TECHNICAL REPORT

ISO/IEC TR 24028

First edition
2020-05

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

Technologies de l'information — Intelligence artificielle — Examen d'ensemble de la fiabilité en matière d'intelligence artificielle

This document is a preview generated by EVS

Reference number
ISO/IEC TR 24028:2020(E)

© ISO/IEC 2020

This document is a preview generated by EVS

COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2020

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
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

 © ISO/IEC 2020 – All rights reserved

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Overview **7**
5 Existing frameworks applicable to trustworthiness **7**
5.1 Background 7
5.2 Recognition of layers of trust 8
5.3 Application of software and data quality standards 8
5.4 Application of risk management 10
5.5 Hardware-assisted approaches 10
6 Stakeholders **11**
6.1 General concepts 11
6.2 Types 12
6.3 Assets 12
6.4 Values 13
7 Recognition of high-level concerns **13**
7.1 Responsibility, accountability and governance 13
7.2 Safety 14
8 Vulnerabilities, threats and challenges **14**
8.1 General 14
8.2 AI specific security threats 15
8.2.1 General 15
8.2.2 Data poisoning 15
8.2.3 Adversarial attacks 15
8.2.4 Model stealing 16
8.2.5 Hardware-focused threats to confidentiality and integrity 16
8.3 AI specific privacy threats 16
8.3.1 General 16
8.3.2 Data acquisition 16
8.3.3 Data pre-processing and modelling 17
8.3.4 Model query 17
8.4 Bias 17
8.5 Unpredictability 17
8.6 Opaqueness 18
8.7 Challenges related to the specification of AI systems 18
8.8 Challenges related to the implementation of AI systems 19
8.8.1 Data acquisition and preparation 19
8.8.2 Modelling 19
8.8.3 Model updates 21
8.8.4 Software defects 21
8.9 Challenges related to the use of AI systems 21
8.9.1 Human-computer interaction (HCI) factors 21
8.9.2 Misapplication of AI systems that demonstrate realistic human behaviour 22
8.10 System hardware faults 22
9 Mitigation measures **23**
9.1 General 23
9.2 Transparency 23
9.3 Explainability 24
9.3.1 General 24

This document is a preview generated by EVS

© ISO/IEC 2020 – All rights reserved

9.3.2 Aims of explanation ... 24
9.3.3 Ex-ante vs ex-post explanation ... 24
9.3.4 Approaches to explainability ... 25
9.3.5 Modes of ex-post explanation ... 25
9.3.6 Levels of explainability ... 26
9.3.7 Evaluation of the explanations ... 27
9.4 Controllability ... 27
9.4.1 General ... 27
9.4.2 Human-in-the-loop control points ... 28
9.5 Strategies for reducing bias ... 28
9.6 Privacy ... 28
9.7 Reliability, resilience and robustness ... 28
9.8 Mitigating system hardware faults ... 29
9.9 Functional safety ... 29
9.10 Testing and evaluation ... 30
9.10.1 General ... 30
9.10.2 Software validation and verification methods ... 30
9.10.3 Robustness considerations ... 32
9.10.4 Privacy-related considerations ... 33
9.10.5 System predictability considerations ... 33
9.11 Use and applicability ... 34
9.11.1 Compliance ... 34
9.11.2 Managing expectations ... 34
9.11.3 Product labelling ... 34
9.11.4 Cognitive science research ... 34
10 Conclusions ... **34**
Annex A (informative) **Related work on societal issues** ... **36**
Bibliography ... **37**

This document is a preview generated by EVS

 © ISO/IEC 2020 – 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 document 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 and IEC 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) or the IEC list of patent declarations received (see http://patents.iec.ch).

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 Joint 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.

This document is a preview generated by EVS

© ISO/IEC 2020 – All rights reserved

Introduction

The goal of this document is to analyse the factors that can impact the trustworthiness of systems providing or using AI, called hereafter artificial intelligence (AI) systems. The document briefly surveys the existing approaches that can support or improve trustworthiness in technical systems and discusses their potential application to AI systems. The document discusses possible approaches to mitigating AI system vulnerabilities that relate to trustworthiness. The document also discusses approaches to improving the trustworthiness of AI systems.

This document is a preview generated by EVS

 © ISO/IEC 2020 – All rights reserved

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

This document is a preview generated by EVS

1 Scope

This document surveys topics related to trustworthiness in AI systems, including the following:

— approaches to establish trust in AI systems through transparency, explainability, controllability, etc.;

— engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and

— approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security and privacy of AI systems.

The specification of levels of trustworthiness for AI systems is out of the scope of this document.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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 http://www.electropedia.org/

3.1
accountability
property that ensures that the actions of an *entity* (3.16) may be traced uniquely to that entity

[SOURCE: ISO/IEC 2382:2015, 2126250, modified — The Notes to entry have been removed.]

3.2
actor
entity (3.16) that communicates and interacts

[SOURCE: ISO/IEC TR 22417:2017, 3.1]

3.3
algorithm
set of rules for transforming the logical representation of *data* (3.11)

[SOURCE: ISO/IEC 11557:1992, 4.3]

3.4
artificial intelligence
AI
capability of an engineered *system* (3.38) to acquire, process and apply knowledge and skills

Note 1 to entry: Knowledge are facts, *information* (3.20) and skills acquired through experience or education.

© ISO/IEC 2020 – All rights reserved