TECHNICAL REPORT

ISO/IEC TR 24027

First edition
2021-11

Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

Technologie de l'information — Intelligence artificielle (IA) — Tendance dans les systèmes de l'IA et dans la prise de décision assistée par l'IA

Reference number
ISO/IEC TR 24027:2021(E)

© ISO/IEC 2021



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2021

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

© ISO/IEC 2021 – All rights reserved

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
3.1 Artificial intelligence 1
3.2 Bias 2
4 Abbreviations **3**
5 Overview of bias and fairness **3**
5.1 General 3
5.2 Overview of bias 3
5.3 Overview of fairness 5
6 Sources of unwanted bias in AI systems **6**
6.1 General 6
6.2 Human cognitive biases 7
6.2.1 General 7
6.2.2 Automation bias 7
6.2.3 Group attribution bias 8
6.2.4 Implicit bias 8
6.2.5 Confirmation bias 8
6.2.6 In-group bias 8
6.2.7 Out-group homogeneity bias 8
6.2.8 Societal bias 9
6.2.9 Rule-based system design 9
6.2.10 Requirements bias 10
6.3 Data bias 10
6.3.1 General 10
6.3.2 Statistical bias 10
6.3.3 Data labels and labelling process 11
6.3.4 Non-representative sampling 11
6.3.5 Missing features and labels 11
6.3.6 Data processing 12
6.3.7 Simpson's paradox 12
6.3.8 Data aggregation 12
6.3.9 Distributed training 12
6.3.10 Other sources of data bias 12
6.4 Bias introduced by engineering decisions 12
6.4.1 General 12
6.4.2 Feature engineering 12
6.4.3 Algorithm selection 13
6.4.4 Hyperparameter tuning 13
6.4.5 Informativeness 14
6.4.6 Model bias 14
6.4.7 Model interaction 14
7 Assessment of bias and fairness in AI systems **14**
7.1 General 14
7.2 Confusion matrix 15
7.3 Equalized odds 16
7.4 Equality of opportunity 16
7.5 Demographic parity 17
7.6 Predictive equality 17
7.7 Other metrics 17

© ISO/IEC 2021 – All rights reserved

8 Treatment of unwanted bias throughout an AI system life cycle **17**
8.1 General 17
8.2 Inception 17
8.2.1 General 17
8.2.2 External requirements 18
8.2.3 Internal requirements 19
8.2.4 Trans-disciplinary experts 19
8.2.5 Identification of stakeholders 19
8.2.6 Selection and documentation of data sources 20
8.2.7 External change 20
8.2.8 Acceptance criteria 21
8.3 Design and development 21
8.3.1 General 21
8.3.2 Data representation and labelling 21
8.3.3 Training and tuning 22
8.3.4 Adversarial methods to mitigate bias 23
8.3.5 Unwanted bias in rule-based systems 24
8.4 Verification and validation 24
8.4.1 General 24
8.4.2 Static analysis of training data and data preparation 25
8.4.3 Sample checks of labels 25
8.4.4 Internal validity testing 25
8.4.5 External validity testing 25
8.4.6 User testing 26
8.4.7 Exploratory testing 26
8.5 Deployment 26
8.5.1 General 26
8.5.2 Continuous monitoring and validation 26
8.5.3 Transparency tools 27
Annex A (informative) **Examples of bias** **28**
Annex B (informative) **Related open source tools** **31**
Annex C (informative) **ISO 26000 – Mapping example** **32**
Bibliography **36**

© ISO/IEC 2021 – All rights reserved

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

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

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

This document was prepared by 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.

© ISO/IEC 2021 – All rights reserved

Introduction

Bias in artificial intelligence (AI) systems can manifest in different ways. AI systems that learn patterns from data can potentially reflect existing societal bias against groups. While some bias is necessary to address the AI system objectives (i.e. desired bias), there can be bias that is not intended in the objectives and thus represent unwanted bias in the AI system.

Bias in AI systems can be introduced as a result of structural deficiencies in system design, arise from human cognitive bias held by stakeholders or be inherent in the datasets used to train models. That means that AI systems can perpetuate or augment existing bias or create new bias.

Developing AI systems with outcomes free of unwanted bias is a challenging goal. AI system function behaviour is complex and can be difficult to understand, but the treatment of unwanted bias is possible. Many activities in the development and deployment of AI systems present opportunities for identification and treatment of unwanted bias to enable stakeholders to benefit from AI systems according to their objectives.

Bias in AI systems is an active area of research. This document articulates current best practices to detect and treat bias in AI systems or in AI-aided decision-making, regardless of source. The document covers topics such as:

— an overview of bias (5.2) and fairness (5.3);

— potential sources of unwanted bias and terms to specify the nature of potential bias (Clause 6);

— assessing bias and fairness (Clause 7) through metrics;

— addressing unwanted bias through treatment strategies (Clause 8).

© ISO/IEC 2021 – All rights reserved

Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

1 Scope

This document addresses bias in relation to AI systems, especially with regards to AI-aided decision-making. Measurement techniques and methods for assessing bias are described, with the aim to address and treat bias-related vulnerabilities. All AI system lifecycle phases are in scope, including but not limited to data collection, training, continual learning, design, testing, evaluation and use.

