ETSI GR SAI 007 V1.1.1 (2023-03)

Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing

Reference

DGR/SAI-007

Keywords

artificial intelligence

ETSI

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

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

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

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

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

Foreword

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

Modal verbs terminology

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

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

1 Scope

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

NOTE: The present document uses the term explicability but recognizes that many other publications use the term explainability. The terms are interchangeable with the proviso that the latter term is not a commonly accepted UK English word.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

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

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

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

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

[i.2] ETSI GR SAI 002: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

[i.3] ETSI GR NFV-SEC 003: "Network Functions Virtualisation (NFV); NFV Security; Security and Trust Guidance".

[i.4] Auguste Kerckhoffs: "La cryptographie militaire" Journal des sciences militaires, vol. IX, pp. 5-83, January 1883, pp. 161-191, February 1883.

[i.5] ETSI GR SAI 001: "Securing Artificial Intelligence (SAI); AI Threat Ontology".

[i.6] COM/2021/206 final: "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts".

[i.7] DARPA eXplainable AI project summary.

[i.8] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. Conference on Fairness, Accountability, and Transparency: "Model Cards for Model Reporting", January 29-31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA, 10 pages.

[i.9] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K. R. (eds.) (2019): "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning". Cham, Springer.

[i.10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford: "Datasheets for Datasets" (Commun. ACM 64, 12 (December 2021), 86-92.

[i.11] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. R. (2019): "Unmasking Clever Hans predictors and assessing what machines really learn". Nat. Commun. 10, 1-8. doi: 10.1038/s41467-019-08987-4.

[i.12] Molnar, C. (2022): "Interpretable Machine Learning-A Guide for Making Black Box Models Explainable".

[i.13] Samek, W., Montavon, G., Binder, A., Lapuschkin, S., and Müller, K. R. (2016): "Interpreting the predictions of complex ML models by layer-wise relevance propagation", arXiv abs/1611.08191.

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.1] and the following apply:

explicability: property of an action to be able to be accounted for or understood

transparency: property of an action to be open to inspection with no hidden properties

3.2 Symbols

Void.

3.3 Abbreviations

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

AI	Artificial Intelligence
BTT	Build-Train-Test
DARPA	Defence Advanced Research Projects Agency
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
OECD	Organisation for Economic Cooperation and Development
RTE	Run Time Explicability
TA	Trust Association
XAI	eXplainable AI

4 Explicability and transparency

The SAI problem statement [i.1] identifies explicability as being a contributor in establishing trust in AI systems as one element of achieving transparency. However, in computer science the concept of transparency is somewhat at odds with explicability and can be interpreted as "*functioning without the user being aware of its presence*" when referring to a process. The term transparent (and its associated noun form, transparency) when applied to AI is, for the purposes of the present document, the core concept of being open to examination, or having no part hidden.

The term explicability is, in very crude terms, being able to show how any result was achieved ("*show your working*"), which when combined with transparency gives assurance that nothing is hidden.

NOTE 1: In ETSI GR SAI 004 [i.1] the term explainability is used whereas in the present document the more common term in UK English, explicability, is used.

NOTE 2: It is recognized that many processes are protected from disclosure by mechanisms that protect the intellectual property that the processes contain and such protections are not intended to be impacted by the requirement to maintain attributes of transparency and explicability.

The outcome of applying constraints of explicability and transparency to systems is that trust can be conferred as a system attribute that is open to examination and verification by 3rd parties.

It is recognized that in many systems, such as in telecommunications, the role of AI is often at a component level. The role of most applications is not to explicitly design or develop intelligence as a primary goal.

One purpose of transparency and, particularly, explicability is to prevent the AI components of a system from denying that they took part in an action, and to prevent the AI component denying they were the recipient of the output of an action from any other part of the system.

NOTE 3: The description above is very close to the common definition of non-repudiation but there is a subtly different intent in the scope of explicability and transparency, hence for the present document this is not referred to as non-repudiation.

In ETSI GR SAI 001 [i.5], it is stated that there are a number of characteristics associated to intelligence the key elements of which are given below, and in the context of transparency and explicability it is expected that each of these characteristics, if they are present in the AI component or system, is described.

