Technical Specification

ISO/IEC TS 8200

Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems

Technologies de l'information — Intelligence artificielle — Contrôlabilité des systèmes d'intelligence artificiels automatisés

First edition
2024-04

Reference number
ISO/IEC TS 8200:2024(en)

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) 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 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*.

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.

Introduction

Artificial intelligence (AI) techniques have been applied in domains and markets such as health care, education, clean energy and sustainable living. Despite being used to enable systems to perform automated predictions, recommendations or decisions, AI systems have raised a wide range of concerns. Some characteristics of AI systems can introduce uncertainty in predictability of AI system behaviour. This can bring risks to users and other persons. For this reason, controllability of AI systems is very important. This document is primarily intended as a guidance for AI system design and use in terms of controllability realization and enhancement.

Controllability characteristics (see Clause 6) and principles of AI systems are identified in this document. This document describes the needs of controllability in a domain-specific context and strengthens the understanding of an AI system's controllability. Controllability is an important fundamental characteristic supporting AI systems' safety for users and other persons.

Automated systems as described in ISO/IEC 22989:2022, Table 1 can potentially use AI. The degree of external control or controllability is an important characteristic of automated systems. Heteronomous systems range over a spectrum from no external control to direct control. The degree of external control or controllability can be used to guide or manipulate systems at various levels of automation. This can be satisfied by the use of controllability features (see Clause 7) or by taking specific preventive actions within each stage of the AI system life cycle as defined in ISO/IEC 22989:2022, Clause 6. This document refers to the controllability by a controller, that is a human or another external agent. It describes controllability features (what and how), but does not presuppose who or what is in charge of the controlling.

Unwanted consequences are possible if an AI system is permitted to make decisions or take actions without any external intervention, control or oversight. To realize controllability (see Clause 8), key points of system state observation and state transition are identified. The exact points where transfer of control is enabled can be considered during the design and implementation of an AI system.

Ideally, the transfer of control for an intervention occurs within reasonable time, space, energy and complexity limits, with minimal interruption to the AI system and the external agent. Stakeholders can consider the cost of control transfer (see 6.9) of automated AI systems. Uncertainty during control transfer can exist on the AI system and the external agent sides. Thus, it is important to carefully design the control transfer processes to remove, minimize, or mitigate uncertainty (see 6.8) and other undesired consequences.

The effectiveness of control can be tested. Such testing takes into account the design and development of the control transfer. This calls for principles and approaches for validation and verification of AI systems' controllability (see Clause 9).

Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems

1 Scope

This document specifies a basic framework with principles, characteristics and approaches for the realization and enhancement for automated artificial intelligence (AI) systems' controllability.

The following areas are covered:

- state observability and state transition;
- control transfer process and cost;
- reaction to uncertainty during control transfer;
- verification and validation approaches.

This document is applicable to all types of organizations (e.g. commercial enterprises, government agencies, not-for-profit organizations) developing and using AI systems during their whole life cycle.

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 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989, ISO/IEC 23053 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
ontology
conceptualisation of a domain

[SOURCE: ISO/IEC 5392:2024, 3.9]

3.2
knowledge representation
process that designs and constructs symbolic *systems* (3.9), rules, frameworks, or other methodologies used to express knowledge which machines can recognize and process

[SOURCE: ISO/IEC 5392:2024, 3.18]

3.3
knowledge computing
process that obtains new knowledge based on existing knowledge and their relationships

[SOURCE: ISO/IEC 5392:2024, 3.28]

3.4
knowledge fusion
process that merges, combines and integrates knowledge from different resources into a coherent form

[SOURCE: ISO/IEC 5392:2024, 3.21]

3.5
control, verb
<controllability>in engineering, the monitoring of system output to compare with expected output and taking corrective action when the actual output does not match the expected output

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.846.1]

3.6
controller
authorized human or another external agent that performs a control

Note 1 to entry: A controller interacts with the control points of an AI system.

3.7
disengagement of control
control disengagement
process where a *controller* (3.6) releases a set of *control points* (3.16)

3.8
engagement of control
control engagement
process where a *controller* (3.6) takes over a set of *control points* (3.16)

Note 1 to entry: Besides taking over a set of control points, an engagement of control can also include a confirmation about the transfer of control to a controller.

