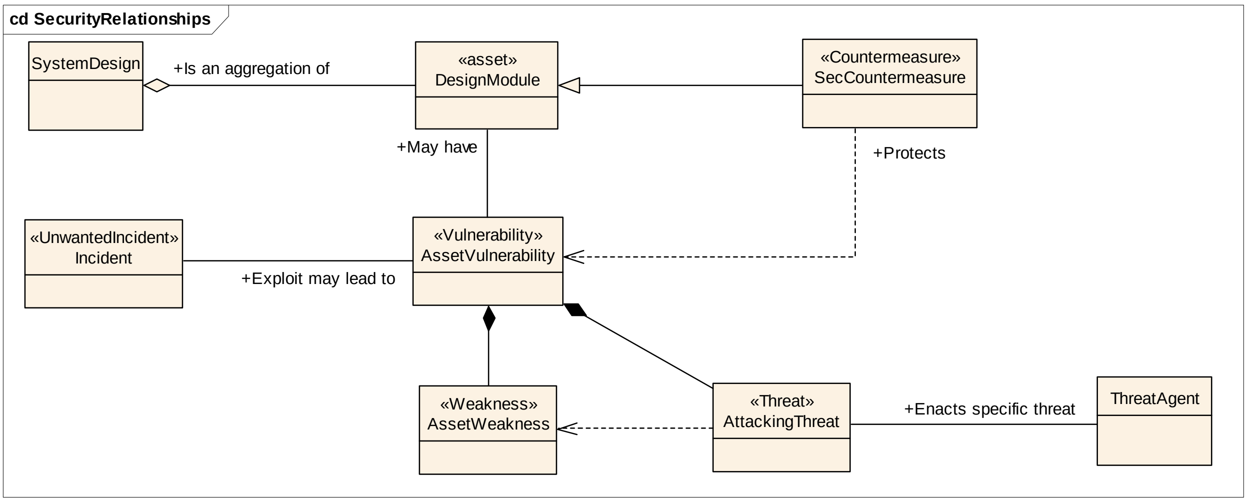

Figure 10: Model of threat from TS 102 165-1 [i.5]

The approach of RDF requires statements in the form:

```
<subject> <predicate> <object>
```

This is given in the example of figure 8 as:

```
ThreatAgent → Enacts → Threat,
```

or in reverse

```
Threat → is enacted by → ThreatAgent.
```

The present document is not intended to re-write the content of ETSI TS 102 165-1 [i.5] and ETSI TS 102 165-2 [i.20] in ontological format although such an exercise is strongly recommended.

Table 5: RDF language constructs

Construct	Syntactic form	Description
Class (a class)	C rdf:type rdfs:Class	C (a resource) is an RDF class
Property (a class)	P rdf:type rdf:Proterty	P (a resource) is an RDF property
type (a property)	I rdf: type C	I (a resource) is an instance of C (a class)

In addressing these RDF constructs Confidentiality, Integrity, Availability (including Authenticity) (the CIA paradigm) are properties of systems or of objects in the systems. AI is in this model also a characteristic of the attacker, i.e. a property of the attack, where an attack is an instance of a threat agent enacting a threat.

An AI agent is modelled as an instance of the Threat Agent when in adversarial mode. The classes of the AI enhanced threat model are then as shown in figure 8 (threat, threat agent, countermeasure, and system asset are all classes).

Thus:

- An AI threat agent is an instance of a threat agent.

In a defensive mode an AI-enabled countermeasure is modelled as an instance of a countermeasure.

- An AI-enabled countermeasure is an instance of a countermeasure.

The significant difference to the simple model given in figure 8 is the learning and reactivity of the AI-enabled threat agent, or the AI-enabled countermeasure. Thus, the AI-enabled entity can observe the impact of the attack and modify the attack or the defence accordingly. This does not alter the core semantic relationship of threat agent to threat, or of a vulnerability being identified by a threat/threat-agent.

Observation is then added as a property of the AI-enabled threat agent and of the AI-enabled countermeasure. This expands upon the model of data learning given in clause 5:

1) data gathering;

2) data processing;

3) applying insights gained from the data processing.

In adversarial systems the data gathering can be applied to the rate of success of attacks and the way that attacks are countered. Data gathering can also be used to develop the ability of the threat agent by examination of the system as a whole.

NOTE: In an AI-enabled system for either adversarial or defensive application the core threat/threat-agent and countermeasure have to be adaptable. In other words, if there is only one way to apply the attack and it can be successively defended then no amount of intelligence can overcome the rigidity of the attack in that scenario.

