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

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

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

Contents

Page

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.

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.

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

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.

3.5
asset
anything that has *value* (3.46) to a *stakeholder* (3.37)

Note 1 to entry: There are many types of assets, including:

a) *information* (3.20);

b) software, such as a computer program;

c) physical, such as computer;

d) services;

e) people and their qualifications, skills and experience; and

f) intangibles, such as reputation and image.

[SOURCE: ISO/IEC 21827:2008, 3.4, modified — In the definition, "the organization" has been changed to "a stakeholder". Note 1 to entry has been removed.]

3.6
attribute
property or characteristic of an object that can be distinguished quantitatively or qualitatively by human or automated means

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.2]

3.7
autonomy
autonomous
characteristic of a *system* (3.38) governed by its own rules as the result of self-learning

Note 1 to entry: Such systems are not subject to external *control* (3.10) or oversight.

3.8
bias
favouritism towards some things, people or groups over others

3.9
consistency
degree of uniformity, standardization and freedom from contradiction among the documents or parts of a *system* (3.38) or component

[SOURCE: ISO/IEC 21827:2008, 3.14]

3.10
control
purposeful action on or in a *process* (3.29) to meet specified objectives

[SOURCE: IEC 61800-7-1:2015, 3.2.6]

3.11
data
re-interpretable representation of *information* (3.20) in a formalized manner suitable for communication, interpretation or processing

Note 1 to entry: *Data* (3.11) can be processed by human or automatic means.

[SOURCE: ISO/IEC 2382:2015, 2121272, modified — Notes 2 and 3 to entry have been removed.]

3.12
data subject
individual about whom *personal data* (3.27) are recorded

[SOURCE: ISO 5127:2017, 3.13.4.01, modified — Note 1 to entry has been removed.]

3.13
decision tree
supervised-learning model for which inference can be represented by traversing one or more tree-like structures

3.14
effectiveness
extent to which planned activities are realized and planned results achieved

[SOURCE: ISO 9000:2015, 3.7.11, modified — Note 1 to entry has been removed.]

3.15
efficiency
relationship between the results achieved and the resources used

[SOURCE: ISO 9000:2015, 3.7.10]

3.16
entity
any concrete or abstract thing of interest

[SOURCE: ISO/IEC 10746-2:2009, 6.1]

3.17
harm
injury or damage to the health of people or damage to property or the environment

[SOURCE: ISO/IEC Guide 51:2014, 3.1]

3.18
hazard
potential source of *harm* (3.17)

[SOURCE: ISO/IEC Guide 51:2014, 3.2]

3.19
human factors
environmental, organizational and job factors, in conjunction with cognitive human characteristics, which influence the behaviour of persons or organizations

3.20
information
meaningful *data* (3.11)

[SOURCE: ISO 9000:2015, 3.8.2]

3.21
integrity
property of protecting the accuracy and completeness of *assets* (3.5)

[SOURCE: ISO/IEC 27000:2018, 3.36, modified — In the definition, "protecting the" has been added before "accuracy" and "of assets" has been added after "completeness".]

3.22
intended use
use in accordance with *information* (3.20) provided with a product or *system* (3.38) or, in the absence of such information, by generally understood *patterns* (3.26) of usage.

[SOURCE: ISO/IEC Guide 51:2014, 3.6]

3.23
machine learning
ML
process (3.29) by which a functional unit improves its performance by acquiring new knowledge or skills or by reorganizing existing knowledge or skills

[SOURCE: ISO/IEC 2382:2015, 2123789]

3.24
machine learning model
mathematical construct that generates an inference or prediction, based on input *data* (3.11)

3.25
neural network
computational model utilizing distributed, parallel local processing and consisting of a network of simple processing elements called artificial neurons, which can exhibit complex global behaviour

[SOURCE: ISO 18115-1:2013, 8.1]

3.26
pattern
set of features and their relationships used to recognize an *entity* (3.16) within a given context

[SOURCE: ISO/IEC 2382:2015, 2123798]

3.27
personal data
data (3.11) relating to an identified or identifiable individual

[SOURCE: ISO 5127:2017, 3.1.10.14, modified — The admitted terms and Notes 1 and 2 to entry have been removed.]

3.28
privacy
freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of *data* (3.11) about that individual

[SOURCE: ISO/IEC 2382:2015, 2126263, modified — Notes 1 and 2 to entry have been removed.]

3.29
process
set of interrelated or interacting activities that use inputs to deliver an intended result

[SOURCE: ISO 9000:2015, 3.4.1, modified — The notes to entry have been omitted.]