3.9
system
arrangement of parts or elements that together exhibit a stated behaviour or meaning that the individual constituents do not

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

Note 2 to entry: In practice, the interpretation of its meaning is frequently clarified by the use of an associative noun (e.g. aircraft system). Alternatively, the word “system” is substituted simply by a context-dependent synonym (e.g. aircraft), though this potentially obscures a system's principles perspective.

Note 3 to entry: A complete system includes all of the associated equipment, facilities, material, computer programs, firmware, technical documentation, services, and personnel required for operations and support to the degree necessary for self-sufficient use in its intended environment.

[SOURCE: ISO/IEC/IEEE 15288:2023, 3.47]

3.10
system state
state
one of several stages or phases of system operation

Note 1 to entry: A system state is represented by related internal parameters and observable characteristics.

[SOURCE: ISO 21717:2018, 3.3, modified states as state]

3.11
system state stability
stable system state
degree to which a system's parameters and observable characteristics remain invariable during a specified period of time or another dimension such as space

Note 1 to entry: Invariableness can be defined by means of a variableness tolerance based on business requirements.

Note 2 to entry: When leaving a stable system state, the system's parameters or observable characteristics change, regardless of whether the next stable state is safe or unsafe, when the *system* (3.9) enters an unstable system state.

Note 3 to entry: A *system* (3.9) can be described as stable, if the system is in a stable state.

3.12
safe state
state (3.10) that does not have or lead to unwanted consequences or loss of control

3.13
unsafe state
state (3.10) that is not a *safe state* (3.12)

Note 1 to entry: Uncertain states are a subset of unsafe states.

3.14
failure
loss of ability to perform as required

[SOURCE: IEC 60050-192:2015, 192-03-01, modified — notes to entry have been deleted.]

3.15
success
simultaneous achievement by all characteristics of required performance

[SOURCE: ISO 26871:2020, 3.1.62]

3.16
control point
part of the interface of a *system* (3.9) where controls can be applied

Note 1 to entry: A control point can be a function, physical facility (such as a switch) or a signal receiving subsystem.

3.17
span of control
subset of control points, upon which controls for a specific purpose can be applied

3.18
interface
means of interaction with a component or module

3.19
transfer of control
control transfer
process of the change of the *controller* (3.6) that performs a control over a *system* (3.9)

Note 1 to entry: Transfer of control does not entail application of a control, but it is a handover of control points of the system interface between agents.

Note 2 to entry: Engagement of control and disengagement of control are two fundamental complementary parts of control transfer.

3.20
finite state machine
FSM
computational model consisting of a finite number of *states* (3.10) and transitions between those states, possibly with accompanying actions

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.1604]

3.21
system state transition
transition
process in that a *system* (3.9) changes from one *state* (3.10) to another state or to the same state

Note 1 to entry: A transition takes place when a condition is satisfied, including an intervention from a controller.

[SOURCE: ISO/IEC 11411:1995, 2.2]

3.22
cost of control
resources spent and associated external effects by performing control over an AI system

Note 1 to entry: Resources include time, space, energy, material and any other consumable items.

Note 2 to entry: External effects include all possible effects and side effects of control, e.g. environment change.

3.23
test completion report
test summary report
report that provides a summary of the testing that was performed

[SOURCE: ISO/IEC/IEEE 29119-1:2022, 3.87]

3.24
process
set of interrelated or interacting activities that transform inputs into outputs

[SOURCE: ISO/IEC/IEEE 15288:2023, 3.27]

3.25
function
defined objective or characteristic action of a *system* (3.9) or component

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.1677.1]

3.26
functionality
capabilities of the various computational, user interface, input, output, data management, and other features provided by a product

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.1716.1, modified — Note 1 to entry has been removed.]

3.27
functional safety
part of the overall safety relating to the EUC (Equipment Under Control) and the EUC control system that depends on the correct functioning of the E/E/PE (Electrical/Electronic/Programmable Electronic) safety-related *systems* (3.9) and other risk reduction measures

[SOURCE: IEC 61508-4:2010, 3.1.12]

3.28
system state observation
observation
act of measuring or otherwise determining the value of a property or *system* (3.9) state

3.29
transaction
set of related operations characterized by four properties: atomicity, consistency, isolation and durability

[SOURCE: ISO/IEC TR 10032:2003, 2.65, modified — Note 1 to entry has been removed.]

3.30
atomic operation
operation that is guaranteed to be either performed or not performed

3.31
out of control state
unsafe state (3.13) in which the *system* (3.9) cannot listen for or execute feasible control instructions

Note 1 to entry: The reasons for out of control state include but are not limited to communication interruption, system defection, resource limitation and security.

