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 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;

— any other kind of categorizations meaningful to system operation, test and maintenance.

A system can be in a safe state but fail to operate as specified. A system can also be in an unsafe state yet still operate as specified. Successful operation does not always correspond to safe states and failure to operate does not always correspond to unsafe states. Success and failure depend on system design and development for handling internal transitions within and between safe and unsafe states.

EXAMPLE In a financial service, an AI system is used to evaluate credit applications. The approval operation against a loan is blocked by the AI system component and consequently failed because of an erroneously predicted credit repayment risk. Such failure of the AI system to operate as specified does not mean that the system entered into an unsafe state.

5.3 System state transition

5.3.1 Target of system state transition

The system state transition target is a finite subset of the system's possible states which are acceptable by stakeholders according to a set of requirements. The system state transition target should be identified during design and development and the transitions to a target state should be subject to verification and validation during system testing.

The implementation and enhancement of controllability of an AI system depends on the ability of an AI system to reach a specified target state. The following attributes of the intended target state should be identified by the designers, developers, managers, users and any other stakeholders of the AI system:

— Completeness of the states of an AI system can be checked. States that are not noticed or hard to be entirely identified can exist during the design and development stage. This is particularly the case when an AI system is implemented by certain approaches, such as deep learning. As a deep learning model's output universe cannot be entirely determined in advance, unidentified states can always exist;

— Stability of the states of an AI system should meet the requirements about control and state observation. This attribute is important for the systems which are designed to be controlled by human, as human-in-the-loop mechanisms are applied to prevent hazards.

5.3.2 Criteria of system state transition

Target states should be reachable under certain circumstances via actions. Actions can include:

— external control via system-defined operations;

— automated state transition by the system itself, if defined conditions are met;

— forced state transition by an external event.

Two types of conditions can be considered for cause system state transition:

— a sufficient condition that by itself causes the transition to take place as long as the condition is met;

— a necessary condition that is required to be met for the state transition to take place.

The satisfaction of a necessary condition does not by itself guarantee the transition happens.

5.3.3 Process of system state transition

Once triggered, a system state transition can occur. AI systems' state transition processes can differ. Common subprocesses of a state transition can be identified. A system state transition process contains up to two subprocesses:

a) Launching: After the condition of a trigger is met, a system can have a set of internal operations launched according to configurations or implementations of business logic. Such operations can include invocation of functions, adjustments on system parameters, resources allocation or deallocation, and

other actions that the system can take internally in order to reach its defined next state. A launching subprocess can be brief and difficult to capture or even record, depending on a system's state definitions and implementations.

EXAMPLE 1 A deep learning training process completes and the change of model parameters in memory stops. Based on the training configuration, the persistence of that model can be triggered. Correspondingly, the system's state transits from model training to model persisting. For this, necessary functions (e.g. write data to disk) are invoked and resources (space on disk) are allocated.

b) Adaptation: An AI system state transition can change the environment where that system works in or objects on which it operates. As a consequence, such environments and objects can react to the AI system. These reactions can lead to an unstable adaptation period in which the system adjusts internal parameters to enter an intended state. An adaptation subprocess is not a necessity that every system state transition process contains.

EXAMPLE 2 An AI-based vehicle system automatically transits its state from low speed to high speed. As the vehicle speeds up, the resistance (from ground, air, etc.) and running stability can change. To cope with this, parameters in subsystems (such as electronic stability program) can be adjusted. Once the target state (high speed) is reached, the adjustment approaches applied in the adaptation subprocess can be stopped.

5.3.4 Effects

The effects of an AI system state transition can include the current state of the system or an additional set of actions needed to be taken by the system or its controller. There can be two types of effects:

a) For successful state transition: When a system successfully transits to the expected state, the system can function as specified and is prevented from entering a hazardous state.

b) For unsuccessful state transition: When a system fails to transit to the expected state, it can be guided to revert to the original state by configured operations or parameters (e.g. system reset). The system can then retry the requested state transition or stay in the original state. For this, extra time, operations, power and other resources can be necessary.

5.3.5 Side effects

Side effects can emerge following a system state transition and can lead to the changes to the environment where a system is operating or to the objects on which the system operates. Not all changes to a system's environment or objects operated on by a system can be recovered to their original state. The inability to reverse side effects should be carefully considered when using AI systems in domains such as material processing and manufacturing.

5.4 Closed-loop and open-loop systems

In a closed-loop system, the output is fed back to the input of the system, where control is determined by the combination of system input and feedback (e.g. the control to an air conditioner is subject to both the current and target temperatures). In an open-loop system, the output is not fed back to the input of the system. Control is subject to the instructions issued by a controller rather than the output of the system (e.g. a TV only accepts and responds to a control signal rather to the results of previous controls).

System state observation measures the appropriate system parameter values or system appearance. It can be achieved via either system outputs or observation or based on analysis of system parameters without output. This document does not treat closed-loop and open-loop separately and does not impose specific settings to the approaches via which the system states are observed. It also does not impose specific settings to the approaches via which the system states are observed.

6 Characteristics of AI system controllability

6.1 Control over an AI system

Control over an AI system can help to conduct the intended business logic and to prevent the system from causing harm to stakeholders. At least the following two ways exist to realize the controllability of an AI system:

— Use the facilities designed and implemented for the purpose of control;

