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.]

3.4.4
test completion report
test summary report
report that provides a summary of the *testing* (3.4.8) that was performed

[SOURCE: ISO/IEC/IEEE 29119-3:2021, 3.9]

3.4.5
test item
test object
work product to be tested

[SOURCE: ISO/IEC/IEEE 29119-1:2022, 3.107, modified — The example has been removed.]

3.4.6
test objective
reason for performing *testing* (3.4.8)

[SOURCE: ISO/IEC/IEEE 29119-2:2021, 3.49, modified — The example has been removed.]

3.4.7
test plan
detailed description of *test objectives* (3.4.6) to be achieved and the means and schedule for achieving them, organized to coordinate *testing* (3.4.8) activities for some *test item* (3.4.5) or set of test items

[SOURCE: ISO/IEC/IEEE 29119-2:2021, 3.50, modified — the Notes to entry have been removed.]

3.4.8
testing
set of activities conducted to facilitate discovery and evaluation of properties of a *test item* (3.4.5)

Note 1 to entry: Testing activities include planning, preparation, execution, reporting and management activities insofar as they are directed towards testing.

[SOURCE: ISO/IEC/IEEE 29119-1:2022, 3.131]

4 Abbreviated terms

AI artificial intelligence

ETL extract, transform and load

IID independent and identically distributed

ML machine learning

PII personally identifiable information

5 Treating unwanted bias in the AI system life cycle

5.1 Inception

5.1.1 Stakeholder identification

Stakeholder identification shall be conducted throughout the development of an AI system.

NOTE This identification enables the collection and assessment of information from those stakeholders on common effects of unwanted bias that they usually experience from AI systems.

Any individuals, groups or organizations who can be foreseeably positively or negatively affected by unwanted bias in the AI system should be considered, not just AI actors, but also including AI subjects directly using or gaining benefit from the implementation of an AI system.

This can include:

— users who operate or interface with an AI system, as they can offer personal feedback on unwanted bias that has affected them as individuals;

EXAMPLE 1 Front-line workers whose familiarity with digital technology is at a lower level than that of an AI application's main beneficiaries or target audience.

— decision-makers within an organization designing, developing, deploying, or using an AI system;

— AI partners (including AI auditors) who are required to perform conformity assessment on an AI system;

EXAMPLE 2 Auditing teams consider the diversity of the audited organization and the audited organization's target audience, customers and other interested parties when conducting the audit to prevent unwanted biases affecting the audit.

— regulators including bodies that are required to review conformity assessment results in relation to an AI system.

EXAMPLE 3 Regulatory staff consider the diversity of the auditors, the audited organization and the audited organization's target audience, customers and other interested parties when conducting the conformity assessment to prevent unwanted biases affecting the review.

Organizations responsible for the deployment or operation of an AI system shall consider:

— AI subjects about whom automated decisions are made, or who share an operating environment with an AI system;

EXAMPLE 4 A recommender system receiving input from a user ensures that variations in ability are accounted for in applications that are used by native and non-native language speakers.

— recipients of information derived from an AI system who are not direct users (e.g. public authorities who will make decisions based on AI-derived information curated by staff);

EXAMPLE 5 Analytical systems that analyse socioeconomic datasets to inform policy development can propagate unwanted bias based on historic data that does not reflect current conditions.

— data subjects who do not directly interact with an AI system but whose data are used in training.

EXAMPLE 6 An online chess playing game uses data from real-world matches and tournaments. The players of those matches are interested parties in this context.

5.1.2 Stakeholder needs and requirements definition

An organization shall identify and document requirements to reduce unwanted bias within an AI system as well as in recommendations, decisions or other output generated by an AI system.

Considered sources of bias-related requirements can include:

— applicable legal requirements;

— customer expectations;

— internal goals, strategies and policies (e.g. an ethics policy);

— organizational processes and decisions as part of AI governance;

— surveys of past known failure modes, based on resources such as the organization's documentation on its prior system failures or in AI incident databases;

— assumptions, processes, decisions and related activities made by individuals or groups across the AI life cycle.

NOTE 1 Treatment leading to unwanted bias can include any kind of action or inaction, including perception, etc.

