



BSI Standards Publication

Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks

National foreword

This Published Document is the UK implementation of CEN/CLC ISO/IEC/TS 12791:2024. It is identical to ISO/IEC TS 12791:2024.

The UK participation in its preparation was entrusted to Technical Committee ART/1, Artificial Intelligence.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

This publication is not to be regarded as a British Standard.

© The British Standards Institution 2024
Published by BSI Standards Limited 2024

ISBN 978 0 539 33922 2

ICS 35.020

Compliance with a Published Document cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 30 November 2024.

Amendments/corrigenda issued since publication

Date	Text affected
------	---------------

November 2024

ICS 35.020

English version

Information technology - Artificial intelligence - Treatment
of unwanted bias in classification and regression machine
learning tasks (ISO/IEC TS 12791:2024)

Technologies de l'information - Intelligence artificielle
- Traitement des biais indésirables dans les tâches
d'apprentissage automatique de classification et de
régression (ISO/IEC TS 12791:2024)

Informationstechnik - Künstliche Intelligenz -
Behandlung von unerwünschtem Bias bei
Klassifizierungs- und Regressionsaufgaben des
maschinellen Lernens (ISO/IEC TS 12791:2024)

This Technical Specification (CEN/TS) was approved by CEN on 12 October 2024 for provisional application.

The period of validity of this CEN/TS is limited initially to three years. After two years the members of CEN and CENELEC will be requested to submit their comments, particularly on the question whether the CEN/TS can be converted into a European Standard.

CEN and CENELEC members are required to announce the existence of this CEN/TS in the same way as for an EN and to make the CEN/TS available promptly at national level in an appropriate form. It is permissible to keep conflicting national standards in force (in parallel to the CEN/TS) until the final decision about the possible conversion of the CEN/TS into an EN is reached.

CEN and CENELEC members are the national standards bodies and national electrotechnical committees of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Türkiye and United Kingdom.



CEN-CENELEC Management Centre:
Rue de la Science 23, B-1040 Brussels

European foreword

This document (CEN/CLC ISO/IEC/TS 12791:2024) has been prepared by Technical Committee ISO/IEC JTC 1 "Information technology" in collaboration with Technical Committee CEN-CENELEC/ JTC 21 "Artificial Intelligence" the secretariat of which is held by DS.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. CEN-CENELEC shall not be held responsible for identifying any or all such patent rights.

Any feedback and questions on this document should be directed to the users' national standards body/national committee. A complete listing of these bodies can be found on the CEN and CENELEC websites.

According to the CEN-CENELEC Internal Regulations, the national standards organizations of the following countries are bound to announce this Technical Specification: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Türkiye and the United Kingdom.

Endorsement notice

The text of ISO/IEC TS 12791:2024 has been approved by CEN-CENELEC as CEN/CLC ISO/IEC/TS 12791:2024 without any modification.

Contents

Page

Foreword.....iv

Introduction.....v

1 Scope.....1

2 Normative references.....1

3 Terms and definitions.....1

3.1 General.....1

3.2 Artificial intelligence.....3

3.3 Bias.....4

3.4 Testing.....5

4 Abbreviated terms.....6

5 Treating unwanted bias in the AI system life cycle.....6

5.1 Inception.....6

5.1.1 Stakeholder identification.....6

5.1.2 Stakeholder needs and requirements definition.....7

5.1.3 Procurement.....8

5.1.4 Data sources.....9

5.1.5 Integration with risk management.....11

5.1.6 Acceptance criteria.....11

5.2 Design and development.....12

5.2.1 Feature representation.....12

5.2.2 Metadata sufficiency.....12

5.2.3 Data annotations.....12

5.2.4 Adjusting data.....13

5.2.5 Methods for managing identified risks.....13

5.3 Verification and validation.....13

5.3.1 General.....13

5.3.2 Static testing of data used in development.....14

5.3.3 Dynamic testing.....14

5.4 Re-evaluation, continuous validation, operations and monitoring.....15

5.4.1 General.....15

5.4.2 External change.....16

5.5 Disposal.....17

6 Techniques to address unwanted bias.....17

6.1 General.....17

6.2 Algorithmic and training techniques.....17

6.2.1 General.....17

6.2.2 Pre-trained models.....18

6.3 Data techniques.....19

7 Handling bias in a distributed AI system life cycle.....19

Annex A (informative) Life cycle processes map.....21

Annex B (informative) Potential impacts of unwanted bias on different types of specific user.....22

Bibliography.....23

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 documents 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*, in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/CLC/JTC 21, *Artificial Intelligence*, in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement).

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.

Introduction

This document describes steps that can be taken to treat unwanted bias during the development or use of AI systems.