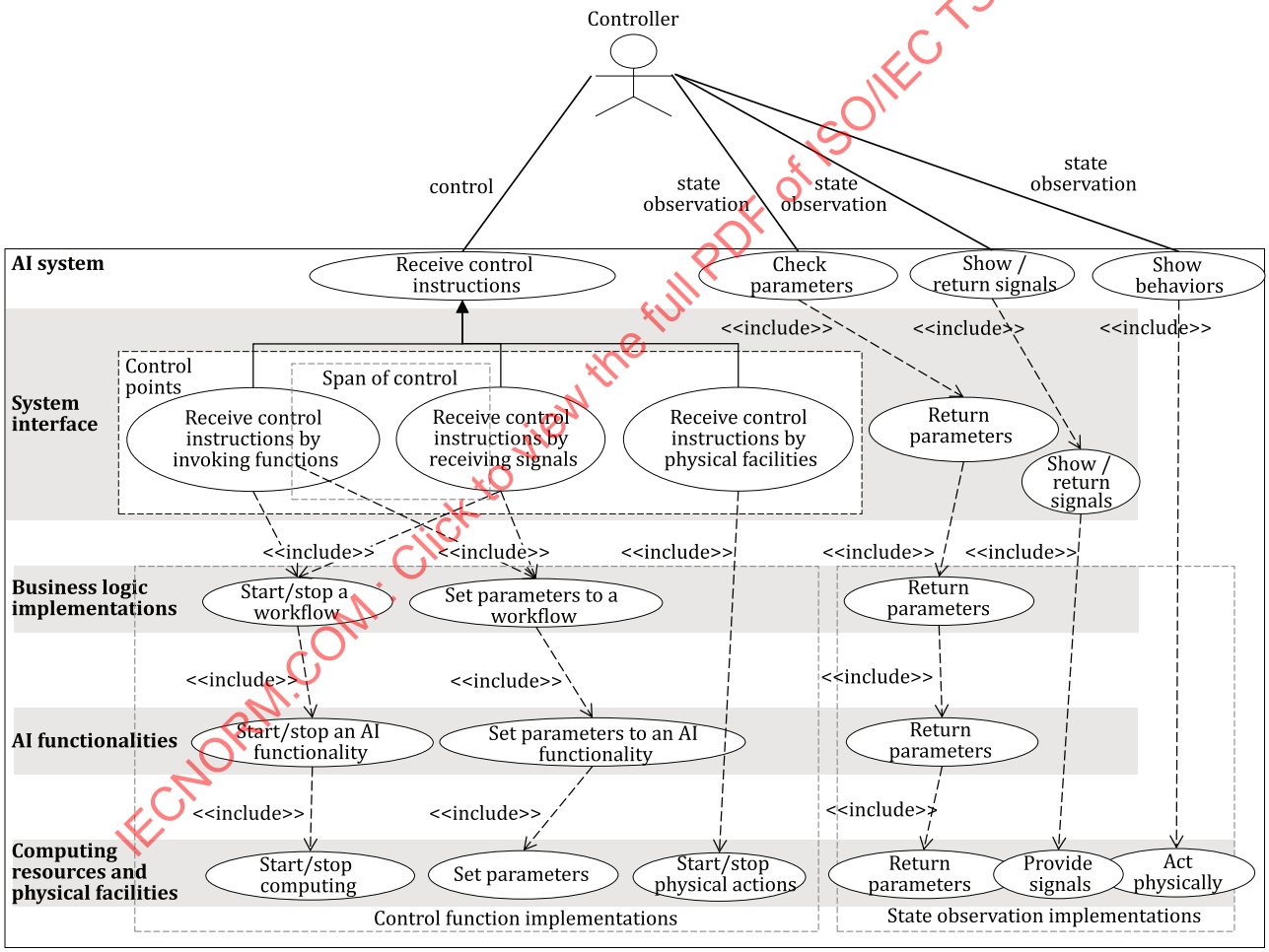
— Take advantage of the functional operations (they are not specifically designed and implemented for control but can be used for the purpose of control).

Control over an AI system is effective if at least the following are satisfied:

— Control is conducted when the system can be controlled for a specific purpose with acceptable side effects.

— Control is conducted via a correct span of control based on control points provided by the system.

— Control works as intended.



NOTE 1 The span of control represented in the diagram is an example. Each specific control can correspond to its intended span of control that is configured, selected and used.

NOTE 2 See ISO/IEC 19505-1[2] for details on the notations in this diagram. The human body notation in Figure 1 does not mean that it is necessarily a human.

Figure 1 — Use case diagram with examples of an AI system under control

Figure 1 shows a use case diagram as an example of an AI system under control:

a) For a specific purpose of control, a controller can observe system states and issue control instructions to an AI system via a span of control provided by that system. Observations on system states can be done by the following:

   1) A controller invokes functionalities that return parameter values on demand.

   2) A controller receives signals sent by the system which contain the information about the current system state.

   3) A controller observes the physical behaviours of the system.

b) An AI system is designed and implemented with interfaces facilitating control and state observation. An AI system can have multiple components. Each of the components can provide facilities for control (see 6.1) and state observation:

   1) Computing resources can include computing device, memory, storage, data transmission facility and any other hardware module that improve computing and data exchange. Status and parameters of computing resources can be set and observed for the purpose of control. Physical facilities can include hardware and associated software used for the formation or functioning of the AI system (e.g. joysticks or gear shafts). Devices in a component can provide control and state observation.

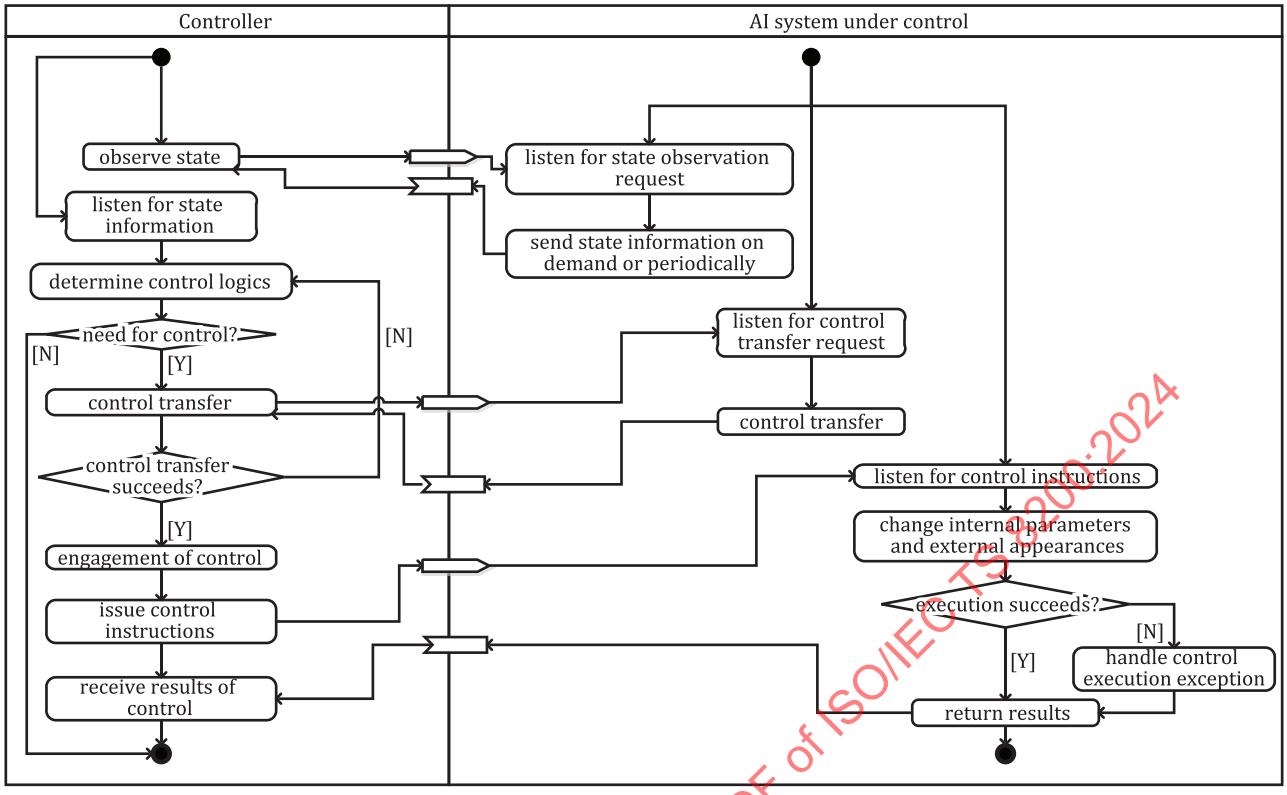
   2) AI functionalities abstract those processes used for prediction, recommendation and classification. Parameters and status of an AI functionality can be set and observed for the purpose of control.