7.4 Device based ontologies

The core ontologies for devices can be found in the oneM2M Base Ontology [i.18] and in the closely related SAREF Ontology [i.19]. In each case AI can be treated at the function and service level to enhance the function or service, in each case by monitoring the effect/affect of the function or service.

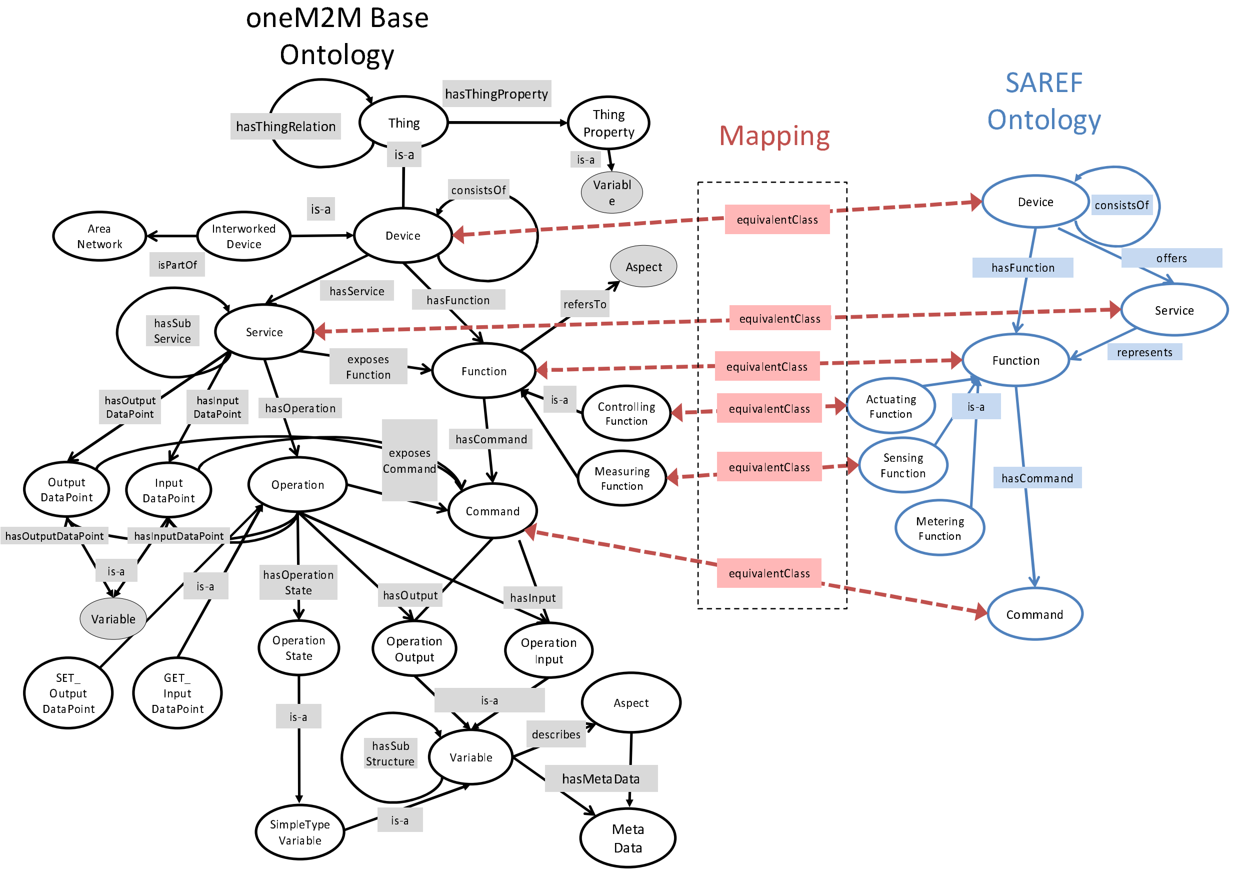

Figure 11: Mapping between SAREF and the oneM2M Base Ontology (from ETSI TS 118 112 [i.18])

With respect to existing machine ontologies such as SAREF and the oneM2M base ontology any ontology for SAI is going to look somewhat different as SAI refers to more than just devices.

7.5 Recommendation for further work

The important point of the present document is that AI in threat analysis, in security detection and prevention, and in security reporting, already exists in implicit ontological form. Many of the simple relationships such as that at the core of ETSI TS 102 165-1 [i.5], "A ***threat agent*** enacts a specific ***attack*** against a system ***weakness*** to exploit a ***vulnerability*** ", has a number of relationships between systems, system elements (the assets of the system), the weaknesses and the vulnerabilities which are already ontological in nature. In addition, it is recognized that most measures are already well described in taxonomical form, such as in ETSI TS 102 165-1 [i.5] for consideration of attack trees.