- **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;
- **learning:** the means by which reasoning and other behaviour evolves over time to address new input;
- **communicating:** in natural language (to human third parties), in particular when within the bounds of the system it is unable to process data to a known state.

In terms of explicability it should be clear where reasoning takes place, and on what data and algorithm, such reasoning is based. Similarly the scope of explicability and transparency addresses the means by which the system learns. Finally, in the context of the key characteristics above, the means by which the system's purpose is communicated should be in natural language where the intended recipient should be considered as a lay person (i.e. having no knowledge of any specialized language of AI/ML or of the programming techniques of AI/ML).

Many concerns raised regarding AI/ML (see ETSI GR SAI 004 [i.1]) and addressed as "Design challenges and unintentional factors" can be made visible through the application of specific explicability techniques. An example is the concern of bias (confirmation bias and selection bias in particular) where, by the application of simple checklists (see clauses 5 and 6) the system deployment should be able to answer questions of the form "why was this data source selected?".

EXAMPLE: An AI can be biased by design if the purpose of the AI is to filter candidates for a job based on some personal characteristic (i.e. as opposed to a meritocratic selection engine, the AI acts as a characteristic selection engine). In such a case the explicability and transparency requirements will be able to identify that negative, or trait-based, filtering is at the root of the reasoning engine of the AI.

It is reasonable to suggest that bias in inputs will be reinforced in the output, hence in clause 5 it is stressed that explicability addresses the purpose of data. If data is preselected to achieve a particular result that could be seen to be consistent with selection bias and that would need to be explained as part of the system purpose (as in the example) or removed by design.

5 Static explicability analysis

5.1 Summary of the role of static explicability analysis

The role of static explicability is closely related to giving detailed system documentation. The purpose of explicability is to allow a lay person (i.e. not a professional programmer or system analyst) to gain a reasonable understanding of the main data flows and processing steps in the program.

EXAMPLE: A data set of images is used as training data and routinely classified as images of, say, "Cat", "Dog", "Fox", "Badger" where the purpose is to enable a camera observing a suburban garden to record movements of particular animals at night, thus being able to say that a badger crossed the garden lawn at a particular time of the night.

In a simple scenario such as in the example above the purpose is clear (identify which animal is in the capture range of the camera), it is clear where the training data comes from (the set of images), and it is reasonable to expect a layperson to understand the purpose, the role of data and components in the system, and to make reasonable attempts to verify the veracity of the system (e.g. by getting a dog to pass in front of the camera and be recognized as a dog, or for a deer to pass in front of the camera and not to be recognized as one of the animals it has been trained to recognize).

As more components are added to the system to improve the system's ability in recognition, say by adding gait analysis (dogs and cats move quite differently) static explicability should be maintained.

The components identified in table 1 should be clearly identifiable in the system documentation.

Table 1: System documentation elements in static explicability analysis

Documentation Element	Element	Short description
1	Statement of system purpose	This element of the system documentation is intended to allow a layperson to clearly understand the purpose of the system and to explicitly identify the role of AI in achieving that purpose.
2a	Identification of data source(s)	Where the data comes from and how the authenticity of the data source is verified.
2b	Purpose of data source(s) (in support of system purpose)	The role of the particular data source in the system (e.g. training data containing images of dogs to train the system in recognizing a dog from an image)
2c	Method(s) used to determine data quality	Methods and processes used in determining if the input data is a fair and accurate representation of the desired input. This should address how bias or preference is identified and corrected in the data input.
3	Identity of liable party	For each processing or data element a means to identify liability for correction of errors or for maintenance of the element.

5.2 Advice on documenting the statement of system purpose

The statement of system purpose is critical in allowing a layperson to clearly understand the intent of the system and the role of AI in achieving that purpose or intent.

EXAMPLE 1: AI used in a voice-recognition personal assistant. The purpose of the system is to allow the user to issue spoken commands in natural language and to translate those into machine commands for purposes including machine control, and internet-based information search and retrieval. The AI in the system provides a number of functions in order to achieve its purpose including: AI to enable speech recognition; AI to assist in parsing of recognized speech to commands; AI to drive voice responses to spoken commands; AI to parse and relay the results of search commands into natural language.