2 Normative references

ISO/IEC 22989[1)], *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053[2)], *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the following terms and definitions given in ISO/IEC 22989 and ISO/IEC 23053 and the following apply.

ISO and IEC maintain terminological 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 Artificial intelligence

3.1.1
maximum likelihood estimator
estimator assigning the value of the parameter where the likelihood function attains or approaches its highest value

Note 1 to entry: Maximum likelihood estimation is a well-established approach for obtaining parameter estimates where a distribution has been specified [for example, normal, gamma, Weibull and so forth]. These estimators have desirable statistical properties (for example, invariance under monotone transformation) and in many situations provide the estimation method of choice. In cases in which the maximum likelihood estimator is biased, a simple bias correction sometimes takes place.

[SOURCE: ISO 3534-1:2006, 1.35]

3.1.2
rule-based systems
knowledge-based system that draws inferences by applying a set of if-then rules to a set of facts following given procedures

[SOURCE: ISO/IEC 2382:2015, 2123875]

1) Under preparation. Stage at the time of publication: ISO/DIS 22989:2021.

2) Under preparation. Stage at the time of publication: ISO/DIS 23053:2021.

© ISO/IEC 2021 – All rights reserved

3.1.3
sample
<statistics> subset of a population made up of one or more sampling units

Note 1 to entry: The sampling units could be items, numerical values or even abstract entities depending on the population of interest.

Note 2 to entry: A sample from a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value population will often be referred to as a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value sample, respectively.

[SOURCE: ISO 16269-4:2010, 2.1, modified - added *<statistics>* domain]

3.1.4
knowledge
information about objects, events, concepts or rules, their relationships and properties, organized for goal-oriented systematic use

Note 1 to entry: Information can exist in numeric or symbolic form.

Note 2 to entry: Information is data that has been contextualized, so that it is interpretable. Data are created through abstraction or measurement from the world.

3.1.5
user
individual or group that interacts with a system or benefits from a system during its utilization

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1.52]

3.2 Bias

3.2.1
automation bias
propensity for humans to favour suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct

3.2.2
bias
systematic difference in treatment of certain objects, people, or groups in comparison to others

Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, prediction or decision

3.2.4
human cognitive bias
bias (3.2.2) that occurs when humans are processing and interpreting information

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

3.2.5
confirmation bias
type of human cognitive *bias* (3.2.4) that favours predictions of AI systems that confirm pre-existing beliefs or hypotheses

3.2.6
convenience sample
sample of data that is chosen because it is easy to obtain, rather than because it is representative

3.2.7
data bias
data properties that if unaddressed lead to AI systems that perform better or worse for different *groups* (3.2.8)

© ISO/IEC 2021 – All rights reserved

3.2.8
group
subset of objects in a domain that are linked because they have shared characteristics

3.2.10
statistical bias
type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates

[SOURCE: ISO 20501:2019, 3.3.9]

4 Abbreviations

AI artificial intelligence

ML machine learning

5 Overview of bias and fairness

5.1 General

In this document, the term bias is defined as a systematic difference in the treatment of certain objects, people, or groups in comparison to others, in its generic meaning beyond the context of AI or ML. In a social context, bias has a clear negative connotation as one of the main causes of discrimination and injustice. Nevertheless, it is the systematic differences in human perception, observation and the resultant representation of the environment and situations that make the operation of ML algorithms possible.

This document uses the term bias to characterize the input and the building blocks of AI systems in terms of their design, training and operation. AI systems of different types and purposes (such as for labelling, clustering, making predictions or decisions) rely on those biases for their operation.

To characterize the AI system outcome or, more precisely, its possible impact on society, this document uses the terms unfairness and fairness, instead. Fairness can be described as a treatment, a behaviour or an outcome that respects established facts, beliefs and norms and is not determined by favouritism or unjust discrimination.

While certain biases are essential for proper AI system operation, unwanted biases can be introduced into an AI system unintentionally and can lead to unfair system results.

5.2 Overview of bias

AI systems are enabling new experiences and capabilities for people around the globe. AI systems can be used for various tasks, such as recommending books and television shows, predicting the presence and severity of a medical condition, matching people to jobs and partners or identifying if a person is crossing the street. Such computerized assistive or decision-making systems have the potential to be fairer and the risk of being less fair than existing systems or humans that they will be augmenting or replacing.

AI systems often learn from real-world data; hence an ML model can learn or even amplify problematic pre-existing data bias. Such bias can potentially favour or disfavour certain groups of people, objects, concepts or outcomes. Even given seemingly unbiased data, the most rigorous cross-functional training and testing can still result in an ML model with unwanted bias. Furthermore, the removal or reduction of one kind of bias (e.g. societal bias) can involve the introduction or increase of another kind of bias (e.g. statistical bias)[3], see positive impact described in this clause. Bias can have negative, positive or neutral impact.

© ISO/IEC 2021 – All rights reserved