In the use of data sets as training data, and as live analytic input, the role of ontologies in the form of semantic and relational labelling of data is recognized in enabling AI. Expanding the use of ontological approaches to maximize the exchange of data with each of syntax, semantics, pragmatics and relationships made explicit, enables the application of AI. However, of itself AI is not identified as having characteristics unique to the artificial nature of the intelligence, but rather expands the characteristics of intelligence (reasoning, representing knowledge, planning, learning, communicating, integrating, sensing, acting) to non-biological entities.

In expressing data in systems adoption of a more complete ontological like approach should always be undertaken. Whilst improving interoperability by addressing syntax and semantics, the addition of pragmatics and relationship definitions as per an ontology expresses intent and purpose much more fully and should lead to less error in interpreting data.

Annex A: Cultural origins of ICT based intelligence

There is some debate regarding who first considered the role of machines as intelligent entities, but the broad consensus is that Alan Turing is the father of machine intelligence from his paper "On computable numbers, with an application to the Entscheidungsproblem" [i.1] and its demonstration of an abstract logic that has become known as a "Turing machine", and later in his paper "Computing Machinery and Intelligence" [i.2]. The test of machine intelligence, also from Turing (the Turing Test), is one in which a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human, suggests that the machine is intelligent. Turing's experiment was based on a game, the "Imitation game" in which the computer is open to lie or tell the truth, a human (the opponent), can only tell the truth, and a 3rd party asks questions to them both and based on the responses attempts to determine which is the human and which the computer. There are various rules to inhibit data leakage such that the determination can only be based on the answers to the questions. The same basic form was used in Philip K. Dick's "Do androids dream of electric sheep?" [i.3] in which the determination of human versus machine was based on the way in which the subject responded to questions.

NOTE 1: It is recognized that the fictional Voight-Kampff test of Philip K. Dick's vision is not directly comparable to Turing's original test but the premise that the evaluator, Deckard in the book, equivalent to party C of Turing's test, has the experience and skill to make the distinction is a failure of both tests. In practical application the machine only has to be better than the evaluator which is a lower barrier to success for the machine to be recognized as human.

The general definition of intelligence is somewhat simpler, "the ability to acquire and apply knowledge and skills", and is not comparative. However, it is also normal to consider cognition, "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses", as a synonym of intelligence. How much is understood regarding intelligence in humans is often difficult to gauge but the actions of neurons and their interconnection through synapses is increasingly being understood with their ability to change the weight of connection over time. Thus, whilst Turing's Universal Computing Engine is a binary system the system of neurons and synapses is more akin to an analogue system, or certainly non-binary.

NOTE 2: The connection linking neuron to neuron is termed the synapse with signals flowing in one direction, from the presynaptic neuron to the postsynaptic neuron via the synapse which acts as a variable attenuator giving weight to the signal (either chemical or electronic).

Turing's view was that a machine did not need to exist but that it could exist. It may be reasonably stated that the existence of neural processors, neural networks, and the near infinite resources of memory and processing available through cloud computing systems that such machines do now exist.

In contrast Ada Lovelace, who along with Charles Babbage, developed the concepts behind modern computer programming, stated that "*The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths.*" In some respects, Ada Lovelace's concern regarding machine intelligence fits into a placement of computing (the Analytical Engine is one example of an early computing engine) as a machine able to implement combinational logic, or to act as a finite state machine, in the automata theory (see figure A.1), where such levels of computation do not anticipate the potential of the Turing Machine.