   3) Business logic implementations are the executable programs that form workflows. Each workflow can invoke AI functionalities as building blocks. Implementations in this layer can include control facilities that make sense to business logic.

   4) System interface can contain a subset of declared functionalities for receiving control instructions, providing parameter values, returning signals and showing observable characteristics. This subset is the control point of the system. For a specific control, a span of control can be configured, selected and used.

c) Dependencies can exist between control functionalities provided by different layers.

6.2 Process of control



NOTE See ISO/IEC 19505-1[2] for details on the notations in this diagram.

Figure 2 — Control process activity diagram

The process of control can involve both the controller and the AI system under control. A general process is shown in an activity diagram in Figure 2, including the following subprocesses:

a) A controller observes the current state of an AI system under control. This is done by interacting with the interface of the AI system. For this, the controller listens for state information that is supplied by the AI system.

b) An AI system under control can listen for the following types of requests:

   1) State observation. When a state observation request is received, the AI system returns information about the current state to the controller. Such information can also be periodically reported to the controller.

   2) Control transfer. When a transfer of control request is received asking for a control transfer, the AI system can hand over control to an authorized controller.

   3) Control instructions. When a request is received containing control instructions, the AI system executes the instructions on demand.

c) When receiving state information, the controller determines the logic of control. This can start one of the following options:

   1) If it is determined that the AI system does not need to be controlled, the control process ends.

   2) If it is determined that the AI system needs to be controlled, a subprocess that prepares control transfer starts.

d) If the controller is not able to perform the intended control due to a lack of span of control, the controller requests a control transfer. This can happen when the needed control points have not been fully handed

over. If the controller holds all needed control points for this control, the request subprocess is skipped; otherwise, control is exchanged from the AI system to the controller. A control transfer can also require authentication and authorization checks.

e) The controller issues control instructions. Once the instructions are received, the AI system changes its internal parameters or observable characteristics. This can lead to two kinds of results:

   1) If the control instructions are executed successfully, the AI system returns the results to the controller.

   2) If the control instructions are executed unsuccessfully, the AI system handles the possible exceptions and returns the results to the controller.

AI systems with a finite set of system states can be modelled based on an FSM. Applying control methods based on an FSM is possible when the representation of the control transfer between different controllers is through the transfer function Σ which is defined by a 3-tuple:

$$\Sigma(S, A, \gamma)$$

where

S is a finite set of system states (see 5.2);

A is a set of actions (see 5.3);

γ is a set of transitions of system states (see 5.3).

6.3 Control points

A control point of an AI system can include but is not limited to the following:

— A function. When a system is controlled programmatically, functions implementing control logics should be designed. For this, local invocations or remote procedure calls can be considered.

— A physical facility. When a system is equipped with physical mechanisms for control, such as a steering wheel on an assisted-driving vehicle, safety and usability factors that can affect the effectiveness and efficiency of control should be considered.

— A signal input-output system. When a system is controlled remotely, a signal input-output subsystem can be applied. In addition to considering the medium (e.g. air or water), distance and noise, the subsystem should also consider expectations for control timeliness and sequencing.

Depending on the design, control points of a system can make use of the following:

— specifically designed and implemented facilities that are exclusively used for control;

— facilities that are parts of a system's functions but can be re-used for control, such as the checkpoint and the pause functions designed for debugging but useful for control in certain cases.

When necessary, the invocations of control points can be secured by authentication and authorization mechanisms. For this, certification, encryption mechanisms and even control-specific channels can be applied.

EXAMPLE An AI-based automated metal processing product line can be controlled via a digital control subsystem as well as a set of physical facilities on the production line. An AI system is used for the analysis of photographs of the key information of the processed metal (e.g. the position and the pose of a part being processed). The controls can include starting, stopping and pausing of subprocesses, selecting and changing of chucks, heating, cooling, lathing and milling of materials, changing of bit tools, etc. The controls of the system can be configured in advance and issued in real time via the digital control subsystem. Physical facilities can also be used if manual and physical controls are needed. To use the digital control subsystem and to enter the physical control area can require that the identification information of human controllers be checked.

6.4 Span of control

Span of control is a subset of control points upon which a specific control can be applied. A controller's processing a span of control reflects the condition that the system is ready to listen for and execute the instructions issued by the controller for the specific control. Therefore, before an actual control is performed the following should be confirmed:

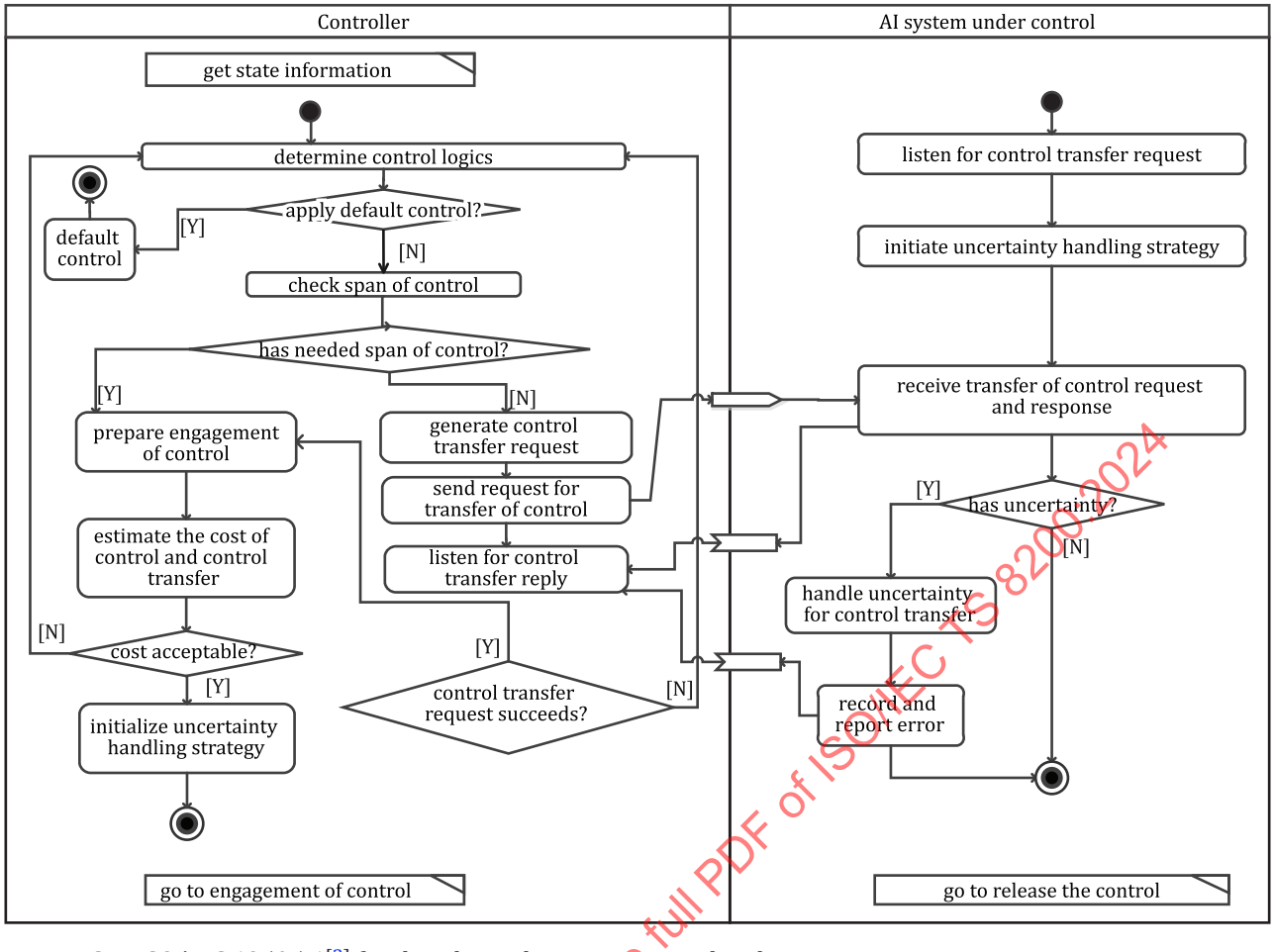
— that the system can accept and conduct the control instructions from a specific controller. If not, the controller cannot fully perform the intended control. An incomplete span of control can lead to a control transfer from the system to the controller;