EXAMPLE 2: AI used in adaptive cruise control in road vehicles. The primary purpose is to ensure that whilst the driver can set a target speed to be maintained it is recognized that strict adherence to the target speed can be unsafe. The role of the AI in this system is to maintain a safe distance between vehicles whilst maximizing the time spent at the target speed. The system therefore adaptively modifies the vehicle speed (not exceeding the target speed) by maintaining a "safe" distance from other vehicles through selective braking and acceleration where data on the presence and actions of other vehicles are obtained from system sensors and driver input.

The statement of system purpose should be written in natural language and be concise as well as precise (i.e. not open to variations in interpretation).

5.3 Methods in documenting the identification, purpose and quality of data sources

As outlined in table 1 where data is used in AI the liable party should ensure that answers are documented for the following questions:

- Where does the data come from?
  - As the purpose of data has been indicated earlier this clarifies explicitly the source of the data. This can include statements such as the following for the example of adaptive cruise control: "the range-data indicating the distance to surrounding vehicles and environmental objects is sourced from a radar array positioned at the front left, centre and right of the vehicle".
- How is the authenticity of the data source verified?
  - The aim here is to ensure that only trusted data (data sources) are used in the system
- What is the role of the particular data source in the system? (e.g. training data containing images of dogs to train the system in recognizing a dog from an image)
- What methods and processes are used in determining if the input data is a fair and accurate representation of the desired input?
- What steps have been taken to determine if the input data has bias?
  - It can be argued that all data is biased and that all designers will have some degree of selection bias in the data chosen to train and run their systems. However it is essential that designers be as objective as possible when documenting their sources. If similar data sources were available it may be necessary for the designer to show why one source was selected over any alternatives (e.g. for reasons of cost, or trust in the source as opposed to the content).
- What steps have been taken to compensate for any bias in the input?
  - As has been noted bias can be a design decision. In many instances it may not. Bias can be compensated in a number of ways including modification of data ranking or direct modification of the source to remove inherent bias. Any steps taken to compensate for bias should be documented in clear, concise, and precise natural language.

The use of Model Cards outlined in [i.8] performs much of the above role and where in [i.8] it is stated that there are no standardized documentation procedures to communicate the performance characteristics of trained Machine Learning (ML) and Artificial Intelligence (AI) models the approaches outlined in the present document and those in [i.8] are part of closing that gap in standardization. In addition, the use of datasheets as outlined in [i.10] provides a means to facilitate communication between dataset creators and consumers that is consistent with the intentions of the present document.

5.4 Identifying who is the liable party

In undertaking analysis and in providing the necessary documentation it should be made clear who is responsible for the AI system, and the system of which it forms a component. This should be consistent with any other obligations when placing products on the market.

6 Run time explicability

6.1 Summary of service

When an AI system is running it applies its AI to data to achieve its purpose. The goal of run time explicability is to ensure that the system developer, and other stakeholders in the supply chain, can identify the role of active processes, and data, in achieving the system purpose.

Static explicability is a pre-requisite to run-time explicability. Run Time Explicability (RTE) is defined in the present document as an explicit service of a running system.

The goal of the explicability service is to collect, maintain, make available and validate irrefutable evidence concerning the purpose of, and data contributing to, an action of the machine in order to assist in determining the validity of the action at the time it was taken.

NOTE: The explicability service is closely related to conventional non-repudiation services but with the intent of explaining actions rather than for solving disputes (see also clause 4).

6.2 Abstraction of AI system

An abstract model of an AI processing system is given in ETSI GR SAI 004 [i.1] from which figure 1 is taken to represent stages in the ML lifecycle.