This document is based on ISO/IEC TR 24027 and provides treatment techniques in accordance with the AI system life cycle as defined in ISO/IEC 22989:2022, Clause 6 and ISO/IEC 5338. The treatment techniques in this document are agnostic of context. This document is based on the types of bias described in ISO/IEC TR 24027.

This document describes good practises for treating unwanted bias and can help an organization with the treatment of unwanted bias in machine learning (ML) systems that conduct classification and regression tasks. The techniques in this document are applicable to classification and regression ML tasks. This document does not address applicability of the described methods outside of the defined ML tasks.

This document does not contain organizational management and enabling processes related to an AI management system, which can be found in ISO/IEC 42001.

[Annex A](#) provides a cross-reference between the life cycle stages and the clauses of this document.

Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks

1 Scope

This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. This document provides mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias. This document is applicable to all types and sizes of organization.

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 5259-4:2024, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*

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

ISO/IEC/IEEE 29119-3:2021, *Software and systems engineering — Software testing — Part 3: Test documentation*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022 and the following 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 General

3.1.1

authoritative record

record which possess the characteristics of authenticity, reliability, integrity and useability

[SOURCE: ISO 30300:2020, 3.2.3]

3.1.2

consumer vulnerability

state in which an individual can be placed at risk of harm during their interaction with or a decision by a service provider due to the presence of personal, situational and market environment factors

[SOURCE: ISO 22458:2022, 3.5, modified — added reference to a decision by a service provider.]

3.1.3

current operating conditions

conditions under which an AI system is currently operating

Note 1 to entry: Conditions can include resource usage, environmental factors, geographic location of use, time of use, training provided to operators and the target population.

3.1.4

data subject

person to whom data refer

[SOURCE: ISO 25237:2017, 3.18]

3.1.5

data quality model

defined set of characteristics which provides a framework for specifying data quality requirements and evaluating data quality

[SOURCE: ISO/IEC 25012:2008, 4.6]

3.1.6

disposition

range of records processes associated with implementing records retention, destruction or transfer decisions which are documented in *disposition authorities* ([3.1.7](#)) or other instruments

[SOURCE: ISO 30300:2020, 3.4.8]

3.1.7

disposition authority

instrument that defines the *disposition* ([3.1.6](#)) actions that are authorized or required for specified records

[SOURCE: ISO 30300:2020, 3.5.4]

3.1.8

intended operating conditions

conditions under which an AI system is meant to function

Note 1 to entry: Conditions can include resource usage, environmental factors, geographic location of use, time of use, training provided to operators and the target population.

3.1.9

management system

set of interrelated or interacting elements of an *organization* ([3.1.10](#)) to establish policies and objectives, as well as processes to achieve those objectives

Note 1 to entry: A management system can address a single discipline or several disciplines.

Note 2 to entry: The management system elements include the organization's structure, roles and responsibilities, planning and operation.

[SOURCE: ISO/IEC 42001:2023, 3.4]

3.1.10

organization

person or group of people that has its own functions with responsibilities, authorities and relationships to achieve its objectives

Note 1 to entry: The concept of organization includes, but is not limited to, sole-trader (sole proprietor), company, corporation, firm, enterprise, authority, partnership, charity or institution or part or combination thereof, whether incorporated or not, public or private.

Note 2 to entry: If the organization is part of a larger entity, the term "organization" refers only to the part of the larger entity that is within the scope of the AI *management system* ([3.1.9](#)).

[SOURCE: ISO/IEC 42001:2023, 3.1]

3.1.11

records process

set of activities for managing authoritative records

[SOURCE: ISO 30300:2020, 3.4.13]

3.1.12

user

individual or group that interacts with a system or benefits from a system during its utilization

[SOURCE: ISO/IEC/IEEE 15288:2023, 3.53, modified — Note 1 to entry has been removed.]

3.2 Artificial intelligence

3.2.1

data quality

characteristic of data that the data meet the *organization's* ([3.1.10](#)) data requirements for a specified context

[SOURCE: ISO/IEC 5259-1:2024, 3.4]

3.2.2

data quality characteristic

category of data quality attributes that bears on *data quality* ([3.2.1](#))

[SOURCE: ISO/IEC 5259-1:2024, 3.5]

3.2.3

data quality measure

variable to which a value is assigned as the result of measurement of a *data quality characteristic* ([3.2.2](#))

[SOURCE: ISO/IEC 5259-1:2024, 3.7]

3.2.4

data provenance

provenance

information on the place and time of origin, derivation or generation of a data set, proof of authenticity of the data set, or a record of past and present ownership of the data set

[SOURCE: ISO/IEC 5259-1:2024, 3.16]

3.2.5

extreme data

type of sample that is an outlier with respect to the real-world distribution

3.2.6

feature

<machine learning> 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 information the algorithm is expected to need.

[SOURCE: ISO/IEC 23053:2022, 3.3.3, modified — Clarification of Note 2 to entry has been added.]