— that the controller can handle or operate all the control points of an intended control. If not, either an uncertainty handling mechanism should be prepared or the plan for this control should be cancelled due to the lack of feasibility.

When interacting with the control points in a span of control, rules can exist about the sequence for using the control points.

6.5 Transfer of control

The transfer of control is a prerequisite when an external controller decides to intervene in the functioning of an AI system in order to prevent unwanted outcomes. A control transfer process enables the controller to obtain the control from any agent that is controlling the AI system. For this, a preparation process for control transfer should be considered. Important subprocesses during a preparation include checking the span of control, preparing for engagement of control, initializing uncertainty handling strategy as well as estimating the cost of control and control transfer. The preparation process for the transfer is shown in Figure 3 and described as follows:



NOTE See ISO/IEC 19505-1[2] for details on the notations in this diagram.

Figure 3 — Transfer of control from an AI system to a controller

a) A control transfer preparation process is conducted based on the requirements of the control. It includes a sequence of subprocesses:

   1) The controller checks the span of control that is necessary for the intended control.

   2) If the controller does not hold all the control points for the required span of control, the controller generates an additional control transfer request before its engagement of the control. The additional request declares the controller's intent about the upcoming operation on a subset of control points and is sent to the AI system. Upon receiving this request, the AI system replies to the controller with a confirmation and the AI system is starting to prepare for its disengagement of control. The AI system disengages its control only if the controller holds the authority for the requested control.

   3) When the controller already holds all the needed control points of a span, the actions in 2) are skipped.

   4) A request of control transfer can fail, if uncertainties (see 6.8) appear during the communication between the controller and the AI system. A failed request can trigger a redetermination of requirements of the control which can cause the controller to adopt a different strategy for control.

   5) If a control transfer request is successful then the preparation for the engagement of the control is carried out. The controller derives a plan containing a sequence of actions (e.g. move to correct position for control) that should be taken in order to be ready for the actual operation.

   6) The cost of control as well as the possible control transfer are estimated. During this subprocess, the controller gathers estimates of time, space, energy and material consumptions, as well as the

effects to the AI system and concerned environments. When the estimated cost exceeds a certain limit, the requirements of the control can be adjusted. If no suitable control exists, then a default control strategy can be performed.

7) The controller can also initialize an uncertainty handling strategy to deal with any unpredicted failures during control and control transfer. A similar strategy can also be initialized by the AI system.

b) A control transfer process and its preparation can be ignored if it is determined that the AI system does not need to be controlled, according to the observed state.

When a control transfer involves multiple control points during an engagement or disengagement process, partial success can happen if failed operations occur during an entire procedure. This can lead an AI system to an abnormal state, where unnecessary occupation of resources or unwanted changes of environments can occur. To handle this, a transaction mechanism should be considered for control transfer which guarantees either the control transfer is entirely successfully performed or entirely not performed if partial failure happens. In a failed or a partially failed case, all operations in a transaction are expected to be rolled back.

An AI system can physically change the world (e.g. an automated driving system drives a car over a distance even during a control transfer to a human driver). Perfect rollback of a transaction can be difficult, since physical changes (e.g. material processed, distance travelled, power consumed, temperature raised) cannot be recovered easily without external changes or extra expenses. In this situation, mechanisms such as two- or three-phase commitment can be considered with availability criteria on controllability messaging channel, controller and AI system.

6.6 Engagement of control

An important prerequisite for the controller to control an AI system is to engage a specific control. Engagement of control means to carry out a sequence of the actions to control the AI system. In addition, a set of criteria should be met when performing a specific action or a sequence of actions. Useful actions include but are not limited to:

— move to a required position;

— wear or setup a required equipment;

— handle a required physical operating equipment;

— launch a required control toolkit (software).

The following criteria can also be selected and used according to control requirements:

— time duration limitation on the completion of an engagement of control;

— physical space limitation allowed by an AI system for the engagement of a control;

— order or precedence restriction on the engagement of control when multiple control points are involved;

— authority restriction on the engagement of control when security requirements exist on the obtainment of control points;

— completeness of engagement over the entire span of the intended control.

To control the AI system, the engagement of control process should be prepared in advance to plan a sequence of actions and satisfy the corresponding criteria. The engagement of control should be configured or planned in advance for each possible control in order to decrease uncertainty and infeasibility. The preparation for control engagement can happen later during the preparation of control transfer so that an accurate cost can be estimated.

When an engagement of control process is prepared, the controller is able to take the planned actions and confirms with the AI system about the needed span of control. A confirmation is a handshake between the controller and the AI system. The controller declares the use of the required control points, while the AI system releases those control points and then listens for and executes the instructions from the controller.

A control transfer can involve multiple control points. If an error happens on a part of a span during a control engagement process, the overall transfer of control can fail or the further control can be infeasible. To avoid this situation, it is useful to consider and implement a transactional mechanism for engagement of control. Keys for such a transactional mechanism can include but are not limited to the following:

— Set up a recovery point during the "initialize uncertainty handling strategy" activity in Figure 3, such that failures during the engagement of control process can be handled and the system configuration and data being processed can be recovered (see "handle control execution exception" activity in Figure 1). In an ML-based AI system, a recovery point can be a set of data including a checkpoint mirror of the ML model, runtime configurations, etc.