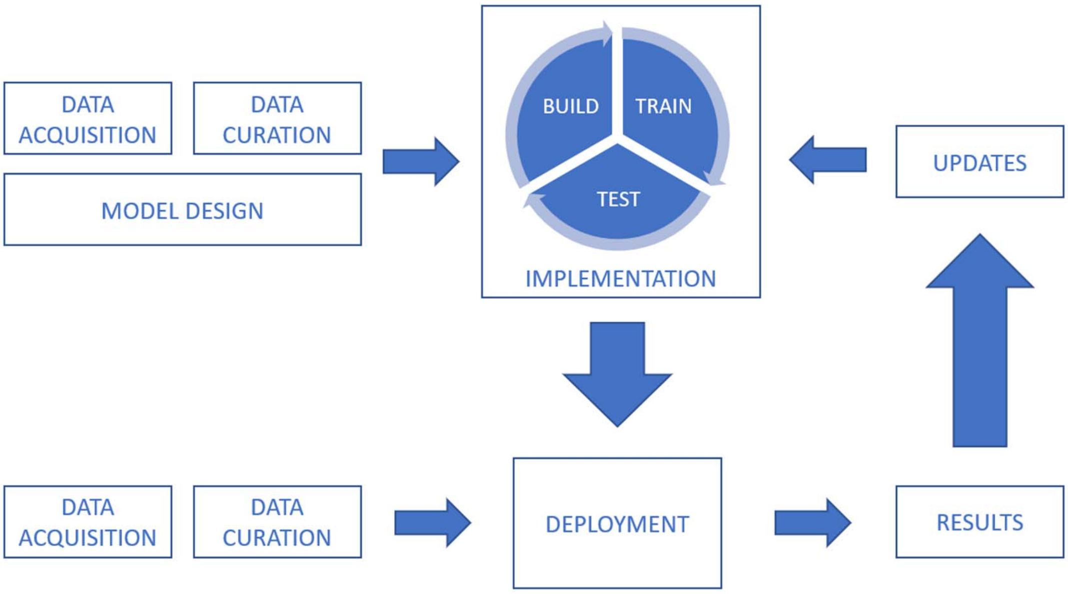

Figure 1: Typical machine learning lifecycle (from [i.1])

Explicability applies to the Build-Train-Test (BTT) cycle during model design, and to the role of the update cycle during deployment that supplements the BTT cycle.

6.3 Evidence requirements for explicability

The requirements for static explicability, outlined in clause 5, apply as a pre-requisite to providing evidence for run-time explicability.

As indicated above, explicability (and transparency as a pre-requisite) aims to prevent the AI components of a system from denying that they took part in an action, and to prevent the AI component denying they were the recipient of the output of an action from any other part of the system. The RTE service expands on the set of questions outlined in clause 5.3 and summarized below:

- What process does data undergo between acquisition and curation?
  - The lifecycle shown in figure 1 identifies data acquisition and curation used in development of the model that is used in implementation (following a BTT cycle), and also in the active deployment phase where results are used in feedback to refine the implemented model. It is reasonable to filter data between acquisition (say where multiple data sources are used) and its curation (say by removing fields from data sources where those fields are not relevant to the model).

- What are the metrics that determine change in the learning/weighting of data?
  - Notwithstanding any intention by the designers to open intellectual property embedded in the feedback and feedforward learning process it should be made clear to the user of the system what is involved in the learning process.

6.4 Performance considerations

An AI/ML system can make decisions at a rate that, if a detailed evidential record was to be created, and retained securely, has potential to overload the system. Rather than take a detailed evidential record for every decision the goal of explicability and transparency is to ensure that the rationale for a decision is clear.

In addition to issues related to performance from audit, the designer should also be able to define the expected accuracy of the system. This can be achieved by explicitly identifying the measure of precision and of recall against both static data and live data.

- Precision, the measure of positive predictive value, measures the correctness of the decision every time the model made a positive decision. Precision can only be reliably measured against a known input (the number of relevant elements in any sample is known).

```
Precision = Number of true positives / (number of true positives + number of false positives)
```

- Recall is the measure of overall success at identifying relevant elements. As for precision, recall can only be reliably measured against a known input.

```
Recall = Number of true positives / (number of true positives + number of false negatives)
```

EXAMPLE 1: An AI system is designed to recognize dogs in an image (dogs are the relevant elements). If the system is presented with an image that contains ten cats and twelve dogs (i.e. there are 22 identifiable animals in the image), and the system identifies eight dogs, of the eight elements identified as dogs, only five actually are dogs (true positives), while the other three are cats (false positives). Seven dogs were missed (false negatives), and seven cats were correctly excluded (true negatives). The program's precision is then 5/8 (true positives/selected elements) while its recall is 5/12 (true positives/relevant elements), i.e. precision of 62,5 % and recall of 42 %.