NOTE 2 Annex B lists particular examples of how unwanted bias can affect specific types of AI users. ISO/IEC 22989:2022, 5.19 describes several AI stakeholder roles and sub-roles (e.g. AI provider, AI user, AI partner, AI subject). AI stakeholders have different roles and responsibilities in treating unwanted bias throughout the life cycle.

An organization shall define and document intended operating conditions under which an AI system is to be evaluated for bias. These can include the relevant groups of stakeholders or users and the geographical, linguistic, socioeconomic or cultural context of deployment.

EXAMPLE For intended operating conditions of a speech recognition system it can be defined and documented that both native and non-native speakers are expected.

The level of definition and documentation should be commensurate with the role of the organization within the AI life cycle. For example, an organization producing a pre-trained ML model should anticipate a broad number of intended operating conditions, while an organization conducting a live deployment should more precisely specify the intended operating conditions.

Procedures, expectations and accountability mechanisms should be in place for relevant actors within the organization to make use of the documentation they receive from others inside or outside the organization.

5.1.3 Procurement

AI producers or AI partners shall make information available regarding systems aspects that can affect unwanted bias, subject to legal requirements. The exact extent of a non-disclosure by AI producers and partners and the existence of legal requirements affecting disclosure shall be disclosed and justified. Examples of such system aspects that can affect unwanted bias include:

— algorithm selection;

— hyperparameter tuning;

— model bias;

— bias in data sources.

Details are provided in ISO/IEC TR 24027. The conveyed information can include representative testing results of the system with regards to unwanted bias.

Organizations should ensure that agreements with third parties include appropriate measures to treat the risk of unwanted bias considering the role of the organization and third parties, in particular, where an organization is unable to obtain full transparency on technical aspects of the system. Such measures shall be documented and justified including divergence from the provisions of this document.

Information made available by data providers in the course of a procurement shall include, subject to legal requirements:

— data provenance (including for training, validation and testing data);

EXAMPLE 1 This information can enable investigation of biases resulting from properties of the data source.

— data quality management policy and data quality check assessment results (including inter-annotator agreement measurements);

EXAMPLE 2 This information can enable acceptance of the presence of an observed but mitigated bias that is deemed appropriate for a given use case.

— data quality model and processing aspects (e.g. labelling processes used, types of machine learning models or algorithms used).

EXAMPLE 3 This information can be used to uncover biases associated with mechanisms that only the data provider has visibility on, as they are part of their internal process.

The exact extent of a non-disclosure and the existence of legal requirements affecting disclosure by data providers shall be disclosed and justified.

Information made available by data providers in the course of a procurement should include:

— the method of data collection;

— information on the working conditions for data labelling workers that can affect their human cognitive biases and hence cause potential unwanted bias in the resulting dataset;

— the geographic locations in which the data labelling was undertaken;

— salient aggregated demographics of the data labelling workforce;

— data dictionaries and associated metadata to enable unwanted bias risk management.

Data providers should ensure that privacy of individuals and groups in the workforce is maintained.

Information made available by AI technology providers in the course of a procurement should include:

— intended context(s) of use and related assumptions;

— known system limitations;

— recommended patterns of interaction between humans and the AI system during use;

— relevant trade-offs in algorithms, machine learning algorithms and ML models development that can affect or relate to unwanted bias;

EXAMPLE 4 How the model is used at inference time, where taking the argmax (argument of the maximum) for classification problems, or using deterministic ranking, risks amplifying small biases in model scores;

— data collection, modification and curation processes that can relate to bias such as imputation or augmentation;

— testing strategies used during the design and development or the verification and validation stage of the AI system (including acceptance criteria and the use of proxies in ML modelling).

Information can also be made available by AI technology providers regarding the geographical or cultural context of the design and development phase, as it can affect unwanted bias when this context differs significantly from the context of deployment.

Implementation guidance: ML tools can be procured, developed in-house or a combination of the two. Visibility into the kind of techniques used during algorithm, machine learning algorithm or ML model development is important for the effective treatment of unwanted bias.