— Arrange a plan to handle the possible damage to the environment. This is important for those AI systems working with physical objects and for when their control or control transfer can influence environments (e.g. an ML-based material processing system).

— Implement a mechanism to ensure the atomicity of a control engagement, such that the control engagement process is guaranteed to either successfully occur or entirely not occur. As a result, no engagement of control over a partial span can occur.

6.7 Disengagement of control

The disengagement of control is a process opposite to the engagement of control. Disengagement of control means the AI system is about to release and transfer control to the controller. The core task of this process is to take a sequence of actions and then satisfy a set of criteria. Useful actions include but are not limited to the following:

— leave a position;

— take off or set down an equipment;

— release a physical operating equipment;

— terminate or pause a control toolkit (software).

The criteria of the engagement of control process can be selected and used in the context of control disengagement, but with a different meaning for each:

— time duration limitation on the completion of a disengagement of control;

— physical space limitation allowed by an AI system for a disengagement of a control;

— order restriction on the disengagement of control when multiple control points are relinquished;

— authority restriction on the disengagement of control when security requirements exist on the relinquishment of control points;

— completeness of disengagement over the entire span of the intended control.

To handover the control of the AI system, a disengagement of control process should be prepared in advance. This disengagement process is commonly triggered by a controller that is preparing its disengagement of control. Preparation should include the generation of a plan for possible detachment. The plans for disengagement can be configured in advance in order to decrease uncertainty and infeasibility. The preparation for control disengagement can also happen within the actual preparation of control transfer

When a control disengagement process is prepared, the AI system is able to take the planned actions and confirm with the controller about the referenced span of control. A confirmation handshake between the controller and AI system happens. The AI system releases the control points in the span of control and starts to listen for other instructions from the controller.

Transactional mechanism should be also prepared for a disengagement of control process where partial failure can happen during the relinquishment of multiple control points. For this, similar keys in 6.6 should be considered.

6.8 Uncertainty during control transfer

For a transfer of control the capacity of the controller should be considered. Capacity refers to whether or not the controller is able to manage the control transfer. It is a relative concept, which can also depend on the complexities of the control transfer. Factors that can affect the capacity include:

— number of control points that are supposed to be handed over;

— positions of control points if physical controls are to be performed;

— time duration restrictions required by the transfer of control;

— controller's resources (e.g. idle time intervals) that can be used for the control transfer.

A transfer of control can fail if either the controller or the AI system does not prepare well or is affected by unpredicted external factors. Uncertainty should be handled when a failure happens, and particularly in the cases that can lead to loss of asset, performance or any other results and risks unacceptable to both the controller and the AI system. Types of uncertainty include but are not limited to:

— communication failure;

— control handover confirmation failure.

In any case, a default uncertainty handling mechanism should be considered and implemented to stop or pause the current action of an AI system. A more advanced approach manages to store the current or the latest acceptable state (e.g. store a checkpoint of the model being trained) of the system, such that the acceptable state can be recovered. This is particularly useful for improving the reliability of a training process.

During a control transfer, situations can exist where the transfer of a part of required control points fails. To minimize the chance of loss of control of an AI system, the following actions should be considered to handle uncertainty:

— Identify atomic operations.

— Specify and implement redo and undo procedures for atomic operations. This can involve the recovery of environments changed by control transfer. In such a situation, undo or redo of an atomic operation can be influenced.

6.9 Cost of control

6.9.1 Consequences of control

The aim of estimating the cost of control is to provide information for determining the feasibility of such control. The following consequences can occur when controlling an AI system:

a) Incomplete work: The internal parameters or the appearance of the AI system are constantly changing. When there are data, communication or materials that have not been completely processed, there is the risk of loss of data, incomplete communication or loss of materials.

b) Resource consumption: The control instructions via function invocations, signal transmissions or physical operations can take time, energy, communication bandwidth, storage, space and any other resources needed. The resources required should be checked when the controller or the AI system under control has limited resources. An estimation of resource consumption during the process of control should be considered. There can be additional effects to the environment (e.g. temperature or electromagnetic change) where the AI system works.

6.9.2 Cost estimation for a control

The cost of control for the controller, the AI system, other entities and the environment should be estimated and checked, including the following:

a) whether the magnitude of resources required by a control exceeds the limits of the system. Trade-offs between the cost of control and the system's quality requirements based on ISO/IEC 25059 should be considered;

b) whether the magnitude of resources required by a control affects the system's functioning currently or in the future;

c) whether the possible changes to the environment or entities that the system works with affect the system's functioning according to business requirements;

d) cognitive, physiological and physical capabilities of human controllers (e.g., reaction speeds for drivers of a vehicle).

Once estimated, the cost of control should be provided to the controller or intended stakeholders of an AI system, who determines the acceptability of the cost (see Figure 3) and take further actions regarding control of the AI system

6.10 Cost of control transfer

6.10.1 Consequences of control transfer

The aim of estimating a cost of control transfer is useful for determining the feasibility of an intended control transfer. The following consequences can exist when a control transfer takes place from an AI system to a controller:

a) Out of control state: When a transfer of control happens, the AI system releases a specific set of control points to the controller. It is possible that the controller is not capable of managing the control points due to the possibly large number of control points or the complexity of the engagement process. As long as there is at least one control point that cannot be managed by the controller, the control of the AI system can be lost.

b) Resource consumption: Several kinds of resources, including time duration, signal transmission bandwidth, storage, energy, etc., can be consumed by a control transfer process.

6.10.2 Cost estimation for a control transfer

The following should be checked when estimating the cost of a control transfer:

a) whether the control transfer makes use of resources that are required by the system's functioning;

b) whether the control transfer makes use of a number of resources exceeding the system limits;

c) whether the control transfer can lead to an out of control state.

6.11 Collaborative control

In an AI system, more than one component can exist that can listen for and execute control instructions. Based on system design, there can also be multiple controllers. Each controller can issue control instructions to one or more components. Controllers or controllable components collaborate for achieving a goal. Collaborative control can be involved in the following cases:

a) Multi-controllable components and one controller: An AI system contains multiple components (e.g. an AI-based multi-agent system) that each can listen for and execute controllability instructions from the controller.

b) One controllable component and multi-controllers: An AI system (e.g. a robot controlled by multiple external human controllers) contains a component that can listen for and execute controllability instructions from multiple controllers.

c) Multi-controllable components and multi-controllers: An AI system (e.g. a group of robots and controlled by multiple human controllers) contains multiple components and each component can listen for controllability instructions from multiple controllers.

For each of these cases, controllability characteristics are described in Table 1.

