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)*

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.