Technical Specification

ISO/IEC TS 8200

First edition
2024-04

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

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

Reference number
ISO/IEC TS 8200:2024(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 ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Abbreviations **5**
5 Overview **5**
5.1 Concept of controllability of an AI system 5
5.2 System state 6
5.3 System state transition 7
5.3.1 Target of system state transition 7
5.3.2 Criteria of system state transition 7
5.3.3 Process of system state transition 7
5.3.4 Effects 8
5.3.5 Side effects 8
5.4 Closed-loop and open-loop systems 8
6 Characteristics of AI system controllability **9**
6.1 Control over an AI system 9
6.2 Process of control 11
6.3 Control points 12
6.4 Span of control 13
6.5 Transfer of control 13
6.6 Engagement of control 15
6.7 Disengagement of control 16
6.8 Uncertainty during control transfer 17
6.9 Cost of control 17
6.9.1 Consequences of control 17
6.9.2 Cost estimation for a control 18
6.10 Cost of control transfer 18
6.10.1 Consequences of control transfer 18
6.10.2 Cost estimation for a control transfer 18
6.11 Collaborative control 18
7 Controllability of AI system **19**
7.1 Considerations 19
7.2 Requirements on controllability of AI systems 20
7.2.1 General requirements 20
7.2.2 Requirements on controllability of continuous learning systems 21
7.3 Controllability levels of AI systems 21
8 Design and implementation of controllability of AI systems **22**
8.1 Principles 22
8.2 Inception stage 23
8.3 Design stage 24
8.3.1 General 24
8.3.2 Approach aspect 24
8.3.3 Architecture aspect 25
8.3.4 Training data aspect 25
8.3.5 Risk management aspect 25
8.3.6 Safety-critical AI system design considerations 25
8.4 Suggestions for the development stage 25
9 Verification and validation of AI system controllability **26**
9.1 Verification 26

© ISO/IEC 2024 – All rights reserved

9.1.1 Verification process 26
9.1.2 Output of verification 26
9.1.3 Functional testing for controllability 26
9.1.4 Non-functional testing for controllability 27
9.2 Validation 28
9.2.1 Validation process 28
9.2.2 Output of validation 28
9.2.3 Retrospective validation 28
Annex A (informative) **Example verification output documentation** **30**
Annex B (informative) **Example validation output documentation** **32**
Bibliography **34**

© 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 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.

© ISO/IEC 2024 – All rights reserved

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).

© ISO/IEC 2024 – All rights reserved

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]

© ISO/IEC 2024 – All rights reserved