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/IEC 2024



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2024

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

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
3.1 General ... 1
3.2 Artificial intelligence ... 3
3.3 Bias ... 4
3.4 Testing ... 5
4 Abbreviated terms ... **6**
5 Treating unwanted bias in the AI system life cycle ... **6**
5.1 Inception ... 6
5.1.1 Stakeholder identification ... 6
5.1.2 Stakeholder needs and requirements definition ... 7
5.1.3 Procurement ... 8
5.1.4 Data sources ... 9
5.1.5 Integration with risk management ... 11
5.1.6 Acceptance criteria ... 11
5.2 Design and development ... 12
5.2.1 Feature representation ... 12
5.2.2 Metadata sufficiency ... 12
5.2.3 Data annotations ... 12
5.2.4 Adjusting data ... 13
5.2.5 Methods for managing identified risks ... 13
5.3 Verification and validation ... 13
5.3.1 General ... 13
5.3.2 Static testing of data used in development ... 14
5.3.3 Dynamic testing ... 14
5.4 Re-evaluation, continuous validation, operations and monitoring ... 15
5.4.1 General ... 15
5.4.2 External change ... 16
5.5 Disposal ... 17
6 Techniques to address unwanted bias ... **17**
6.1 General ... 17
6.2 Algorithmic and training techniques ... 17
6.2.1 General ... 17
6.2.2 Pre-trained models ... 18
6.3 Data techniques ... 19
7 Handling bias in a distributed AI system life cycle ... **19**
Annex A (informative) **Life cycle processes map** ... **21**
Annex B (informative) **Potential impacts of unwanted bias on different types of specific user** ... **22**
Bibliography ... **23**

© ISO/IEC 2024 – All rights reserved

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.

© ISO/IEC 2024 – All rights reserved

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.

© ISO/IEC 2024 – All rights reserved

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.]

© ISO/IEC 2024 – All rights reserved