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 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) 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.

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).

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.

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)

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.

Before discussing aspects of bias in AI systems, it is necessary to describe the operation of AI systems and what unwanted bias means in this context. An AI system can be characterized as using knowledge to process input data to make predictions or take actions. The knowledge within an AI system is often built through a learning process from training data; it consists of statistical correlations observed in the training dataset. It is essential for both the production data and the training data to relate to the same area of interest.

The predictions made by AI systems can be highly varied, depending on the area of interest and the type of the AI system. However, for classification systems, it is useful to think of the AI predictions as processing the set of input data presented to it and predicting that the input belongs to a desired set or not. A simple example is that of making a prediction relating to a loan application as to whether the applicant represents an acceptable financial risk or not to the lending organization.

A desirable AI system would correctly predict whether the application represents an acceptable risk without contributing to systemic exclusion of certain groups. This can mean in some circumstances taking into account considerations of certain groups, such as ethnicity and gender. There can be an effect of bias on the resulting environment where the prediction can change the results of subsequent predictions. Examples of how to determine whether an algorithm has unwanted bias according to the metrics defined in Clause 7, are given in Annex A.

Uncovering bias can involve defining appropriate criteria and analysing trade-offs associated with these criteria. Given particular criteria, this document describes methodologies and mechanisms for uncovering and treating bias in AI systems.

Classification (a type of supervised learning) and clustering (a type of unsupervised learning) algorithms cannot function without bias. If all subgroups are to be treated equally, then these kinds of algorithms would have to label all outputs the same (resulting in only one class or cluster). However, investigation would be necessary to assess whether the impact of this bias is positive, neutral or negative according to the system goals and objectives.

Examples of positive, neutral and negative effects of bias are as follows:

— Positive effect: AI developers can introduce bias to ensure a fair result. For example, an AI system used for hiring a specific type of worker can introduce a bias towards one gender over another in the decision phase to compensate for societal bias inherited from the data, which reflects their historical underrepresentation in this profession.

— Neutral effect: The AI system for processing images for a self-driving car system can systematically misclassify “mailboxes” as “fire hydrants”. However, this statistical bias will have neutral impact, as long as the system has an equally strong preference for avoiding each type of obstacle.

— Negative effect: Examples of negative impacts include AI hiring systems favouring candidates of one gender over another and voice-based digital assistants failing to recognize people with speech impairments. Each of these instances can have unintended consequences of limiting the opportunities of those affected. While such examples can be categorized as unethical, bias is a wider concept that applies even in scenarios with no adverse effect on stakeholders, for example, in the classification of galaxies by astrophysicists.

One challenge with determining the relevance of bias is that what constitutes negative effect can depend on the specific use case or application domain. For example, age-based profiling can be considered unacceptable in job application decisions. However, age can play a critical role in evaluation of medical procedures and treatment. Appropriate customization specific to the use case or application domain can be considered.

In ML systems, the outcome of any single operation is based upon correlations between features in the input domain and previously observed outputs. Any incorrect outputs (including for example, automated decisions, classifications and predicted continuous variables) are potentially due to poor generalization, the outputs used to train the ML model and the hyperparameters used to calibrate it. Statistical bias in the ML model can be introduced inadvertently or due to bias in the data collection and modelling process. In symbolic AI systems, human cognitive bias can lead to specifying explicit

knowledge inaccurately, for example specifying rules that apply to oneself, but not the target user, due to in-group bias.

Another concern about bias is the ease with which it can be propagated into a system, after which it can be challenging to recognize and mitigate. An example of this is where data reflects a bias that exists already in society and this bias becomes part of a new AI system that then propagates the original bias.

Organisations can consider the risk of unwanted bias in datasets and algorithms, including those that at first glance appear harmless and safe. In addition, once attempts at removing unwanted bias have been made, unintended categorisation and unsophisticated algorithms have the potential to perpetuate or amplify existing bias. As a consequence, unwanted bias mitigation is not a “set-and-forget” process.

For example, a resume review algorithm that favours candidates with years of continuous service would automatically disadvantage carers who are returning to the workforce after having taken time off work for caring responsibilities. A similar algorithm can also downgrade casual workers whose working history consists of many short contracts for a wide variety of employers: a characteristic that can be misinterpreted as negative. Careful re-evaluation of the newly achieved outcomes can follow any unwanted bias reduction and retraining of the algorithm.