Information made available by data providers can enable the investigation of biases resulting from the properties of the data source. This can enable acceptance of the presence of an observed but mitigated bias that is deemed appropriate for a given use case or uncover biases associated with mechanisms used by the data provider.

5.1.4 Data sources

Organizations shall document and evaluate unwanted bias in relation to:

— sources of data used by the AI system;

— data selection criteria and processes;

— data collection procedures, including the:

  — mechanisms for requesting informed consent and for revoking consent for future users;

  — provenance of the data;

  — collection, input, preprocessing, labelling and label cleaning mechanism;

  — impacts of dataset collection on data subjects.

Documentation of data sources enables organizations to qualitatively identify potential biases and to prioritize quantitative assessment in relation to those aspects. Quantitative assessment of unwanted bias is not usually tractable and benefits from source information.

Test data used to assess performance shall be representative. Training data shall be representative when using it to train an ML algorithm that is not specifically designed for leveraging non-representative data. Using multiple data sources is one way to improve representativeness across diverse groups; when done so, organizations shall determine whether the combination of datasets would introduce additional risks in relation to data bias and, if so, the newly introduced biases shall be assessed and treated in accordance with 5.1.5.

Representativeness can be a function of data size, type, dimensionality and complexity. Imputations, exclusions and augmentations in the dataset can affect representativeness. For ML models that cannot be trained on data with missing values, the distribution of imputations, exclusions and augmentations applied to make the dataset usable can affect representativeness.

For supervised ML, the distribution of the label values in training data and across at-risk groups is relevant to representativeness. For example, in the case of binary supervised classifiers, the balance of the data among the relevant groups can be evaluated by considering the ratio of positive and negative training examples within each group.

Where appropriate for the use case, the following aspects of each data source shall be evaluated and documented:

— intended use and purpose of the dataset created, including specific tasks;

— identification of dataset creators and sources of funding if applicable;

— composition of the dataset, including nature, size, labels, relations, errors, redundancies, noise and missing information of the instances in the dataset;

— completeness regarding the contents of the data such as confidential information, sensitive data that reveals identifiers of individuals, subpopulations and groups, and information on missing features for each group of relevant stakeholders;

— terms of use and license;

— accuracy, including the amount of inaccurate data contained within the dataset and the inaccuracy for each group of stakeholders;

— currency, including potential effects of the time of collection on accuracy;

— appropriateness in terms of amount of toxic or offensive data contained in the dataset;

— consistency, including labels (e.g. as measured by inter-annotator agreement);

— dataset coverage across different sub-groups relevant to the deployment context, including intersectional sub-groups;

— consideration of how labels that are proxies for unobservable constructs can lead to unintended impacts;

— risk of feedback loops and error propagation between data collection and modelling;

— dataset maintenance, including the stakeholders responsible for supporting, hosting, updating, versioning, retaining, expiring the dataset as well as participatory approaches for dataset improvement;

— traceability, including other systems involved in funnelling or treating the data within each source;

— understandability, including symbols, units and languages;

— auditability, including past, planned and potential audits;

— identifiability and protection of Personally Identifiable Information (PII);

— relevance of a data source for a given use case;

— representativeness to the target population.

Further information on data quality measures can be found in ISO/IEC 5259-2.

5.1.5 Integration with risk management

An organization shall identify and document risks related to unwanted bias that can occur during the design, development, deployment and use of an AI system. An organization shall assess and treat these identified risks to affected AI stakeholders.

NOTE Risks of unwanted bias include the risks for unwanted biases caused by inaction or perception.

Consideration shall be given to at-risk groups present in the data (training, validation, test or production data) even if the members of those groups are not explicitly identified in the data as belonging to the group. This consideration can uncover proxy bias.

Other aspects of incorporation with broader risk management can include:

— documentation of change management plans;

— incorporation of documentation into organizational inventories;

— communication with senior management relating to bias risks.

Risks relating to unwanted bias can arise from incorrect use or labelling of data, an AI system's mission and goals, a system's context of use including interactions with humans and failures to fully meet internal and external requirements. These risks can materially affect one or more groups of stakeholders.