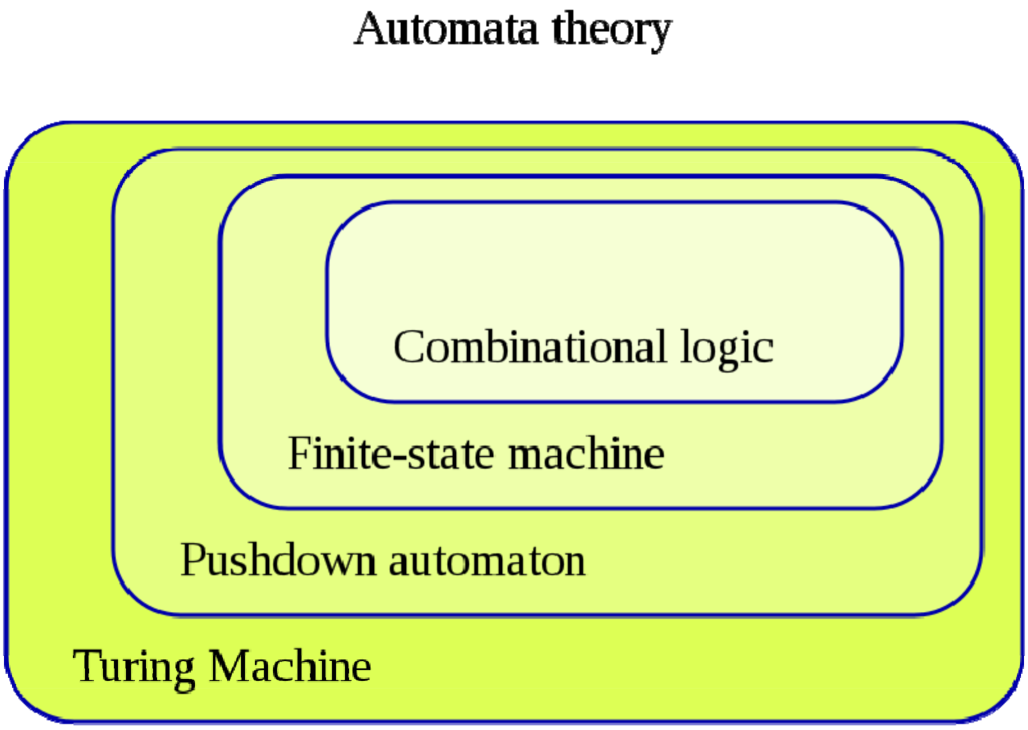

Figure A.1: Automata theory of machine logic (diagram released under Creative Commons license)

There are a number of ways of interpreting how computing engines may exhibit or gain "the ability to acquire and apply knowledge and skills". In normal human development the acquisition of knowledge or skill requires data, test and application.

In other parts of popular culture Asimov proposed three rules of robotics that may be assumed to apply to any form of near sentient AI. To avoid any issues of copying copyrighted material. Refer to the text of i-Robot [i.4], although the rules can be found cited in many locations.

It is possibly useful to note that Asimov proposed the rule of robotics as a moral compass, and that the rules are akin to an ethics framework (do no harm to others, do not let others come to harm, do not allow yourself to be harmed), where the ordering of the rules is critical and "others" is meant to mean humans but is generalized here.

NOTE: If Asimov's robotics rules are inversed such that the self-preservation rule is dominant it allows the machine to allow harm to be caused to others. In the "correct" order self-sacrifice is allowed to prevent harm to others.

In the forming of many ethical or moral rulesets this form of ordering is critical.

Annex B: Machine processing to simulate intelligence

B.1 Overview of the machine intelligence continuum

NOTE: This annex is a guide and readers are also advised to consult the sources in the Bibliography and references of the present document to gather a full picture of the development and application of machine intelligence.

The use of data to build up intelligence can be represented as a continuum in terms of how data is interpreted.

At the far-left of the continuum are expert systems: Guided paths through data. Early forays in AI were often in the form of expert systems, and can be typified as a tree of questions of an "if … then … else" structure, where the enquirer was led to an answer based on their responses to specific questions. The future stages in the continuum presented in figure B.1.

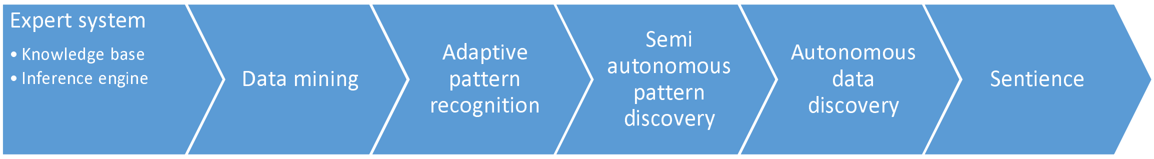
Expert system
• Knowledge base
• Inference engine
Data mining
Adaptive pattern recognition
Semi autonomous pattern discovery
Autonomous data discovery
Sentience

Figure B.1: Continuum of intelligence, one possible model

In figure B.1 it should be noted that sentience is not a simple achievement that would be a trivial next step from autonomous data discovery, rather it has to be recognized that there are substantial intellectual and engineering barriers to overcome between each stage, and that all of these will co-exist for all time.

B.2 Expert systems

The founder of much of the early work on expert systems, Edward Feigenbaum, postulated "intelligent systems derive their power from the knowledge they possess rather than from the specific formalisms and inference schemes they use".