The more automated the system and the less effective the human oversight, the likelihood of unintended negative consequences is heightened. This situation is compounded when multiple AI applications contribute to the automation of a given task. In such multi-application AI systems, greater demand for transparency and explainability regarding the outcomes it produces can be anticipated by the organisations deploying them.

5.3 Overview of fairness

Fairness is a concept that is distinct from, but related to bias. Fairness can be characterized by the effects of an AI system on individuals, groups of people, organizations and societies that the system influences. However, it is not possible to guarantee universal fairness. Fairness as a concept is complex, highly contextual and sometimes contested, varying across cultures, generations, geographies and political opinions. What is considered fair can be inconsistent across these contexts. This document thus does not define the term fairness because of its highly socially and ethically contextual nature.

Even within the context of AI, it is difficult to define fairness in a manner that will apply equally well to all AI systems in all contexts. An AI system can potentially affect individuals, groups of people, organizations and societies in many undesirable ways. Common categories of negative impacts that can be perceived as “unfair” include:

— Unfair allocation: occurs when an AI system unfairly extends or withholds opportunities or resources in ways that have negative effects on some parties as compared to others.

— Unfair quality of service: occurs when an AI system performs less well for some parties than for others, even if no opportunities or resources are extended or withheld.

— Stereotyping: occurs when an AI system reinforces existing societal stereotypes.

— Denigration: occurs when an AI system behaves in ways that are derogatory or demeaning.

— “Over“ or “under“ representation and erasure: occurs when an AI system over-represents or under-represents some parties as compared to others, or even fails to represent their existence.

Bias is just one of many elements that can influence fairness. It has been observed that biased inputs do not always result in unfair predictions and actions and unfair predictions and actions are not always caused by bias.

An example of a biased decision system that can nonetheless be considered fair is a university hiring policy that is biased in favour of people with relevant qualifications, in that it hires a far greater proportion of holders of relevant qualifications than the proportion of relevant qualification holders in the population. As long as the determination of relevant qualifications does not discriminate against particular demographics, such a system can be considered fair.

An example of an unbiased system that can be considered unfair, is a policy that indiscriminately rejected all candidates. Such a policy would indeed be unbiased, as not differentiating between any categories. But it would be perceived as unfair by people with relevant qualifications.

This document distinguishes between bias and fairness. Bias can be societal or statistical, can be reflected in or arise from different system components (see Clause 6) and can be introduced or propagated at different stages of the AI development and deployment life cycle (see Clause 8).

Achieving fairness in AI systems often means making trade-offs. In some cases, different stakeholders can have legitimately conflicting priorities that cannot be reconciled by an alternative system design. As an example, consider an AI system that decides the award of scholarships to some of the graduate programme applicants in a university. The diversity stakeholder in the admissions office wants the AI system to provide a fair distribution of such awards to applications from various geographic regions. On the other hand, a professor, who is another stakeholder, wants a particular deserving student interested in a particular research area to be awarded the scholarship. In such a case, there is a possibility that the AI system denies a deserving candidate from a particular region in order to meet the research objectives. Thus, meeting the fairness expectations of all stakeholders is not always possible. It is therefore important to be explicit and transparent about those priorities and any underlying assumptions, in order to correctly select the relevant metrics (see Clause 7).

6 Sources of unwanted bias in AI systems

6.1 General

This clause describes possible sources of unwanted bias in AI systems. This includes human cognitive bias, data bias and bias introduced by engineering decisions. Figure 1 shows the relationship between these high-level groups of biases. The human cognitive biases (6.2) can cause bias to be introduced through engineering decisions (6.4), or data bias (6.3).

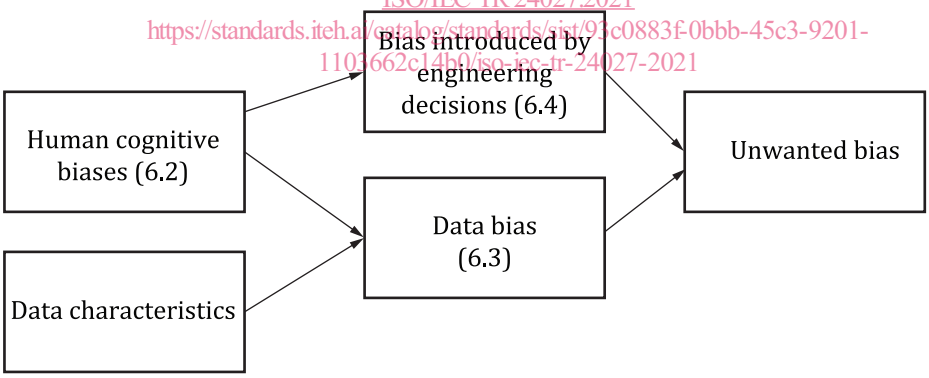

Figure 1 — Relationship between high-level groups of bias

For example, written or spoken language contains societal bias which can be amplified by word embedding models[4]. Because societal bias is reflected in existing language that is used as training data, it in turn causes non-representative sampling data bias (described in 6.3.4), which can lead to unwanted bias. This relationship is shown in Figure 2.