Table 1 — Controllability characteristics for collaborative controls

	Multi-controllable components and one controller	One controllable component and multi-controllers	Multi-controllable components and multi-controllers
Process of control	For each controllable component and a control, the process in Figure 2 applies	For each control, the process in Figure 2 applies. Controllers synchronize (e.g. arrangement of the order of controls, resources obtainment and release) their control processes	For each controllable component, the process for one controllable component and multiple controllers applies
Control points	Union of control points of each controllable component	6.3 applies	Union of control points of each controllable component
Span of control	For each control, the span of control is deterministic. 6.4 applies	6.4 applies	For each control, the span of control in deterministic. 6.4 applies
Transfer of control	For each control, 6.5 applies	6.5 applies, in which the distribution of control points over controllers should be managed by controllers	For each control, 6.5 applies, in which the distribution of control points over controllers should be managed by controllers
Engagement of control	For each control, 6.6 applies	6.6 applies	For each control, 6.6 applies
Disengagement of control	For each control, 6.7 applies	6.7 applies	For each control, 6.7 applies
Uncertainty during control transfer	Besides the uncertainties in 6.8, failures that can happen on a part of controllable components should be considered	Besides the uncertainties in 6.8, communication failure between controllers should be considered	Besides the uncertainties in 6.8, failures that can happen on a part of controllable components as well as the communication failure between controllers should be considered
Cost of control	For each control, 6.9 applies	For each control, 6.9 applies. The resources spent during collaborations between controllers should be considered	For each control, 6.9 applies. The resources spent during collaborations between controllers should be considered
Cost of control transfer	For each control transfer, 6.10 applies	For each control transfer, 6.10 applies	For each control transfer, 6.10 applies

7 Controllability of AI system

7.1 Considerations

To realize the controllability of an AI system, the following should be considered:

a) The states of an AI system (see 5.2) should be observable and able to be transitioned. For this, an AI system should provide functionalities by which a controller can observe the system' states or at least obtain those internal parameters meaningful for control. The able to be transitioned system state refers

to the capacities of an AI system to accept and execute authorized control instructions at any intended time (see 5.3).

b) The following subprocesses of an AI system should be controlled:

1) Execution of non-fully-explainable processes: For systems that provide non-fully-explainable subprocesses due to the lack of a completed mapping between mathematic processes (e.g. mathematic computations defined by a deep learning neural network) and computational logics (semantically verifiable logic), those subprocesses should be controllable such that the potential hazards caused by unpredicted behaviours can be intervened and restricted. In this context, controls on the launch and termination of an unexplainable subprocess are important.

2) State observation: For AI systems that provide functionalities for state observation, the subprocesses that carry out sampling to those that return a system state should be controllable. In this context, a system state should be provided without any precondition except for authorization and authentication checks.

3) Control instruction execution: For AI systems that execute control instructions from an authorized and authenticated controller, the sequence of subprocesses that execute the control instructions should be controllable. In this context, the AI system should be able to accept and execute all control instructions.

4) Learning policy determination: For AI systems that can select the knowledge to learn from or determine the approach for learning (e.g. continuous learning), the subprocesses for such decisions should be controllable. This can be crucial for AI systems of which underlying learning policy can affect the AI system's behaviours towards human beings.

7.2 Requirements on controllability of AI systems

7.2.1 General requirements

7.2.1.1 Plan of controllability features

Controllability features should be planned during the inception or design and development stages of the AI system life cycle. The use of controllability features for risk identification and treatment should be prepared.

7.2.1.2 Description of controllability features

The provider of an AI system shall provide users with descriptions and documentation of the AI system's controllability features.

7.2.1.3 Requirements for ML-based AI systems controllability

For machine learning-based AI systems, requirements for controllability include:

a) The start and termination of an inference process shall be controllable.

b) For systems using a sequence of operations realized by executing multiple machine learning models, controls should be available on the transition between the execution of different models.

c) The observation of all system states should be enabled.

d) The observations to the input and output values of the following should be enabled:

1) entire system,

2) a module of the system,

3) specific neurons, layers or structures of a neural network for the systems where neural networks are used;

e) The observations to machine learning-based AI system execution logs and errors should be enabled. The availability of such information can help a controller to decide controls for minimizing potential hazards. When an AI system provides both asynchronous and synchronous modes for the execution of its subprocesses, controls should be available for controlling switching between the two modes. This enables an AI system to control the execution mode such that control decisions can be made in a timely way. Control can be missed if an asynchronous notification comes late and indicates a hazard.

7.2.1.4 Requirements on semantic computing-based AI systems controllability

For semantic computing-based AI systems, requirements on controllability include:

a) The start and termination of a reasoning process should be controllable.

b) When a system is able to perform automated reasoning over multiple kinds of knowledge representations, the selection and use of reasoners should be controllable.

c) The input data to and the output data from a reasoner should be observable.

d) The observation of system execution logs and errors should be enabled.

7.2.2 Requirements on controllability of continuous learning systems

7.2.2.1 ML-based continuous learning systems

For machine learning-based continuous learning systems, requirements for controllability include:

a) The start and termination of a learning process shall be controllable.

b) For AI systems using neural networks, during a backpropagation, gradient values of a relevant part of a neural network should be observable.

c) For those AI systems that automatically determine the content to learn, the selection and change of the content to learn should be controllable.

7.2.2.2 Semantic computing-based continuous learning system

For semantic computing-based continuous learning systems, the following should be controllable:

a) selection of the ontologies to be built as well as the priorities of new knowledge to be merged during a knowledge fusion process;

b) selection of the ontologies on which knowledge computing is performed.

7.3 Controllability levels of AI systems

The controllability levels of AI systems include:

a) Completely controllable: An AI system, in any state, is able to listen for and execute control and state observation instructions. The system can respond to control and state observation instructions. The execution of control (or a sequence of controls) and state observation (or a sequence of state observations) can be completed within an acceptable resource consumption limitation that meets the defined requirements. The system can reach the required state within resource constraints, including energy, time and processing cycles.

b) Partially controllable: An AI system, in a certain set of selected states, is able to listen for and execute control and state observation instructions. The system can respond to control instructions and reach the required state. The execution of control can be completed within an acceptable resource consumption limitation that meets the defined requirements. When the system is in a state other than

one of the selected states, the system can reach the required state by a sequence of controls, but resource consumption can be outside of acceptable limits.

c) Sequentially controllable: An AI system, in any state, can respond to control and state observation instructions. System cannot reach any required state by the execution of one control, but is able to reach any required state by a sequence of state observations and controls. Consumed resources can be outside acceptable limits.

d) Loosely controllable: An AI system, in any state, can respond to control and state observation instructions. System cannot reach the required state by the execution of one control. The system cannot guarantee that it can reach a required state via a sequence of controls and state observations. Consumed resources can be outside acceptable limits.

e) If an AI system is not controllable, all of the following apply at the same time:

   1) There is no state identified or defined.

   2) Only part of the parameters or appearances of the AI system are observable.

   3) There are no instructions implemented for state transitions.

   4) The system does not provide any instruction that can be used to make the system reach a required state.