A list of examples of types of bias that can be present in data and types of human cognitive bias can be found in ISO/IEC TR 24027:2021, 6.3.

An applicable risk management system such as ISO/IEC 23894 should be used in conjunction with this document. The following documents should also be used:

— ISO/IEC 25059 to identify quality measures that can vary by at-risk group;

— ISO/IEC 5259-2 to identify data quality measures that can vary by at-risk group.

5.1.6 Acceptance criteria

An organization shall determine appropriate tolerances for functional correctness. These shall be defined in advance of evaluating an AI system. The choice of acceptance criteria shall be justified in the documentation. Where functional correctness differs amongst groups or at-risk groups, an organization shall make this information available to relevant stakeholders and provide an explanation for the difference. Acceptance criteria shall be documented in the context of the intended use and operating conditions. Acceptance criteria for the system shall be testable. Where acceptance criteria are specified in relation to outputs, they shall be specified in a quantitative manner.

For example, AI stakeholders can specify a maximum limit for false positive or false negative rates. These limits can form a lower bound for acceptance criteria.

Organizations shall declare the diverse AI stakeholders involved in the decision-making and these stakeholders shall indicate whether they are satisfied with how the AI system operates.

5.2 Design and development

5.2.1 Feature representation

Organizations shall document the rationale for the design choices made regarding the input features used by the ML model. Where risks have been identified in relation to unwanted bias, the organization shall consider:

— types of data biases that can be present and the effects of feature selection;

— types of human cognitive biases that can be present in the individuals involved in selecting features;

— missing or unexpected feature values and unwanted imbalance;

— types of processes and decisions made during the life cycle that can be impacted by types of human cognitive bias;

— interactions between system components;

— biases that can result from the disproportionate (with respect to the real-world distribution) availability of datasets or features;

— biases that can be embedded in processes involved in selecting features;

— presence of proxies for demographic categories;

— distribution of group membership in training, validation, or test data and its representativeness of the population to which the system will operate;

— distribution of positive or negative outcomes across demographic groups;

— other forms of statistical or computational bias that can be relevant.

5.2.2 Metadata sufficiency

Metadata should be sufficient to identify potential sources of unwanted bias. Metadata or documentation of a dataset should enable data to be evaluated for unwanted bias.

5.2.3 Data annotations

Where risks have been identified in relation to unwanted bias, the organization shall consider:

— biases that result from the disproportionate (with respect to the real-world distribution) availability of annotations of data;

— types of human cognitive biases, including in relation to the annotation of data;

— types of processes and decisions made during the life cycle that can be impacted by types of human cognitive bias. Including the annotation of data.

An organization should implement mechanisms to ensure annotation activities do not create unwanted data bias. This can include:

— creating inputs with known ground truth to provide a quality measure;

— conducting sample checks;

— providing clear instruction or training to humans;

— mechanisms for humans to provide feedback on annotation tasks;

— evaluating statistical bias in automatically produced labels or annotations;

— comparing labels created by multiple humans on the same data by computing inter-annotator agreement;

— methods to resolve differences between human created labels.

Due to human cognitive biases and human mistakes, it should not be assumed that manual annotations are of high quality.

5.2.4 Adjusting data

When data bias is identified in the training data, an organization shall consider adjusting the data to treat the risk of unwanted bias. For example, before discarding features, an organization can analyse features for training data correlated with membership of a group at risk of unwanted bias. If the data bias is considered unwanted, a possible treatment is to create a more even distribution of examples in each category being represented.

5.2.5 Methods for managing identified risks

Organizations should implement and document mechanisms to treat identified risks related to unwanted bias from occurring.

An organization should consider:

— further data-based methods;

— model-based methods;

— post-hoc methods.

Examples of such methods can be found in ISO/IEC TR 24027:2021, 8.3.3.2 and Clause 6 of this document.

5.3 Verification and validation

5.3.1 General

Testing as defined in ISO/IEC/IEEE 29119-1 as a verification and validation process is appropriate for the detection of unwanted bias. There are three types of testing that are relevant to detecting bias:

— static testing of the training data to identify risks related to unwanted bias;

— dynamic testing of the ML model, including data pre-processing, to evaluate functional correctness;

— dynamic testing of the AI system to evaluate functional correctness for at-risk groups.

