ISO/IEC ~~TS~~DTS 6254~~:2024(E)~~

ISO/IEC JTC 1/SC- 42~~/WG 3~~

Secretariat: ANSI

Date: ~~2024-10~~2025-03-03

Information technology- — Artificial intelligence — Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems

DTS stage

~~**Warning for WDs and CDs**~~

~~This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.~~

~~Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.~~

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11
~~Email~~E-mail: copyright@iso.org
Website: ~~www.iso.org~~www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) is a and IEC (the International Electrotechnical Commission) form the specialized system for worldwide federation of national standardsstandardization. National bodies (that are members of ISO member bodies). The workor IEC participate in the development of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. Internationalby the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documentsdocument should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawnISO and IEC draw attention to the possibility that some of the elementsimplementation of this document may beinvolve the subjectuse of (a) patent rights. ISO(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.htmlwww.iso.org/members.html and www.iec.ch/national-committees.

Introduction

When AI systems are used to help make decisions that affect people's lives, it is important that people understand how those decisions are made. Achieving useful explanations of the behaviour of AI systems and their components is a complex task. Industry and academia are actively exploring emerging methods for enabling explainability, as well as scenarios and reasons why explainability can be required.

Due to the multitude of stakeholders and communities contributing to this effort, the field is suffering from a certain terminological inconsistency. Most notably, the methods to provide such explanations of the behaviour of an AI system are discussed under the banner of "explainability", "interpretability", (sometimes even other terms like "transparency"), raising the question of how these terms relate to each other. This document aims to provide practical guidance for stakeholders regarding compliance with regulatory requirements labelled one way or another. With this goal in mind, it uses the umbrella term "explainability" and provides a non-exhaustive taxonomy and list of approaches that stakeholders can use to comply with regulatory requirements.

While the overarching goal of explainability is to evaluate the trustworthiness of AI systems, at different stages of the AI system life cycle, diverse stakeholders can have more specific objectives in support of the goal. To illustrate this point, several examples are provided. For developers, the goal can be improving the safety, reliability and robustness of an AI system by making it easier to identify and fix bugs. For users, explainability can help to decide how much to rely on an AI system by uncovering potential sources or existence of unwanted bias or unfairness. For service providers, explainability can be essential for demonstrating compliance with legal requirements. For policy makers, understanding the capabilities and limitations of different explainability methods can help to develop effective policy frameworks that best address societal needs while promoting innovation. Explanations can also help to design interventions to improve business outcomes.

This document describes the applicability and the properties of existing approaches and methods for improving explainability of ML models and AI systems. This document guides stakeholders through the important considerations involved with selection and application of such approaches and methods.

While methods for explainability of ML models can play a central role in achieving the explainability of AI systems, other methods such as data analytics tools and fairness frameworks can contribute to the understanding of AI systems' behaviour and outputs. The description and classification of such complementary methods are out of scope for this document.

Information technology – — Artificial intelligence — Objectives and approaches for explainability ~~of~~ and interpretability of machine learning (ML) models and artificial intelligence (AI) systems

1 Scope

This document describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regard to machine learning (ML) models and artificial intelligence (AI) systems' behaviours, outputs and results. Stakeholders include but are not limited to, academia, industry, policy makers and end users. It provides guidance concerning the applicability of the described approaches and methods to the identified objectives throughout the AI system's life cycle, as defined in ISO/IEC 22989.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the following terms and definitions ~~given in the following~~ apply.

ISO and IEC maintain ~~terminological~~terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at ~~http://www.electropedia.org/~~https://www.electropedia.org/

Field Code Changed

3.1 ~~3.1~~
stakeholder
any individual, group, or organization that can affect, be affected by or perceive itself to be affected by a decision or activity

[SOURCE: ISO/IEC 22989:2022, 3.5.13]

3.2 ~~3.2~~
explainability
property of an *AI system* ~~(3.4)~~(3.4) that enables a given human audience to comprehend the reasons for the system's *behaviour* ~~(3.22)~~(3.22)

Note 1 to entry: Explainability methods are not limited to the production of explanations, but also include the enabling of interpretations.

3.3 ~~3.3~~
transparency
<system> property of a system that appropriate information about the system is communicated to relevant *stakeholders* ~~(3.1)~~(3.1)

Note 1 to entry: Appropriate information for system transparency can include aspects such as features, performance, limitations, components, procedures, measures, design goals, design choices and assumptions, data sources and labelling protocols.

Note 2 to entry: Inappropriate disclosure of some aspects of a system can violate security, privacy, or confidentiality requirements.