3.30
reliability
property of consistent intended behaviour and results

[SOURCE: ISO/IEC 27000:2018, 3.55]

3.31
risk
effect of uncertainty on objectives

Note 1 to entry: An effect is a deviation from the expected. It can be positive, negative or both and can address, create or result in opportunities and *threats* (3.39).

Note 2 to entry: Objectives can have different aspects and categories and can be applied at different levels.

Note 3 to entry: Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood.

[SOURCE: ISO 31000:2018, 3.1]

3.32
robot
programmed actuated mechanism with a degree of *autonomy* (3.7), moving within its environment, to perform intended tasks

Note 1 to entry: A robot includes the *control* (3.10) system and interface of the control *system* (3.38).

Note 2 to entry: The classification of robot into industrial robot or service robot is done according to its intended application.

[SOURCE: ISO 18646-2:2019, 3.1]

3.33
robotics
science and practice of designing, manufacturing and applying *robots* (3.32)

[SOURCE: ISO 8373:2012, 2.16]

3.34
safety
freedom from *risk* (3.31) which is not tolerable

[SOURCE: ISO/IEC Guide 51:2014, 3.14]

3.35
security
degree to which a product or *system* (3.38) protects *information* (3.20) and *data* (3.11) so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization

[SOURCE: ISO/IEC 25010:2011, 4.2.6]

3.36
sensitive data
data (3.11) with potentially harmful effects in the event of disclosure or misuse

[SOURCE: ISO 5127:2017, 3.1.10.16]

3.37
stakeholder
any individual, group or organization that can affect, be affected by or perceive itself to be affected by a decision or activity

[SOURCE: ISO/IEC 38500:2015, 2.24]

3.38
system
combination of interacting elements organized to achieve one or more stated purposes

Note 1 to entry: A system is sometimes considered as a product or as the services it provides.

[SOURCE: ISO/IEC/IEEE 15288:2015, 3.38]

3.39
threat
potential cause of an unwanted incident, which may result in *harm* (3.17) to *systems* (3.38), organizations or individuals

3.40
training
process (3.29) to establish or to improve the parameters of a *machine learning model* (3.24) based on a machine learning *algorithm* (3.3) by using training *data* (3.11)

3.41
trust
degree to which a *user* (3.43) or other *stakeholder* (3.37) has confidence that a product or *system* (3.38) will behave as intended

[SOURCE: ISO/IEC 25010:2011, 4.1.3.2]

3.42
trustworthiness
ability to meet *stakeholders'* (3.37) expectations in a verifiable way

Note 1 to entry: Depending on the context or sector and also on the specific product or service, *data* (3.11) and technology used, different characteristics apply and need *verification* (3.47) to ensure stakeholders expectations are met.

Note 2 to entry: Characteristics of trustworthiness include, for instance, *reliability* (3.30), availability, resilience, *security* (3.35), *privacy* (3.28), *safety* (3.34), *accountability* (3.1), transparency, *integrity* (3.21), authenticity, quality, usability.

Note 3 to entry: Trustworthiness is an *attribute* (3.6) that can be applied to services, products, technology, data and *information* (3.20) as well as, in the context of governance, to organizations.

3.43
user
individual or group that interacts with a *system* (3.38) or benefits from a system during its utilization

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1.52, modified — Note 1 to entry has been removed.]

3.44
validation
confirmation, through the provision of objective evidence, that the requirements for a specific *intended use* (3.22) or application have been fulfilled

Note 1 to entry: The right *system* (3.38) was built.

[SOURCE: ISO/IEC TR 29110-1:2016, 3.73, modified — Only the last sentence of Note 1 to entry has been retained and Note 2 to entry has been removed.]

3.45
value
<data> unit of *data* (3.11)

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.41]

3.46
value
<social> belief(s) an organization adheres to and the standards that it seeks to observe

[SOURCE: ISO 10303-11:2004, 3.3.22]

3.47
verification
confirmation, through the provision of objective evidence, that specified requirements have been fulfilled

Note 1 to entry: The *system* (3.38) was built right.

[SOURCE: ISO/IEC TR 29110-1:2016, 3.74, modified — Only the last sentence of Note 1 to entry has been retained.]

3.48
vulnerability
weakness of an *asset* (3.5) or *control* (3.10) that can be exploited by one or more *threats* (3.38)

[SOURCE: ISO/IEC 27000:2018, 3.77]

3.49
workload
mix of tasks typically run on a given computer *system* (3.38)

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.4618, modified — Note 1 to entry has been removed.]

