Technical Specification

ISO/IEC TS 6254

First edition
2025-09

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

Technologies de l'information — Intelligence artificielle — Objectifs et approches pour l'explicabilité et l'interprétabilité des modèles d'apprentissage automatique (AA) et des systèmes d'intelligence artificielle (IA)

Reference number
ISO/IEC TS 6254:2025(en)

© ISO/IEC 2025



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2025

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

Published in Switzerland

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Symbols and abbreviated terms **5**
5 Overview **6**
6 Stakeholders' objectives **6**
6.1 General 6
6.2 AI user 7
6.3 AI developer 7
6.4 AI product or service provider 7
6.5 AI platform provider 8
6.6 AI system integrator 8
6.7 Data provider 8
6.8 AI evaluator 8
6.9 AI auditor 8
6.10 AI subject 8
6.11 Relevant authorities 8
6.11.1 Policy makers 8
6.11.2 Regulators 8
6.11.3 Other authorities 9
7 Explainability considerations throughout the AI system life cycle **9**
7.1 General 9
7.2 Inception 10
7.3 Design and development 10
7.3.1 General 10
7.3.2 Development of the explainability component 10
7.3.3 Explainability's contribution to development 11
7.4 Verification and validation 11
7.4.1 General 11
7.4.2 Evaluation of the explainability component 11
7.4.3 Explainability's contribution to evaluation 13
7.5 Deployment 14
7.5.1 General 14
7.5.2 Deployment of the explainability component 14
7.5.3 Explainability's contribution to deployment 14
7.6 Operation and monitoring 14
7.7 Continuous validation 14
7.8 Re-evaluation 14
7.9 Retirement 15
8 Property taxonomy of explainability methods and approaches **15**
8.1 General 15
8.2 Properties of explanation needs 16
8.2.1 General 16
8.2.2 Expertise profile of the targeted audience 16
8.2.3 Frame activity of interpretation or explanation 17
8.2.4 Scope of information 17
8.2.5 Completeness 17
8.2.6 Depth 18
8.2.7 Reasoning path 18
8.2.8 Implicit and explicit explanations 19

© ISO/IEC 2025 – All rights reserved

8.3 Forms of explanation ... 19
8.3.1 General ... 19
8.3.2 Numeric ... 19
8.3.3 Visual ... 19
8.3.4 Textual ... 20
8.3.5 Structured ... 20
8.3.6 Example-based ... 20
8.3.7 Interactive exploration tools ... 20
8.4 Technical approaches towards explainability ... 20
8.4.1 General ... 20
8.4.2 Empirical analysis ... 21
8.4.3 Post hoc interpretation ... 21
8.4.4 Inherently interpretable components ... 21
8.4.5 Architecture- and task-driven explainability ... 22
8.5 Technical constraints of the explainability method ... 22
8.5.1 General ... 22
8.5.2 Genericity of the method ... 22
8.5.3 Transparency requirements ... 23
8.5.4 Display requirements ... 23

9 Approaches and methods to explainability ... 23
9.1 General ... 23
9.2 Empirical analysis methods ... 24
9.2.1 General ... 24
9.2.2 Fine-grained evaluation ... 25
9.2.3 Error analysis ... 25
9.2.4 Analysis-oriented datasets ... 25
9.2.5 Ablation ... 26
9.2.6 Known trends ... 26
9.3 Post hoc methods ... 27
9.3.1 Local ... 27
9.3.2 Global ... 32
9.4 Inherently interpretable components ... 36
9.4.1 General ... 36
9.4.2 Legible models ... 37
9.4.3 Meaningful models ... 39
9.4.4 Models with explicit knowledge ... 41
9.5 Architecture- and task-driven methods ... 43
9.5.1 General ... 43
9.5.2 Informative features ... 43
9.5.3 Rich and auxiliary inputs ... 44
9.5.4 Multi-step processing ... 44
9.5.5 Rich outputs ... 45
9.5.6 Rationale-based processing ... 46
9.5.7 Rationale generation as auxiliary output ... 46
9.6 Data explanation ... 47

Annex A (informative) **Extent of explainability and interaction with related concepts ... 48**

Annex B (informative) **Illustration of methods' properties ... 51**

Annex C (informative) **Concerns and limitations ... 61**

Bibliography ... 65

© ISO/IEC 2025 – All rights reserved

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by 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.

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 document 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 or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(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.

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. 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.html and www.iec.ch/national-committees.

© ISO/IEC 2025 – All rights reserved

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 machine learning (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.

© ISO/IEC 2025 – All rights reserved

Information technology — Artificial intelligence — Objectives and approaches for explainability 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 apply.

ISO and IEC maintain 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 https://www.electropedia.org/

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

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

© ISO/IEC 2025 – All rights reserved