4 Abbreviations

AI	artificial intelligence
ML	machine learning

5 Overview

5.1 Concept of controllability of an AI system

Controllability is the property of an AI system which allows a controller to intervene in the functioning of the AI system. The concept of controllability is relevant to the following areas for which International Standards provide terminology, concepts and approaches for AI systems:

a) AI concepts and terminology: This document inherits the definition of controllability from ISO/IEC 22989;

b) AI system trustworthiness: ISO/IEC TR 24028 describes controllability as a property of an AI system that is helpful to establish trust. Controllability as described by ISO/IEC TR 24028 can be achieved by providing mechanisms by which an operator can take over control from the AI system. ISO/IEC TR 24028 does not provide a definition for controllability. Controllability in this document is used in the same sense as in ISO/IEC TR 24028. A controller in the context of this document can be a human. This is the same with the philosophy in ISO/IEC TR 24028. When an AI system is in its operation and monitoring stage, a human can be in the loop of control, deciding control logics and providing feedback to the system for further action;

c) AI system quality model: ISO/IEC 25059 describes user controllability as a sub-characteristic of usability. ISO/IEC 25059 emphasizes the interface of an AI system, which enables the control by a controller, while the controllability defined in this document is more about the functionalities that allow for control;

d) AI system functional safety: ISO/IEC TR 5469 uses the term control with two different meanings:

   1) Control risk: This meaning refers to an iterative process of risk assessment and risk reduction. The term control belongs to the context of management. This meaning differs from the use of control in this document;

   2) Control equipment: This meaning refers to the control of equipment as well as the needs of control by equipment that has a certain level of automation. This meaning of control in ISO/IEC TR 5469 is consistent to the use of control in this document;

e) AI risk management: ISO/IEC 23894[12] uses the term control in the context of organization management, meaning the ability of an organization to influence or restrict certain activities identified to be risk sources. This meaning is different from the meaning of control or controllability in this document;

f) AI system using machine learning: The meaning of control in this document is the same as the meanings in ISO/IEC 23053, where reinforcement learning is described as an approach to realize control purpose. In the context of this document, an external agent can make use of reinforcement learning to realize control logic.

Based on the definition of controllability in ISO/IEC 22989:2022, 3.5.6, an AI system does not control itself but is controlled by an external agent. In this document, an AI system that has realized controllability functionalities is regarded as a system of systems. It is composed of a system realizing AI and a system realizing controllability. The latter is defined as an external agent in ISO/IEC 22989. This concept is applied in this document.

Controllability can be important for AI systems whose underlying implementation techniques cannot provide full explainability or verifiable behaviours. Controllability can enhance the system's trustworthiness, including its reliability and functional safety.

No matter the automation level of an AI system, controllability of an AI system is important, so an external agent can ensure that the system behaves as expected and to prevent unwanted outcomes.

The design and implementation of controllability of an AI system can be considered and performed in each stage of the AI system life cycle defined in ISO/IEC 22989:2022, Clause 6.

Controllability is a technical prerequisite of human oversight of an AI system, so that the human-machine interface can be technically feasible and enabled. The design and implementation of controllability should be considered and practiced by stakeholders of an AI system that can impact users, the environment and societies.

Controllability of an AI system can be achieved if the following two conditions are met:

- The system can represent its system states (e.g. internal parameters or observable characteristics) to a controller such that the controller can control the system.
- The system can accept and execute the control instructions from a controller, which causes system state transitions.

5.2 System state

In a system, interacting elements can exchange data and cooperate with each other. These interactions can lead to different sets of values for the system's internal parameters and consequently can result in different observable characteristics.

A system can have several different states. The different states of a system can indicate a mapping from the continuous parameter space to a discrete state space. When designing the different states of a system, at least the following recommendations apply:

- All states are meaningful to the system's business logic.
- The duration of a state is sufficient so that tests and specific operations against the state can be made.
- A state is observable by qualified stakeholders, via technical means, such as system logging, debugging and breakpoints, etc.
- Entry into a state is possible via a set of defined operations on the system.

The states of an AI system can be identified during the design and development stage in the AI system life cycle as described in ISO/IEC 22989:2022, Figure 3. The identification of the states of an AI system is important for the implementation of controllability and can therefore affect the trustworthiness of the AI system. According to the results of the design and development stages, the states of an AI system can be organized into the following three categories:

- safe and unsafe;
- operating or failing to operate as specified;