Technical Specification

# ISO/IEC TS 12791

First edition
2024-10

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

*Technologies de l'information — Intelligence artificielle — Traitement des biais indésirables dans les tâches d'apprentissage automatique de classification et de régression*



Reference number
ISO/IEC TS 12791:2024(en)



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

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.]

**3.2.7**
**functional correctness**
degree to which a product or system provides the correct results with the needed degree of precision

Note 1 to entry: AI systems, and particularly those using machine learning methods, do not usually provide functional correctness in all observed circumstances.

[SOURCE: ISO/IEC 25059:2023, 3.2.3]

**3.2.8**
**intended use**
use in accordance with information provided with an AI system, or, in the absence of such information, by generally understood patterns of usage

[SOURCE: ISO/IEC Guide 51:2014, 3.6, modified — "a product or system" has been changed to "an AI system".]

**3.2.9**
**inter-annotator agreement**
degree of consensus or similarity among the annotations made by different annotators on the same data

## 3.3 Bias

**3.3.1**
**AI subject**
organization, person or entity that is affected by an AI system, service or product

**3.3.2**
**automation bias**
propensity for humans to favour suggestions from automated decision-making systems and to ignore contradictory information from non-automated sources, even if it is correct

[SOURCE: ISO/IEC TR 24027:2021, 3.2.1, modified — "made without automation" was changed to "from non-automated sources".]

**3.3.3**
**coverage bias**
type of *data bias* (3.3.4) that occurs when a population represented in a dataset does not match the population that the machine learning model is making predictions about

**3.3.4**
**data bias**
data properties that if unaddressed lead to AI systems that perform better or worse for different objects, people or *groups* (3.3.5)

[SOURCE: ISO/IEC TR 24027:2021, 3.2.7]

**3.3.5**
**group**
subset of objects in a domain that are linked because they have shared characteristics

[SOURCE: ISO/IEC TR 24027:2021, 3.2.8]

**3.3.6**
**human cognitive bias**
bias that occurs when humans are processing and interpreting information

Note 1 to entry: human cognitive bias influences judgement and decision-making.

[SOURCE: ISO/IEC TR 24027:2021, 3.2.4]

**3.3.7**
**representativeness**
qualitative assessment of degree to which a given dataset's properties approximate the statistical properties of the *target population* (3.3.10) of interest

Note 1 to entry: Representativeness can be quantified through the use of one or more measures pertaining to the size, distribution or composition of the data.

Note 2 to entry: Representative test data enables verification that an AI system achieves an acceptable level of *functional correctness* (3.2.7) for the *target population* (3.3.10).

Note 3 to entry: Representative training data can enable training a machine learning model that achieves an acceptable level of *functional correctness* (3.2.7) for the *target population* (3.3.10).

**3.3.8**
**selection bias**
type of *data bias* (3.3.4) that can occur when a dataset's samples are not collected in a way that is representative of their real-world distribution

**3.3.9**
**statistical bias**
type of consistent numerical offset in an estimate relative to the true underlying value

Note 1 to entry: The offset is inherent to most estimates

[SOURCE: ISO 20501:2019, 3.3.9, modified — "inherent to most estimates" was moved to Note 1 to entry.]

**3.3.10**
**target population**
*group* (3.3.5) of *AI subjects* (3.3.1) that the AI system will process data in relation to

Note 1 to entry: The target population can include organizations or other objects.

**3.3.11**
**at-risk group**
subset of stakeholders that can be adversely affected by unwanted bias

Note 1 to entry: at-risk groups can also emerge from intersections of groups as described in ISO/IEC TR 24027.

Note 2 to entry: unforeseen at-risk groups can emerge due to the use of AI systems, as described in 5.1.5.

## 3.4 Testing

**3.4.1**
**dynamic testing**
*testing* (3.4.8) in which a *test item* (3.4.5) is evaluated by executing it

[SOURCE: ISO/IEC/IEEE 29119-2:2021, 3.3]

**3.4.2**
**model testing**
*testing* (3.4.8) in which the behaviour of a model is examined against a set of qualities or other criteria

Note 1 to entry: Model testing is usually performed by executing the model on a systematic set of inputs and evaluating how well its outputs achieve some measure of task performance, such as matching canonical answers or being rated highly by humans.

**3.4.3**
**static testing**
*testing* (3.4.8) in which a *test item* (3.4.5) is examined against a set of quality or other criteria without the test item being executed

[SOURCE: ISO/IEC/IEEE 29119-1:2022, 3.20, modified — The example has been removed.]