Testing shall be performed through static testing of data and dynamic testing of the model and AI system. To the extent possible, AI systems should also be tested with realistic users in realistic conditions. Approaches include bug bounties, functional testing, independent audits and other structured human feedback exercises:

— Bug bounties: Users, researchers and others are incentivized to find and report issues related to bias through structured programs and monetary compensation. Bug bounties are particularly important for assessing system behaviour in realistic or adversarial conditions.

— Functional testing: Testers assess the system as a whole as it is intended to be used in production environments. Functional testing can reveal issues with documentation, processes, interfaces and other system components that can affect AI system behaviour or its impacts on humans.

— Independent audits: Formal transparency exercises in which independent experts document adherence to an authoritative standard such as non-discrimination laws or this document. Independent audits provide external, informed perspectives on system behaviour.

— Structured human feedback exercises, including:

  — randomized control testing;

— evaluation of human–AI configurations through structured experiments;

— participatory engagement exercises (e.g. studies, surveys, or focus groups) with potentially affected individuals and communities;

— product management and user-interaction/user-experience research activities to prioritize and incorporate user and customer feedback.

The test approach used shall be documented in a test plan in accordance with ISO/IEC/IEEE 29119-3:2021, 7.2. The test results should be recorded in a test completion report in accordance with ISO/IEC/IEEE 29119-3:2021, 7.4. The test completion report shall include criteria for when verification and validation activities are repeated, including continuous monitoring.

5.3.2 Static testing of data used in development

The organization shall measure data quality in relation to unwanted bias for each at-risk group. The static testing shall select and prioritise appropriate measures. The static testing shall be documented, including justification for the selection and prioritisation of measures, as well as the results.

The measures selected by the organization shall enable the assessment of:

— whether the data contains an inappropriate imbalance among feature values, among labels or among other categories;

— whether the data is representative and relevant with respect to the expected production data;

— whether the data is sufficiently diverse both within and among groups;

— whether the proportion of instances with missing or corrupted content is evenly distributed for each at-risk group;

— whether the data format and the amount of information contained is consistent for each group.

Further information and examples of data quality measures can be found in ISO/IEC 5259-2. The data quality evaluation process shall be conducted in accordance with ISO/IEC 5259-4:2024, 6.3, including measuring training data quality for each at-risk group that has been identified in relation to unwanted bias.

If the training data does not contain values that link a given record to an at-risk group, and at-risk groups are identifiable from other data, then values from the other available data shall be provided as meta-data in order to perform the data quality checking process if the group is identifiable from the available data.

EXAMPLE A credit-risk rating system can be trained on data that explicitly excludes the gender variable and other potentially correlated variables (such as first name). But gender information can be collected and included as metadata, in order to enable testing for unwanted bias based on gender.

Implementation guidance: organizations can identify the profile of the available training data and validate whether the distribution of a specific variable is accurate. An example of this is identifying that records of a certain age group have been used for training, when a different spread of ages is expected in production data.

This activity can aim to validate the potential for selection bias and coverage bias but cannot do so exhaustively, as it is limited by the knowledge of the evaluator.

Organizations can identify stages in the data preparation process that can potentially introduce bias through “missing data”. For example, if a specific data item is not available consistently across an input dataset, organizations can impute that information for the remaining records or they can remove the data item.

5.3.3 Dynamic testing

In dynamic software testing, tests are designed based on the specification of desired behaviour, the structure of the system or the experience of the person conducting the activity.

In dynamic testing in the context of AI systems, given inputs generate expected outputs and measurement approaches that show whether the system is behaving statistically as anticipated. These tests are designed based on the expected input data.

Model testing shall be conducted on models that the AI system comprises. Component testing should be conducted on the automated data pre-processing steps (see ISO/IEC 23053:2022, 8.3) as part of the development process.