EXAMPLE 2: An AI system is designed to grant people access to a secure building using facial recognition. The system recognizes 150 people and grants access to 100 of them with a precision of 98 % meaning that of 100 people granted access, 2 were not supposed to enter the building. However if in the 150 people recognized there were in fact 120 that should have been granted access the recall rate is 98/120 or only 82 %.

There are many other ways of measuring the system performance using other statistical measures but the key point is that the system documentation should clearly indicate the measure by which the system claims to be accurate. A run-time measure of accuracy should be considered to be developed and implemented as part of an AI system's design.

NOTE: Accuracy can be used as a component in developing trust, see also Annex A.

6.5 Application of XAI approaches

Complementing the approaches presented above, academic research on more complex technical methods for gaining insights into the behaviour and decisions of AI models is performed in the field of eXplainable AI (XAI). Depending on the use case, different methods can be used. An overview of the different approaches is given in [i.9]. When making predictions from structured data, probabilistic methods are generally considered promising [i.12], whereas applications from computer vision rely on more advanced methods such as Layer-wise Relevance Propagation (LRP) [i.13].

Some XAI methods provide global explanations, while others explain individual (local) model decisions. One useful application of XAI methods has been to uncover spurious patterns in data sets learned by AI models and leading to wrong decisions [i.11].

A number of projects have been created under the DARPA XAI [i.7] leadership to address the following aspects of AI as applied to ML:

- produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and
- enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

Whilst the XAI programme is not complete and does not directly produce standards the goals are aligned to both the static explicability analysis (clause 5) and the RTE service (clause 6) of the present document.

It is noted that the XAI program is focused on the development of multiple systems by addressing challenging problems in two areas:

1) ML problems to classify events of interest in heterogeneous, multimedia data; and
2) ML problems to construct decision policies for an autonomous system to perform a variety of simulated missions.

These have been chosen to represent the intersection of classification and reinforcement learning, and also address the intersection of gathered data analysis and autonomous systems.

A third major element of the XAI project is to gain a better understanding of the psychology of explanation which reinforces the intent of the present document to provide the user with greater understanding of the role and scope of AI in systems.

7 Data transparency

ETSI GR SAI 002 [i.2] identifies the role of understanding the data supply chain as a link in integrity and availability assurance. As stated in clause 4 transparency when applied to AI is related to being "*open to examination*". The value of integrity checks, e.g. using cryptographic hashes, in transparency is that they are able to indicate unauthorized change between sender and receiver.

Thus a general requirement for data transparency with respect to integrity is as follows:

- The recipient of data should be able to determine if the data has been manipulated by a 3rd party before receipt (i.e. in the period from the sender releasing data to the recipient receiving it).

In addition to determining integrity the recipient, in support of transparency, needs to determine the source of the data. This may in turn require additional technical measures as follows:

- The recipient of data should be able to identify the source of data.
- The recipient of data should be able to verify the identity of the source of data.
- The recipient of data should be able to verify that the data source has authority to share data with the recipient.

Data transparency in ML systems applies in particular to the Data Acquisition and Data Curation phases, i.e. where the data comes from.

Annex A:
Trust in AI for transparency and explicability

In the context of AI the model of trust that is offered by the AI is part of the overall relationship of the AI and its dependent users. How AI entities build trust is complex and can differ from the trust measures used in simpler, non-AI, systems. In practice a number of security assurance elements are combined to determine an overall trust level. Such elements include identity, attribution, attestation and non-repudiation. In the context of AI a number of objectives for trust apply, alongside transparency and explicability.

The assignment of trust in conventional discourse is the decision that an entity A should trust entity B in one or more particular contexts. Key criteria for assigning trust are:

- The identity of the entity to be trusted.
- The contexts within which the trust should be constrained.