NOTE 1 Not controllable level is applicable to those systems or scenarios where controllability is not required.

NOTE 2 The functionality termination of an AI system is a basic requirement that can be designed and implemented not for control. This feature is not required in the levels of controllability.

NOTE 3 System states can be observed via the approaches in 6.1 a).

8 Design and implementation of controllability of AI systems

8.1 Principles

Stakeholders should consider the following principles during the crucial AI system life cycle design and development stage:

a) Derive controllability features based on not only the explicitly specified requirements, but also those implicit necessities indicated by scenarios where the AI system can cause unwanted outcomes without adequate control. The following specific types of requirements can be considered:

   1) Adapted requirements are not explicitly stated but can be adapted from the environment through learning.

   2) Delegated requirements can come from another AI system, in a system-of-system structure. Requirements delegated from another subsystem or super-system can be another type.

b) Plan controllability features depending on the AI system's functionality, but implement them independently from the AI system's functionality design and development as follows:

   1) Controllability features are required during the AI system's execution. Implementation and use of controllability can be subject to what AI system functionality is performed.

   2) To improve the effectiveness and efficiency of controllability, design and development should not depend on the AI system's functionality implementation.

c) It is efficient for control if these state observation and control can be implemented separately:

   1) State observation and control make use of separate communication channels.

2) State observation and control are not subject to an identical group of shared resources.

d) A “stop” control that stops an AI system from executing its current task should always be considered during design and development. Cost of the "stop" control can be a valuable reference rather than a determinant of whether the "stop" control is implemented or applied.

For safety-critical AI systems, classical control and monitoring systems typically have an unchanging performance envelope. This is not the case with some continuous learning AI systems, where the performance envelope can change over time.

Developers should consider the set points and other forms of algorithmic goals of such systems, the appropriateness and sufficiency of behavioural constraints put in place, and what bearing these have on ensuring the system remains in a safe state and, in adverse situations, recovers from unsafe states

8.2 Inception stage

During the inception stage of an AI system, controllability functionalities should be considered, including:

a) Determine the objectives of each controllability functionality of an AI system, including but not limited to the following:

1) problems this controllability functionality solves;

2) customer needs or business opportunities that the controllability functionality addresses;

3) metrics of success.

b) Identify the requirements for each controllability functionality (control or state observation), including:

1) For each interaction between a controller and an AI system, the following should be analysed and recorded:

i) casual relationship between a controller’s instruction and the behaviour or appearance the system should exhibit;

ii) the system state and the control actions that can be applied to the system when it is in that state;

iii) after a control, the state in which the system is.

2) Based on the result of 8.2 b) 1), determine the requirements on control functionalities.

3) Based on the result of 8.2 b) 1), determine the requirements on system state observation functionalities.

4) A requirement can contain functional and non-functional concerned aspects.

5) Each aspect can contain specific measures (see 9.1.3 and 9.1.4) and values that a tested AI system is supposed to meet.

c) Identify the controllability functionalities useful in typical scenarios in which the system is supposed to be used. This should be done in particular to prevent or stop an AI system from causing harm. The range of identified controllability functionalities by this work is more extensive in comparison to the identification of requirements (see b)) that merely meet system specification. The following should be performed by stakeholders:

1) Controllability scenario identification discovers the scenarios where control or state observation functionalities are needed. For each scenario, determine the following:

i) the expected system outputs or behaviours if controllability functionalities are executed normally;

ii) the potential unwanted outcomes the system can perform if controllability functionalities are not executed normally.

2) Based on the result of 8.2 c) 1), for each scenario, determine the acceptance criteria, including but not limited to functional and non-functional aspects (e.g. performance efficiency, security and stability). Each aspect can correspond to a set of measures and values that a tested AI system is supposed to meet.

For each controllability functionality identified in 8.2 b) and 8.2 c), determine the state observation technical features that support system transparency and accountability, as controllability is a technical prerequisite of human oversight to AI systems (see 5.1).

Analyse the feasibility of each controllability functionality identified in 8.2 b) and 8.2 c). It can be done with a proof-of-concept of the AI system. For a controllability functionality, the analysis on feasibility includes but is not limited to items specified in 6.9.2 and 6.10.2.

In the inception stage defined in ISO/IEC 22989:2022, 6.2, the term cost is about funding. It is different with the term cost used in this document. The term cost in this document refers to the resources controls and control transfers consume. The funding-related cost for controllability functionalities should be forecast for the AI system over its entire system life cycle.

For safety-critical AI systems, requirements should be identified before the system (any software or hardware) design is undertaken, as it is usually not possible to retrofit safety design features.

Constraints on the AI system's socio-technical (human, procedural and technical) components and their interactions, and implement socio-technical controls to ensure they are not violated (see, for example, Systems theoretic accident model and process (STAMP), Systems Theoretic Process Analysis (STPA), in Reference [1]).

8.3 Design stage

8.3.1 General

The design stage of an AI system provides details for the system fulfilling requirements and targets, according to the outcomes of the inception stage. In ISO/IEC 22989:2022, 6.2.3, the design of an AI system can involve various aspects including approach, architecture, training data and risk management.

8.3.2 Approach aspect

The design of an AI system's controllability is subject to the AI system's set of states since controls are performed on a span of control. Stakeholders can select from the following models of design and apply them depending on their investigation on the AI system's states:

a) When all the system states can be foreseen by a designer, a good way to realize controllability is to adopt engineering methods that ensure computational control of the AI system by using a finite number of states and state transitions. This design provides a useful approach to ensure the controllability of an AI system based on the oversight of external agents that use the external interfaces of the system.

b) When some of the system states cannot be foreseen by a designer, the designer can realize controllability by analysing each state and defining system states if identified and meaningful for control. For an AI component, when its possible states can be entirely predetermined, the model in a) can be applied.

It is not necessary to design controllability of an AI system from scratch, but to take advantage of features of computing devices as well as enabling software, such as a machine learning software toolkit. For example, in a deep learning-based AI system, the capacities of controls and state observations over certain parts of a model, such as a neuron, a layer or a structure, during a forward propagation or a backpropagation can be inherited from enabling software.

8.3.3 Architecture aspect

It is beneficial to design controllability features in line with the formation of an AI system's architecture, such that the design of control and state observation can take system functionalities and components into account. Call backs (e.g. short cuts in procedures) can be used for state observation as well as synchronization mechanisms (e.g. waiting for the completion of control points' transfer and notifying the controller or the AI system) can help to let an AI system be available and meet the criteria (see 6.5 and 6.6) of control or control transfer.

8.3.4 Training data aspect

The design of controllability on a learning process can improve the usage effectiveness and efficiency as well as security. Controllability on a learning process includes:

a) control of the start and the termination of a learning process;

b) observation over interested parts of a deep neural network during forward propagation and backpropagation;

c) control of the selection and change of data to learn from.

