ETSI TS 104 224 V1.1.1 (2025-03)

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

Reference
RTS/SAI-0016

Keywords
artificial intelligence

ETSI

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

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

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

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

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

Foreword

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

NOTE: The present document updates and extends ETSI GR SAI 007 prepared by ISG SAI.

Modal verbs terminology

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

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

1 Scope

The present document identifies steps to be taken by designers and implementers of AI platforms in order to give assurance of the explicability and transparency of AI processing. AI processing includes AI decision making and AI data processing.

NOTE: The present document uses the term explicability but recognizes that many other publications use the term explainability as a synonym. The terms are interchangeable with the proviso that the latter term is not a commonly accepted UK English word but that it has been used in the specific context of AI (see also clause 3.1 of the present document).

2 References

2.1 Normative references

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

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

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

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

[1] ETSI TS 104 050: "Securing Artificial Intelligence (SAI); AI Threat Ontology and definitions".

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

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 TR 104 221: "Securing Artificial Intelligence (SAI); Problem Statement".

NOTE: An earlier version of the above document is available as ETSI GR SAI 004.

[i.2] ETSI TR 104 048: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

NOTE: An earlier version of the above document is available as ETSI GR SAI 002.

[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] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[i.6] DARPA: "XAI: Explainable Artificial Intelligence".

[i.7] 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", 29 January 2019, Atlanta, GA, USA. ACM, New York, NY, USA.

[i.8] 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.9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford: "Datasheets for Datasets", Communications of the ACM, Volume 64, Issue 12, pp. 89-92, November 2021.

[i.10] 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, doi: 10.1038/s41467-019-08987-4.

[i.11] Molnar C.: "Interpretable Machine Learning-A Guide for Making Black Box Models Explainable".

[i.12] 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.

[i.13] ETSI TR 104 102: "Cyber Security (CYBER); Encrypted Traffic Integration (ETI); ZT-Kipling methodology".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in ETSI TS 104 050 [1] and ISO/IEC 22989 [2] and the following apply:

AI system: engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives

NOTE: Definition from ISO/IEC 22989 [2].

explainability: property of an AI system to express important factors influencing the AI system results in a way that humans can understand

NOTE: Definition from ISO/IEC 22989 [2].

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	Organization for Economic Cooperation and Development
RTE	Run Time Explicability
TA	Trust Association
XAI	eXplainable AI

4 Explicability and transparency

The SAI problem statement, ETSI TR 104 221 [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 TR 104 221 [i.1] and in ISO/IEC 22989 [2] 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 third 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. Trust should not be attributed where purpose is not clear.

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 TS 104 050 [1], 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 TR 104 221 [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 shall be maintained (i.e. at all times static explicability shall be a characteristic of the current system).

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

Table 1: System documentation elements in static explicability analysis

Documentation Element	Element	Mandatory	Short description
1	Statement of system purpose	Yes	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)	Yes	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)	Yes	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	Strongly recommended	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	Yes	For each processing or data element a means to identify liability for correction of errors or for maintenance of the element.

5.2 Requirements for 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.

NOTE 1: In the above example multiple AI capabilities are identified even if the perception of the user is of a single AI being applied.

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).

The following characteristics shall be identifiable in the statement of system purpose:

- **Unambiguous:** it should be impossible to interpret the system purpose in more than one way.
- **Complete:** the system purpose should contain all the information necessary to understand it without requiring reference to other documents.

NOTE 2: The above requirement may be seen to contradict best practice in standards development where referencing is used to ensure succinctness, whereas in the statement of system purpose a little more verbosity may be beneficial.

- **Precise:** the system purpose should be worded clearly and exactly, without unnecessary detail that might confuse the reader.

- **Well-structured:** any individual elements of the system purpose should be included in an appropriate and easy-to-read manner.

The present document provides a template for the documenting of the system purpose in Annex D.

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 (this is also addressed in the ZT-Kipling method defined in ETSI TR 104 102 [i.13] and in Annex A):

- 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.7] performs much of the above role and where in [i.7] 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.7] are part of closing that gap in standardization. In addition, the use of datasheets as outlined in [i.9] 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.

NOTE: This is addressed in part in the AI Act [i.5] as part of the transparency requirements in Article 13.