The security relationships of an AI, in addition to countering risks and attacks on the system, are used to reinforce trust relationships. A number of trust models are commonly used in technology:

- Delegated trust
    - entity A is unable to evaluate the appropriate level of trust for a relationship with another entity B, thus entity A can choose to delegate the decision to another entity C.
- Collaborative trust
    - two entities (entities A and C) work together to decide whether to trust another (entity B) - the final goal can be for both entity A and entity C to have a trust relationship with entity B.
- Transitive trust
    - entity A trusts entity B because entity C trusts it.

A more complete description of the role of trust in networks is found in ETSI GR NFV-SEC 003 [i.3].

In the context of an AI the role of trust is somewhat complex as there is not a single root of trust, rather there has to be trust in the process of learning, of data sources, and of the actions taken. The relying party, that is the party dependent on the AI output, should be able to build a trust model of the AI system. There are therefore a number of Trust Associations (TA) in the AI/ML system each with an independent quantitative (and qualitative) assessment of their Trust Value. The metrics for determining the trust value are for further study, but it is considered that the Trust Value assigned to the overall system is given as the (vector) sum of the set of Trust Values of each TA in the system.

$$SystemTV = \sum TrustValue.TA$$

In addition trust can be associated to accuracy (e.g. the combination of precision and recall), or to other metrics associated to the processing.

It should be assumed that a zero-trust model applies and that every TA is verified.

Annex B:
Threats arising from explicability and transparency

B.1 Overview

There is a legitimate concern that by making processes more open by adopting measures that make a system more explicable or more transparent that it also makes those systems more vulnerable to attack.

The principle from crypto-security described by Auguste Kerckhoffs "*A cryptosystem should be secure even if everything about the system, except the key, is public knowledge*" [i.4] can be extended to AI systems. In applying Kerckhoffs' principle to AI the aim is that the purpose of algorithms, data and the intelligence model, when they are public do not impact on system security, where system security includes the ability to demonstrate and prove the explicability and transparency of the system.

B.2 Model extraction

In ETSI GR SAI 004 [i.1] it is inferred that most AI systems are opaque, where the systems accept inputs, and generate outputs without ever revealing the internal logic, algorithms or parameters. In addition, training data sets, which effectively contain all the knowledge of the trained system, are also usually kept confidential. The role of transparency and explicability however challenges the inference of [i.1].

If opacity is removed in favour of transparency it can reasonably be asked: How transparent? The short answer is that it depends on context and some examples below can assist in determining to what extent an AI system can remain opaque, or its data sets remain confidential.

EXAMPLE: An AI system that is categorized as High Risk under the EU's AI Act [i.6] can be required to undergo compliance testing against certain mandatory requirements and an ex-ante conformity assessment. In such cases it would be reasonable to expect the AI system to be fully open, at least to the assessors.

NOTE: An open system does not infer an insecure or unsafe system. Rather by adopting Kerckhoffs' principle [i.4] the AI system is expected to be designed in such a way that it is secure and safe, and its secrets secret, whilst also being open.

If the AI system is transparent and explicable it should not infer that it can be easily extracted. The intention is therefore to encourage transparency and explicability whilst at the same time offering assurance to developers that the model itself will not be open to abuse (e.g. by theft). Methods to achieve this are still under study and development.

Annex C: Data quality in AI/ML

Many of the commonly perceived threats in AI/ML systems can be classified as arising from data quality issues. The aim of transparency and explicability as outlined in the present document is part of the quality metric of the system.

The provisions recommended and identified in ETSI GR SAI 002 [i.2] apply in support of element 2c of the static explicability analysis (see clause 5 of the present document).

Table C.1

Documentation Element	Element	Short description
2c	Method(s) used to determine data quality	Methods and processes used in determining if the input data is a fair and accurate representation of the desired input. This should address how bias or preference is identified and corrected in the data input

Common methods of data quality assessment include table C.2, where the AI/ML concern is noted.

Table C.2

Metric	Definition	Role in AI/ML
Accuracy	Measures the number (and type) of errors in a dataset. Typically measured as a percentage of errors across all the records.	
Completeness	Checks if all elements in a data record are complete.	
Consistency	Measured across datasets to determine if the same data is presented in the same way.	
Timeliness	Determines if the data is fresh (for the context it is consumed in).	
Uniqueness	Tracks duplicate data with a view to eliminating duplicates.	Whilst often a necessary constraint in relational databases there is often a different view in statistical analysis where a cleaned data source may actually give misleading results (there is some value in ensuring that complete records are not duplicated within single datasets but care has to be taken to validate duplication versus repetition).
Validity		

The ISO 8000 series of standards also address data quality as identified by their titles below with the most relevant elements for transparency and explicability highlighted in bold type (these are not cited as explicit references but are listed in the bibliography).