[SOURCE: ISO/IEC 22989:2022, 3.5.15]

3.4 ~~3.4~~
artificial intelligence system
AI system
engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives

Note 1 to entry: The engineered system can use various techniques and approaches related to artificial intelligence to develop a model to represent data, knowledge, processes, etc. which can be used to conduct tasks.

[SOURCE: ISO/IEC 22989:2022, 3.1.4]

3.5 ~~3.5~~
machine learning
ML
process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the data or experience

[SOURCE: ISO/IEC 22989:2022, 3.3.5]

3.6 ~~3.6~~
trustworthiness
ability to meet *stakeholder* ~~(3.1)~~(3.1) expectations in a verifiable way

Note 1 to entry: Depending on the context or sector and also on the specific product or service, data and technology used, different characteristics apply and need verification to ensure *stakeholders'* ~~(3.1)~~(3.1) expectations are met.

Note 2 to entry: Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality and usability.

Note 3 to entry: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

[SOURCE: ISO/IEC TR 24028:2020, 3.42, modified — Stakeholders' expectations replaced by stakeholder expectations; comma between quality and usability replaced by "and".]

3.7 ~~3.7~~
feature
measurable property of an object or event with respect to a set of characteristics

Note 1 to entry: Features play a role in training and prediction.

Note 2 to entry: Features provide a machine-readable way to describe the relevant objects. As the algorithm will not go back to the objects or events themselves, feature representations are designed to contain all useful information.

[SOURCE: ISO/IEC 23053:2022, 3.3.3]

3.8 ~~**3.8**~~
global
property of an *explanation* ~~(3.27)~~(3.27) or an *interpretation* ~~(3.28)~~(3.28) that describes how model predictions are determined

Note 1 to entry: A global explanation provides an overall understanding of the model's typical operation. For instance, a list of rules or *features* ~~(3.7)~~(3.7) that determine the model outputs is an example of global explanation.

3.9 ~~**3.9**~~
local
property of an *explanation* ~~(3.27)~~(3.27) or an *interpretation* ~~(3.28)~~(3.28) that describes how a single model prediction was determined

Note 1 to entry: Compared to a global explanation, a local explanation does not try to explain the whole model.

3.10 ~~**3.10**~~
post-hoc explanation
explanation ~~(3.27)~~(3.27) built by applying analysis on the model after it has been trained or developed

Note 1 to entry: Post-hoc explanations are often used with *opaque box* ~~(3.16)~~(3.16) models, but they are not limited to opaque box models.

3.11 ~~**3.11**~~
feature-based explanation
explanation ~~(3.27)~~(3.27) of model *behaviour* ~~(3.22)~~(3.22) based on input *features* ~~(3.7)~~(3.7)

Note 1 to entry: For instance, a measure of how much each input feature contributes to a model's output for a given data point is an example of feature-based explanation.

Note 2 to entry: An input feature for a model does not necessarily correspond to the inputs a user gives as entry as several layers of processing can be applied.

3.12 ~~**3.12**~~
application programming interface
API
boundary across which a software application uses facilities of programming languages to invoke software services

[SOURCE: ISO/IEC 13522-6:1998~~(en),~~ 3.3]

3.13 ~~3.13~~
backpropagation
neural network training method that uses the error at the output layer to adjust and optimise the weights for the connections from the successive previous layers

[SOURCE: ISO/IEC 23053:2022, 3.2.1]

3.14 ~~3.14~~
classification model
<machine learning> machine learning model whose expected output for a given input is one or more classes

[SOURCE: ISO/IEC 23053:2022, 3.1.1]

3.15 ~~3.15~~
closed box
black box
<access> property of an *AI system* ~~(3.4)~~(3.4) or a model within an AI system, whereby only its outputs can be obtained programmatically

3.16 ~~3.16~~
opaque box
black box
<explainability> property of an *AI system* ~~(3.4)~~(3.4) or a model within an AI system, whereby it does not offer *intrinsic interpretability* ~~(3.17)~~(3.17)

3.17 ~~3.17~~
intrinsic interpretability
inherent interpretability
property of an AI model that holds its criteria and *decision process* ~~(3.21)~~(3.21) in an intelligible way in its structure or content

Note 1 to entry: Intrinsic interpretability is not limited to access only, but also implies an ability to understand the provided information. For instance, a structure of millions of parameters does not usually constitute an intelligible way of holding it.

Note 2 to entry: Intrinsic interpretability is opposed to *opaque box* ~~(3.16).~~(3.16).

3.18 ~~3.18~~
decision
content or item produced by the *AI system* ~~(3.4)~~(3.4) as a fulfilment of its task, based on a given input

Note 1 to entry: The decision can be a class, but also any other form of structured or unstructured data (e.g. a sentence, an image).