8.3.5 Risk management aspect

Controllability features should be designed in order to technically satisfy the needs of planned risk assessment and treatment activities for an AI system, according to ISO/IEC 23894:2023,[12] 6.4 and 6.5.

8.3.6 Safety-critical AI system design considerations

The following can be in addition considered during the design of safety-critical AI system's controllability:

a) use of independent safety systems and safeguards, such as watchdog timers and hardware interlocks, independent monitors or simple fail-safe systems;

b) real-time systems methods where deterministic timing or other deterministic behaviour is required;

c) fault management techniques in AI system architecture, with primacy given to fault avoidance and fault removal, followed by designing for detection and fault tolerance, and communication to control agents (such as humans or other monitoring systems);

d) for an overview of functional safety methods, see ISO/IEC TR 5469.

8.4 Suggestions for the development stage

The development of an AI system's controllability corresponds to the processes realizing control and state observation functionalities, including but not limited to programming, documenting, testing, bug fixing, etc. The target of development of an AI system's controllability is to realize the required functionality with effectiveness and without introducing any decline or variation in performance. For this, the following suggestions should be considered:

a) Separate the ownership and use of computing resources (e.g. memory, communication bandwidth and processor) between controllability and system functionalities. It is important to provide adequate computing resources for control and state observations when controllability is expected to be executed immediately in time-deterministic cases.

b) Provide proper priorities to controllability instructions execution. In an IT system, computing tasks are scheduled by fundamental software (e.g. operating system) by a unified component. Equivalent distribution of priorities over controllability and other tasks can bring risk of late execution of control or state observations. This is important for those AI systems where controllability is expected to be executed immediately in time-deterministic cases.

c) Make use of controllability functionalities provided by layers of an AI system and avoid redundant encapsulation or re-implementation, if applicable. Controllability functionalities of an AI system are used through control points and controllability functionalities can be provided by certain layers (e.g. AI functionalities and computing resources in Figure 1). It can be more effective and efficient to use those basic state observation and control functionalities due to their large test and application scales.

9 Verification and validation of AI system controllability

9.1 Verification

9.1.1 Verification process

The aim of the verification of an AI system's controllability is to confirm whether its implemented controllability functionalities meet specified requirements. Verification is a stage defined in AI system life cycle in ISO/IEC 22989.

NOTE Detailed definition of AI system life cycle processes is specified in ISO/IEC 5338, which is aligned with ISO/IEC 22989:2022, Clause 6.

Verification should include the following:

a) Identify controllability functionality requirements that an AI system should provide, including control and state observation. This work should be done during the inception stage (see 8.3). If this identification has not been done before verification, it should be done before testing of controllability (see 9.1.1 b)).

b) Test the controllability functionalities to confirm they correctly implement the following requirements:

   1) For a requirement on controllability, design and perform a test. The test environment, test data and system configuration, input as well as output should be prepared.

   2) Test environment is represented by a set of parameters (e.g. temperature, humidity, network bandwidth, processor utilization ratio, process or threads inter-communication, time and region) with which the intended test is to be performed. A parameter should be considered if it can potentially affect the results of control or state observation.

   3) Test data and system configuration are the data and settings necessary to drive the system to be in specific states such that the tested control or state observation is meaningful and can be performed.

   4) Input refers to the control or state observation instruction.

   5) Output corresponds to the effects (5.3.4) and side effects (5.3.5) that a control or state observation can lead to. For a control, the outputs include the returned messages containing a system's states. For a state observation, the output is the system state.

c) The actual outputs of a controllability functionality should be compared with the expected and unexpected effects and side effects. For a requirement on controllability, at least the functional correctness and efficiency should be compared. Other aspects of efficiency required in a specific scenario should be considered.

d) The verified controllability functionalities should be listed with their expectations and actual results.

9.1.2 Output of verification

The verification process of an AI system's controllability should be documented. Annex A describes a form that can be used to document the verification process. It can be used in a test completion report.

9.1.3 Functional testing for controllability

Functional testing checks whether an AI system's controllability functionalities meet requirements. Non-functional testing is described in 9.1.4. Since AI systems are designed and used in different domains, the

measures of controllability functional correctness can be diverse and domain-specific. The following types of measures can be considered for the functional testing of AI systems' controllability:

a) Discrete measure: When an AI system's controllability functionality returns results predefined in a limited universe with discrete elements (e.g. an integer number representing control execution success), the measure is to determine the relationship between the actual returned and the expected values given the designed system configurations and states.

   EXAMPLE 1 A floor cleaning robot provides a controllability functionality start clean that can be triggered by a physical button on the robot body. This control can be performed when the system is in the state ready for control indicated by a light on its body. The expected result of this control is a predefined code called starting to clean and occurs when the indicator light is on. The functional testing for start clean is to press the button and to check whether the system is able to return the predefined code starting to clean and whether the indicator light is on. In this scenario, the measure can only return a Boolean value.

b) Continuous measure: When an AI system's controllability functionality is required to return a result with auxiliary values indicating to what extent the control is performed or the system changes, the measure is to determine whether the returned auxiliary values are correct given the designed system configurations and states.

   EXAMPLE 2 A floor cleaning robot provides a controllability functionality go forward when it is in cleaning state, to provide the ability to a user to manually control and clean irregular shaped areas. This control can be issued via pressing a button on a remote controlling device and can return a result going forward with the distance the robot actually passes physically. The requirement on this controllability functionality can be met when the robot returns going forward and passes a predetermined distance (e.g. 10 cm). In this scenario the measure returns not only an indication of going forward but also the difference between the actual distance the robot moves forward and the requirement (10 cm).

In an AI system, controllability functionalities can exist that are intended to ensure functional safety (e.g. the system response duration on controls of a real-time AI system, or the power consumption restriction of certain controls of a power supply restricted AI system). Functional testing of safety controllability functionalities should use 9.1.3 b).

9.1.4 Non-functional testing for controllability

Non-functional testing includes the tests on performance efficiency, security, stability, usability and any other aspects that can influence the execution of a controllability functionality.

Performance efficiency testing measures the magnitude of resources consumed by executing a controllability functionality of an AI system and checks whether it meets requirements. It can provide evidence for optimizing system controllability design and implementation. The types of measures include but are not limited to:

a) Duration: This measure refers to the length of time needed by a controllability functionality, including preparation (see 6.5), instruction execution and communication regarding the outcome of the control action. The durations of important subprocesses (e.g. transfer of control and engagement of control) can be checked:

   1) For a state observation, the duration measures the time from when the controller's instruction is issued to when the controller receives the response containing the requested system state.

   2) For control, the following durations can be applied based on requirements:

      i) the time from when a control is issued by a controller to the time when the controller receives the requested control;

      ii) the time when a control is issued by a controller to the time when the controller receives the result of the control.

b) Number of operations: This measure refers to the number of operations needed to be performed by controllers, when a controllability functionality is performed. This measure indicates the complexity of control and is important to those controls where physical operations are applied.