- ISO/TS 8000-1:2011: "Data quality - Part 1: Overview"
- ISO 8000-2:2017: "Data quality - Part 2: Vocabulary"
- ISO 8000-8:2015: "Data quality - Part 8: Information and data quality: Concepts and measuring"
- ISO 8000-61:2016: "Data quality - Part 61: Data quality management: Process reference model"
- ISO 8000-63:2019: "Data quality - Part 63: Data quality management: Process measurement"
- ISO 8000-100:2016: "Data quality - Part 100: Master data: Exchange of characteristic data: Overview"
- ISO 8000-102:2009: "Data quality - Part 102: Master data: Exchange of characteristic data: Vocabulary" (Withdrawn)

- ISO 8000-110:2009: "Data quality - Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification"
- ISO 8000-115:2017: "Data quality - Part 115: Master data: Exchange of quality identifiers: Syntactic, semantic and resolution requirements"
- **ISO 8000-120:2016: "Data quality - Part 120: Master data: Exchange of characteristic data: Provenance"**
- **ISO 8000-130:2016: "Data quality - Part 130: Master data: Exchange of characteristic data: Accuracy"**
- **ISO 8000-140:2016: "Data quality - Part 140: Master data: Exchange of characteristic data: Completeness"**
- ISO/TS 8000-150:2011: "Data quality - Part 150: Master data: Quality management framework"
- ISO/TS 8000-311:2012: "Data quality - Part 311: Guidance for the application of product data quality for shape (PDQ-S)"

It is suggested in ETSI GR SAI 002 [i.2] that poisoning as an attack can be determined by identifying data values significantly outside of the norm for base data. However it is also known that influencing opinion, e.g. on social media and in news articles, does not require significant modification of data, but that data is stressed differently. Thus the methods of data quality assessment in ETSI GR SAI 002 [i.2] may not always be practical if such filtering also misidentifies long term, or short term, actual variation in data.

Annex D: Bibliography

D.1 Data Quality

- OECD: "Quality Framework and Guidelines for OECD Statistical Activities", Version 2011/1.
- ISO/TS 8000-1:2011: "Data quality - Part 1: Overview".
- ISO 8000-2:2017: "Data quality - Part 2: Vocabulary".
- ISO 8000-8:2015: "Data quality - Part 8: Information and data quality: Concepts and measuring".
- ISO 8000-61:2016: "Data quality - Part 61: Data quality management: Process reference model".
- ISO 8000-63:2019: "Data quality - Part 63: Data quality management: Process measurement".
- ISO 8000-100:2016: "Data quality - Part 100: Master data: Exchange of characteristic data: Overview".
- ISO 8000-102:2009: "Data quality - Part 102: Master data: Exchange of characteristic data: Vocabulary" (Withdrawn).
- ISO 8000-110:2009: "Data quality - Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification".
- ISO 8000-115:2017: "Data quality - Part 115: Master data: Exchange of quality identifiers: Syntactic, semantic and resolution requirements".
- ISO 8000-120:2016: "Data quality - Part 120: Master data: Exchange of characteristic data: Provenance".
- ISO 8000-130:2016: "Data quality - Part 130: Master data: Exchange of characteristic data: Accuracy".
- ISO 8000-140:2016: "Data quality - Part 140: Master data: Exchange of characteristic data: Completeness".
- ISO/TS 8000-150:2011: "Data quality - Part 150: Master data: Quality management framework".
- ISO/TS 8000-311:2012: "Data quality - Part 311: Guidance for the application of product data quality for shape (PDQ-S)".
- ETSI GR SAI 005: "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".
- ISO/IEC TR 24028: "Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence".
- ISO/IEC 22989: "Artificial intelligence concepts and terminology".

History

Document history		
V1.1.1	March 2023	Publication