4 Overview

This document provides an overview of topics relevant to building trustworthiness of AI systems. One of the goals of this document is to assist the standards community with identifying specific standardization gaps in the area of AI.

In Clause 5, the document briefly surveys existing approaches being used for building trustworthiness in technical systems and discusses their potential applicability to AI systems. In Clause 6, the document identifies the stakeholders. In Clause 7, it discusses their considerations related to the responsibility, accountability, governance and safety of AI systems. In Clause 8, the document surveys the vulnerabilities of AI systems that can reduce their trustworthiness. In Clause 9, the document identifies possible measures that improve trustworthiness of an AI system by mitigating vulnerabilities across its lifecycle. Measures include those related to improving AI system transparency, controllability, data handling, robustness, testing and evaluation and use. Conclusions are presented in Clause 10.

5 Existing frameworks applicable to trustworthiness

5.1 Background

For the purposes of this document, it is important to provide working definitions of artificial intelligence (AI) systems and trustworthiness.

Here, we consider an AI system to be any system (whether a product or a service) that uses AI. There are many different kinds of AI systems. Some are implemented completely in software, while others are mostly implemented in hardware (e.g. robots).

A working definition of trustworthiness is the ability to meet stakeholders' expectations in a verifiable way. This definition can be applied across the broad range of AI systems, technologies and application domains.

As with security, trustworthiness has been understood and treated as a non-functional requirement specifying emergent properties of a system — i.e. a set of characteristics with their attributes — within the context of quality of use. This is indicated in ISO/IEC 25010[20].

Additionally, like with security, trustworthiness can be improved through an organizational process with specific measurable outcomes and key performance indicators (KPIs).

In summary, trustworthiness has been understood and treated as both an ongoing organizational process as well as a (non-functional) requirement.

According to UNEP[26], the "precautionary principle" means that where there are threats of serious or irreversible harm, lack of full scientific certainty shall not be used as a reason for postponing effective measures to prevent harm. In safety engineering, a process for capturing and then sizing, stakeholder "value" requirements includes the understanding of the system's context of use, the risks of harm and, when applicable, an application of the "precautionary principle" as a risk mitigation technique against potential unintended consequences, such as harm to rights and freedom of natural persons, life of any kind, the environment, a species or a community.

AI systems are often existing systems enhanced with AI capabilities. In this case, all the approaches and considerations regarding trustworthiness that applied to the old version of the system, continue to apply to enhanced system. These include approaches to quality (both metrics and measurement methodologies), safety and risk of harm and risk management frameworks (such as those existing for security and privacy). Subclauses 5.2 to 5.5 present different frameworks for contextualizing the trustworthiness of AI systems.

5.2 Recognition of layers of trust

An AI system can be conceptualized as operating in an ecosystem of functional layers. Trust is established and maintained at each layer in order for the AI system to be trusted in its environment. For example, the ITU-T report on Trust Provisioning[27] introduces three layers of trust: physical trust, cyber trust and social trust, taking into account the physical infrastructure for data collection (e.g. sensors and actuators), IT infrastructure for data storage and processing (e.g. cloud) and end-applications (e.g. ML algorithms, expert systems and applications for end-users).

Regarding the layer of physical trust, the concept is often synonymous to the combination of reliability and safety because the metrics are based on a physical measurement or test. For instance, the technical control of a car makes the car and its inner mechanisms trustworthy. In this context, the level of trust can be determined through the level of fulfilment of a checklist. In addition, some processes such as sensor calibration can guarantee the correctness of measurements and, therefore, the data produced.

At the cyber trust layer, concerns often shift to IT infrastructure security requirements, such as access control and other measures to maintain AI system integrity and to keep its data safe.

Trust at the end-applications layer of an AI system requires, among other things, software that is reliable and safe. In the context of critical systems, the production of software is framed by a set of processes to verify and validate the "product"[28]. The same is true for AI systems and more. With the stochastic nature of AI systems based on machine learning, trustworthiness also implies fairness of the system's behaviour, corresponding to the absence of inappropriate bias.

Moreover, social trust is based on a person's way of life, belief, character, etc. Without a clear understanding of the internal functioning, its operating principles are not transparent to the non-technical segment of population. In this case, the establishment of trust might not be dependent on objective verification of the AI system's performance, but rather based on a subjective pedagogical explanation of the AI system's observed behaviour.

5.3 Application of software and data quality standards

Software has an important effect on the trustworthiness of a typical AI system. As a result, identifying and describing the quality attributes of its software can help to improve trustworthiness of the whole system[29]. These attributes can contribute to both cyber and social trust. For example, from a societal