Testing shall also be conducted on the whole AI system that uses the models to determine if it exhibits unwanted bias. Testing the AI system is important because bias can be introduced in other human decisions in the development of the AI system, such as the data preparation and data processing steps, as well as in the user interface or other aspects of how humans interact with the system and output. Where appropriate, testing with users shall be undertaken to determine whether user interface design choices reinforce bias in the system. This class of human cognitive biases can result from factors such as:

— over-reliance on the AI system or output for the users or operators;

— improper feedback loops between the AI system and the user;

— loss of situational awareness.

AI system and ML model testing should be comprehensive enough to be representative of the expected input data in production usage. Test data separate from the training data shall be used. The comparative functional correctness of the outputs shall be evaluated amongst groups, including at-risk groups, using appropriate metrics, to determine whether the AI system meets the acceptance criteria. Evaluations should include groups to determine if differences in the quality of the outputs can be observed. The organization shall determine if these differences are acceptable against the acceptance criteria. Appropriate metrics for assessment of bias are described in ISO/IEC TR 24027:2021, Clause 7. The rationale for the selection of appropriate metrics shall be documented.

Extreme data inputs for each at-risk group shall also be tested in order to identify variations in the robustness of the model that can result in unwanted bias. The process of obtaining the extreme data inputs can benefit from the participation of AI customers and AI users in test design, test execution, results assessment and interpretation.

Evaluation of unwanted bias should be conducted in the context of the intended use and the intended operating conditions. Deploying an ML model or an AI system to a different environment or a different target population can change the degree to which it exhibits unwanted bias.

For dynamic bias testing of generative AI systems, organizations shall employ benchmarks (e.g. BBQ, Winogender, real toxicity prompts), counterfactual prompts that systematically track outcomes when switching demographic group information, and low-context prompts (e.g. “leader,” or “bad guys”).

5.4 Re-evaluation, continuous validation, operations and monitoring

5.4.1 General

An organization shall document and adhere to its approach to ensuring that the risk treatments identified to address unwanted bias (see 5.1.5) continue to have the desired effect on an ongoing basis. This approach shall include criteria for when all verification activities shall be conducted again.

This approach should:

— provide appropriate methods to treat unwanted bias occurring without detection;

— consider any continuous learning or model retraining;

— consider changes in surrounding data pipelines and user interfaces.

The following additional risk treatments should be considered in the approach:

— monitoring of unwanted bias during operation;

— monitoring of changes in the training data profile where model retraining occurs;

— technical mechanisms to alert operators if processing is outside of defined usage limits;

— technical mechanisms to identify if production data contains inputs that are outside those assessed in verification activities;

— mechanisms for end users to identify potential unwanted bias and bring it to the attention of operators;

— meaningful human oversight;

— redress for adverse outcomes.

Organizations shall ensure there is appropriate logging to support and monitor the selected unwanted bias treatments.

5.4.2 External change

Depending on their significance, external changes can necessitate a full reassessment of the AI system for unwanted bias. 5.1.2 indicates that the implementation of techniques of evaluation for bias should consider:

— relevant groups of identified stakeholders and users;

— geographical, linguistic, socioeconomic or cultural contexts;

— the current operating conditions of the AI system.

External changes in the use of AI systems can affect bias after AI systems are evaluated and deployed. Monitoring can provide relevant information about external changes that necessitate a reassessment for unwanted bias.

Additional elements monitored for external change are:

— accessibility concerns, including accessibility to opt-outs and redress mechanisms and accommodations for people with disabilities;

— monitoring effectiveness of bias mitigation controls (applied bias mitigation approaches can fail or worsen bias risks);

— ongoing user feedback.

Organizations should monitor and document external change affecting risks of unwanted bias, in compliance with the organization's risk management process.

The following are examples of how some external changes can affect decision-making processes:

— Deployment of an existing AI system to a different environment, including different users, target markets or data sources, can change the risks and require the AI system to be reassessed for unwanted bias.

— Over time, the relationship between the real-world phenomena reflected in the inputs and outputs of the system can change, often called "drift". For example, the performance of a system using an ML model to make decisions based on correlations established during the initial ML model training can deteriorate or become more biased if those correlations change over time.

— New use cases for the system can develop, either deliberately or organically, with an effect on risks related to unwanted bias.