3.19 ~~3.19~~
outcome
one of the various options that the *AI system* ~~(3.4)~~(3.4) considers when choosing a given *decision* ~~(3.18)~~(3.18)

Note 1 to entry: Outcomes are candidate decisions.

3.20 ~~3.20~~
output
any data or information returned by the *AI system* ~~(3.4)~~(3.4) when processing a given input

Note 1 to entry: Outputs encompass decisions but also any additional data or information that is returned together with a decision, e.g. to contextualize or explain it.

3.21 3.21
decision process
set of steps and criteria used by the *AI system* (3.4)(3.4) to analyse an input and choose the *decision* (3.18)(3.18) among the possible *outcomes* (3.20)(3.20)

Note 1 to entry: Depending on the design of the AI system, that decision process can be embedded in part or in whole, implicitly or explicitly, into the AI system's models.

3.22 3.22
behaviour
<AI system> any observable effect of a given *decision process* (3.21),(3.21), such as a particular *decision* (3.18),(3.18), the preferences made among different *outcomes* (3.19),(3.19), a relationship among multiple decisions or a statistical property of the complete set of decisions made by the *AI system* (3.4)(3.4) (including future decisions)

Note 1 to entry: Depending on the design of the AI system, the behaviour of the AI system can be attributed to the behaviour of the AI system's models or to their interplay.

3.23 3.23
factor
element, property or other characteristic that is considered during the *decision process* (3.21)(3.21) and can have an effect on the chosen *decision* (3.18)(3.18)

3.24 3.24
cause
any type of circumstance that can lead to a given *decision* (3.18),(3.18), including for instance the presence, absence or value of a *factor* (3.23),(3.23), but also the analysis made of that factor, its similarity or interaction with other factors, or the presence or absence of a given step or criterion in the *decision process* (3.21)(3.21)

3.25 3.25
rationale
piece of information or the analysis made of that information, based on which *decisions* (3.18)(3.18) are made

Note 1 to entry: A rationale provided for a single decision identifies one or more causes as having affected the *decision process* (3.21)(3.21) of the *AI system* (3.4)(3.4) when choosing that particular decision. A rationale provided without the context of a specific decision identifies a set of *causes* (3.24)(3.24) that can affect the *behaviour* (3.22)(3.22) of the AI system during past or future decisions.

3.26 3.26
justification
piece of information or the analysis made of that information, that is sufficient to choose a given *decision* (3.18)(3.18) among the possible *outcomes* (3.19)(3.19)

Note 1 to entry: A justification identifies causes relevant to a given decision, without assumption on the set of causes that have affected the decision process of the *AI system* (3.4).(3.4).

3.27 3.27
explanation
result of expressing a given *rationale* (3.25)(3.25) or *justification* (3.26)(3.26) in a way that humans can understand

Note 1 to entry: Explanations can pertain to a *decision* (3.18)(3.18) or to an *AI system* (3.4).(3.4).

3.28 ~~3.28~~
interpretation
result of understanding (by a human) a given *rationale* ~~(3.25)~~(3.25) or *justification* ~~(3.26)~~(3.26)

Note 1 to entry: Interpretations can pertain to a *decision* ~~(3.18)~~(3.18) or to an *AI system* ~~(3.4)~~(3.4).

Note 2 to entry: Interpretations can be produced either based on a received explanation, or directly from observation without an explicit act of expression.

3.29 ~~3.29~~
behavioural accuracy
adequacy between the *outcomes* ~~(3.19)~~(3.19) to which the *explanation* ~~(3.27)~~(3.27) leads and the actual *decisions* ~~(3.18)~~(3.18) made by the *AI system* ~~(3.4)~~(3.4)

3.30 ~~3.30~~
simulatability
ability of humans to process the provided information and apply the corresponding criteria mentally to obtain the output

4 Symbols and abbreviated terms

~~CEM contrastive explanations method~~

~~CEM-MAF contrastive explanations method with monotonic attribute functions~~

~~CNN convolutional neural networks~~

~~CV computer vision~~

~~ML machine learning~~

~~XAI explainable artificial intelligence~~

CEM	contrastive explanations method
CEM-MAF	contrastive explanations method with monotonic attribute functions
CNN	convolutional neural networks
CV	computer vision
ML	machine learning
XAI	explainable artificial intelligence

5 Overview

Explainability is the property of an AI system that enables a given human audience to comprehend the reasons for the system's behaviour. Reasons are rationales or justifications, as defined in this document with respect to the system's behaviour. The appropriate way of achieving explainability depends on the context and stakeholder characteristics. Stakeholder-appropriate explainability helps to achieve concrete objectives such as:

— identifying the causes of an incorrect decision;

— ensuring that a decision was taken for the right reasons;