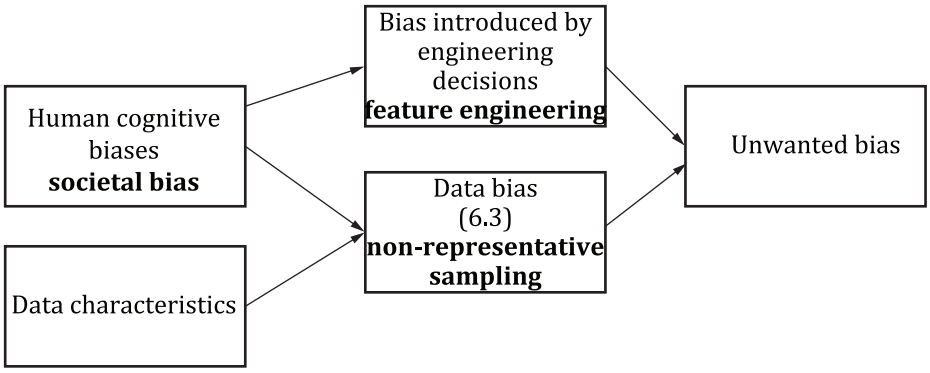

Figure 2 — Example of societal bias manifesting as unwanted bias

Systems are likely to exhibit multiple sources of bias simultaneously. Analysing a system to detect one source of bias is unlikely to uncover all. In the same example, multiple models are used for natural language processing. The outputs of the word embedding model that may be affected by non-representative sampling bias are then further processed by a secondary model. In this case, the secondary model is vulnerable to bias in feature engineering because a choice was made to use word embeddings as features of this model.

Not all sources of bias start with human cognitive biases, bias can be caused exclusively by data characteristics. For example, sensors that are attached to a system may fail and produce signals that can be considered outliers (see 6.3.10). This data, when used for training or reinforcement learning, can introduce unwanted bias. This is shown in Figure 3.

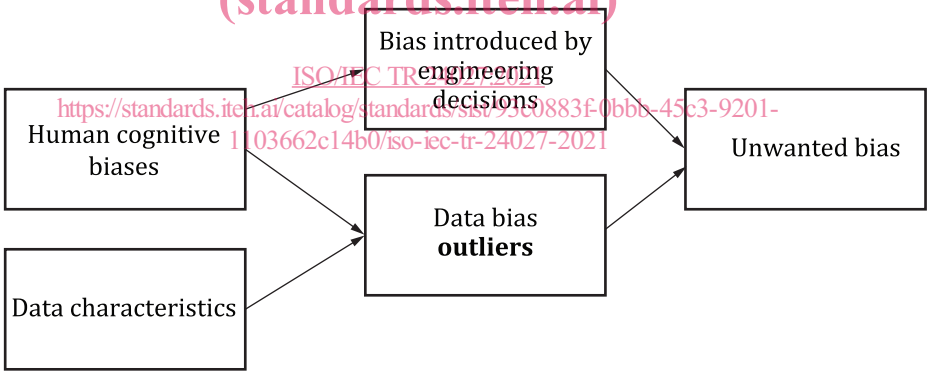

Figure 3 — Example of data characteristics manifesting as unwanted bias

6.2 Human cognitive biases

6.2.1 General

Human beings can be biased in different ways, both consciously and unconsciously, and are influenced by the data, information and experiences available to them for making decisions[5]. Thinking is often based on opaque processes that lead humans to make decisions without always knowing what leads to them. These human cognitive biases affect decisions about data collection and processing, system design, model training and other development decisions that individuals make, as well as decisions about how a system is used.

6.2.2 Automation bias

AI assists automation of analysis and decision-making in various systems, for example in self-driving cars and health-care systems, that can invite automation bias. Automation bias occurs when a human