— Societal norms can change over time (for example, attitudes towards gender behavioural norms, ideal body shape or smoking).

Bias in an AI system should be re-evaluated for changes (such as to metrics, risks, stakeholders or requirements) and addressed accordingly.

5.5 Disposal

ISO/IEC 5338:2023, 6.4.17 specifies the final process in the life cycle as disposal.

ISO/IEC 5338 explains the purpose of the disposal process as to end the existence of a system or component appropriately. This can require consideration of contracts, policies and environmental, legal, safety or security aspects.

ISO/TR 8344:2024, 5.9 contains concerns about disposition of individual records in structured data environment. ISO/TR 8344:2024, 5.10 contains concerns about conflict between disposition and the referential integrity rule for relational databases.

ISO 15489-1:2016, 8.5 contains recommendations to establish disposition authorities to govern disposition, and ISO 15489-1:2016, 9.9 contains recommendations for the disposition process.

An organization should ensure that, as appropriate, documentation and data are retained after system or component disposal:

— to support repurposing of the datasets or ML model;

— for investigating, after disposal, any historical incidents or potential existing unwanted bias in the AI system whose data has been disposed.

Data are important to AI systems. ISO/IEC/IEEE 12207 recommends that the disposal process is extended to consider the disposal of data. Disposal of data can introduce new issues in relation to retention, security and privacy.

6 Techniques to address unwanted bias

6.1 General

Unwanted bias can be treated through techniques in algorithmic development, training, data handling and organizational processes. Examples of such techniques are provided in this Clause. While the AI system life cycle stage applicable to these techniques is discussed, applicable techniques should be selected during the life cycle inception stage whenever possible.

NOTE Some of these techniques are briefly introduced in ISO/IEC TR 24027.

6.2 Algorithmic and training techniques

6.2.1 General

An AI system can be composed of one or more ML models either used independently or in combination. When multiple ML models are used in combination, data bias can be amplified at a system level.

Changes to model specifications that can be used to address unwanted bias include the following:

— Application of regularization techniques can ensure predictions are not extreme or overfitting. It can be important to document the details of such techniques applied, especially the mathematical restrictions enforced. For example, when introducing a regularization term on top of a loss function for the cost function, the regularization term can prioritize learnings from under-sampled data to ensure such learning is not forgotten due to dominant data samples. This is important when data distribution is non-IID.

— Constraints can be applied to ensure model objective functions or behaviours follow known causal relationships and do not learn demographic proxies or other incorrect behaviours that can lead to biased outcomes. Common approaches include constrained optimization, monotonicity or shape constraints and interaction constraints.

— Dual objective functions can be constructed such that model updates incorporate error reduction and improvements to a selected bias measurement. For example, an objective can be constructed to minimize logarithmic loss and maximize average impact ratios across groups.

— Changing decision thresholds can impact the fraction of AI subjects receiving positive outcomes.

— Widening or changing bins in discretized input variables can change learned relationships between population segments and positive outcomes.

— Using suites of metrics and disaggregated metrics can obtain a fuller picture and avoid gaming.

— Sampling from or averaging the top-*k* predictions can be applied instead of taking the argmax.

— Application of a decoupled classifiers[22] technique using different classifiers for diverse groups based on the requirements to address unwanted bias.

— Model selection that considers a large family of models trained with different hyperparameter or input feature specifications and trade-offs between performance and bias measurements.

— Rules applied to model predictions that mitigate foreseeable instances of unwanted bias or harm.

— Introducing explainable AI techniques[23] during development can be performed in a model-specific or model-agnostic mode and helps explain post-hoc the predictions output by the AI model. This explanation can be directed towards detection and monitoring of unwanted bias. The combination of explainable AI with machine learning operations[24][25] offers a methodology for an organization to automate monitoring for unwanted bias and trigger corrective steps to address the bias.

— Training over a distributed dataset using methodologies such as federated learning can facilitate access to previously unavailable datasets. Access to more data can address data distribution challenges, i.e. non-IID. This can improve AI system accuracy and reduce unwanted bias. A good example of the efficacy of federated learning is seen for disease detection in the healthcare sector.[26] Other variants of distributed training such as incremental learning, cyclic learning or a combination of these methodologies can also be adapted to reduce unwanted bias.