Expert systems are not autonomously intelligent, rather their observed intelligence is captured in the design of the inference systems. A crude simplification of an expert system is one in which a question is offered with a finite set of possible answers, such as in fault diagnosis, and based on the answer to the first question a second and then a third and further layers of questions and answers lead to a diagnosis. The "intelligence" is in the construct of the questions and answers and there is no processed intelligence.

B.3 Data mining and pattern extraction

The statistical analysis of large volumes of data to match patterns, or to discover patterns is the forefather of much of ML (see also clause 6.5 of the present document). The application of significant levels of processing power and memory to such statistical analysis led to the ability to sift through data quickly, in addition the introduction of alternative database models allowed changes in data analysis to be developed. Conventional database wisdom of relational databases with highly structured search patterns was challenged by offering pooled or unstructured data to be searched using statistical approaches. Adding self-described data into the mix, using semantic labelling and similar advances, allowed data resources to be searched or analysed without the structural rigidity of conventional relational databases.

Annex C: Bibliography

C.1 AGI analysis

[A] L Chen, Z Yi, X Chen: "Research on Network Security Technology Based on Artificial Intelligence", - Recent Trends in Intelligent Computing…, 2020 - Springer.

[B] W Wu, T Huang, K Gong: "Ethical Principles and Governance Technology Development of AI in China"; Engineering, 2020 - Elsevier.

[C] R Girasa: "International Initiatives in AI", Artificial Intelligence as a Disruptive Technology, 2020 - Springer.

[D] FB Pizzini, F Pesapane, W Niessen: "ESMRMB Round Table report on "Can Europe Lead in Machine Learning of MRI-Data?"", 2020 - Springer.

[E] A Abuarqoub: "D-FAP: Dual-Factor Authentication Protocol for Mobile Cloud Connected Devices", Journal of Sensor and Actuator Networks, 2020 - mdpi.com.

[F] W Hoffmann-Riem: "Artificial Intelligence as a Challenge for Law and Regulation", Regulating Artificial Intelligence, 2020 - Springer.

[G] D Feldner: "Designing a Future Europe", Redesigning Organizations, 2020 - Springer.

[H] SM Lee, SC Jeong: "A study on strategy for invigorating utilization of HPC in industry based on business building blocks model", Nonlinear Theory and Its Applications, IEICE, 2020 - jstage.jst.go.jp.

[I] R Girasa: "Bias, Jobs, and Fake News", Artificial Intelligence as a Disruptive Technology, 2020 - Springer.

[J] J Schemmel: "Artificial Intelligence and the Financial Markets: Business as Usual?", Regulating Artificial Intelligence, 2020 - Springer.

[K] RK Saini, C Prakash, A Dua: "A Survey on Artificial Intelligence Techniques for Cybersecurity", Current Trends in …, 2020 - computerjournals.stmjournals.in.

[L] RT Kreutzer, M Sirrenberg: "Fields of Application of Artificial Intelligence - Security Sector and Military Sector" Understanding Artificial Intelligence, 2020 - Springer.

[M] S Saif, S Biswas: "On the Implementation and Performance Evaluation of Security Algorithms for Healthcare", Proceedings of ^{t}he 2nd International Conference on …, 2020 - Springer.

[N] X Liu, Y Lin, H Li, J Zhang: "A novel method for malware detection on ML-based visualization technique", Computers & Security, 2020 - Elsevier.

[O] T Chen, J Liu, Y Xiang, W Niu, E Tong, Z Han: "Adversarial attack and defense in reinforcement learning-from AI security view", Cybersecurity, 2019 - Springer.

[P] Russell, Stuart J.; Norvig, Peter (2003): Artificial Intelligence: A Modern Appro^{ac}h (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2.

[Q] Luger, George; Stubblefield, William (2004): Artificial Intelligence: Structures and Strategies for Complex Problem Solving (5th ed.), The Benjamin/Cummings Publishing Company, Inc., p. 720, ISBN 978-0-8053-4780-7.

C.2 AI in the context of threat analysis

[R] "Threat Modeling AI/ML Systems and Dependencies", Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar; November 2019.

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

[S] N. Papernot, P. McDaniel, A. Sinha and M. P. Wellman, "SoK: Security and Privacy in Machine Learning", 2018 IEEE European Symposium on Security and Privacy (EuroS&P), 2018, pp. 399-414, doi: 10.1109/EuroSP.2018.00035.

NOTE: Available at https://ieeexplore.ieee.org/document/8406613.

History

Document history		
V1.1.1	January 2022	Publication