ISO/IEC DTS 8200:2023

ISO/IEC JTC 1/SC 42/WG 3

Secretariat: ANSI

Date: 2023-10-1012-11

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

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

FDIS stage

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

A model manuscript of a draft International Standard (known as "The Rice Model") is available at https://www.iso.org/iso/model_-rice_model.pdf

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11
EmailE-mail: copyright@iso.org
Website: www.iso.orgwww.iso.org

Published in Switzerland

Contents

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

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 hazards. 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 system 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.

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, i.e. a human or another external agent. –It describes controllability features (what and how), but does not predetermine who or what is in charge of the controlling.

Unwanted consequences are possible if an AI system is permitted to take decisions or actions without any external intervention, control or oversight. To realize controllability (see ~~clause 8),~~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 both 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).~~Clause 9).

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

1 Scope

This document ~~defines~~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~~:2022~~, *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~~:2022~~, ISO/IEC 23053~~: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
ontology
~~a~~ conceptualisation of a domain

[SOURCE: ISO/IEC 5392:—[1], 3.9]

[1] Under preparation. Stage at the time of publication: ISO/IEC FDIS 5392:2023.

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

[SOURCE: ISO/IEC 5392~~:—[1],~~:—, 3.18]

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

[SOURCE: ISO/IEC 5392~~:—[1],~~:—, 3.~~2328~~]

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

[SOURCE: ISO/IEC 5392~~:—[1],~~:—, 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)*~~(3.6) releases a set of *control points* ~~*(3.16)*~~(3.16)

3.8
engagement of control

control engagement
process where a *controller* ~~*(3.6)*~~(3.6) takes over a set of *control points* ~~*(3.16)*~~(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 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)*~~(3.9) enters an unstable system state.

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

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

3.13
unsafe state
state ~~*(3.10)*~~(3.10) that is not a *safe state* ~~*(3.12)*~~(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)~~(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 effects to the external 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.