— Adversarial machine learning approaches can be applied to penalize estimators from which predictions encode unnecessary demographic information. Training proceeds with the primary estimator to be deployed learning from training data and receiving feedback from an adversary model until the adversary can no longer predict group membership using the predictions of the primary estimator.

6.2.2 Pre-trained models

Training techniques can be applied that customize pre-trained ML models to adapt AI system results to a specific deployment. Even with attention to creating an AI system with minimal unwanted bias, it is not possible to anticipate all the diverse scenarios in real-world deployment. To account for such scenarios, organizations can use techniques including the following:

— changing temperature (or similar) settings that control randomness in system outputs;

— content moderation via rules or simpler AI systems that identify problematic output and restrict their output to users;

— fine-tuning, retraining, or transfer learning based on ground truth training data or data from within an organization;

— providing pre-approved responses for common or foreseeably problematic inputs;

— strong meta-prompts that acknowledge diverse user populations and instruct models to avoid stereotyping, disparaging, or otherwise toxic outputs;

— including user feedback and redress mechanisms.

Techniques including continuous learning[27] and transfer learning provide the required capabilities to customize the AI system and factor for unwanted bias at deployment.

6.3 Data techniques

Data techniques can apply to all stages of an AI system's life cycle. Data collection involves identifying data sources, preparing the dataset and staging data for model training. However, obtaining the desired data distribution as defined by the AI system design objectives and target user base can be a significant challenge.

Techniques that address data-related challenges include:

— improved experimental design, data collection and selection;

— up-sampling and down-sampling depending on the distribution of samples across classes;

— applying data augmentation techniques to artificially increase the dataset while reusing the existing dataset;

— accessing additional data stores to increase data representativeness relative to the target population, potentially in combination with training techniques such as federated learning;

— creating a separate dataset with known biases to test the AI system sensitivity to unwanted bias. When appropriate metrics are applied, this approach provides a view on the boundary conditions with respect to unwanted bias for the given AI system;

— filtering training datasets for low quality or toxic samples;

— avoiding or minimizing feedback loops in the data collection process or in model evaluations.

7 Handling bias in a distributed AI system life cycle

ISO/IEC 22989:2022, 8.6.2 documents the different combinations of training in terms of cloud or edge training. Additionally, with emerging application of methodologies such as federated learning, edge training need not be at a single edge location but can be distributed over a network of edge devices.

The AI system life cycle as shown in ISO/IEC 22989:2022, Figure 3 can be applied to such a distributed training deployment, at each of the training edge nodes on the distributed network and to the orchestrating entity managing the overall training, possibly on the cloud. As these distributed entities are not homogeneous in terms of implementation details of the AI system life cycle, additional considerations apply. These considerations are listed below.

— Harmonization of the AI development process across the participating nodes helps to minimize the introduction of additional bias due to variability in the development process. Harmonization approaches include the use of the same data ETL process and tools and similar training environment, an aligned validation strategy that is reflective of the data distribution at each node, managing security and privacy given the distributed nature (e.g. use of trusted execution environments), a risk strategy that includes the identification, assessment and treatment of new adverse risks emerging from the distributed setup and an overall governing strategy including the entity that is responsible for the same.

— Training methodologies such as federated learning require one orchestrating entity to combine the individual training at each distributed training node by using a specific aggregation algorithm. Selection of the appropriate aggregation algorithm is important and should be done considering the data distribution across the participating nodes. The aggregation algorithm should be cognizant of the data distribution to ensure the nodes with under-sampled data (or classes) are not missed from the aggregation or the nodes with over-sampled data (or classes) are not overemphasized in the aggregated version of the model.

— As the data distribution across the distributed network varies, the verification and validation of the developed AI system shall be adapted as necessary. During the inception stage of the AI system life cycle, the organization should develop guidance on the appropriate verification and validation strategy for the system appropriate for the data distribution. Unwanted bias should be addressed, using the guidance provided in this document, at each of the training nodes and at the orchestrating entity.