Technical Report

ISO/IEC TR 5469

First edition
2024-01

Artificial intelligence — Functional safety and AI systems

Intelligence artificielle — Sécurité fonctionnelle et systèmes d'intelligence artificielle

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

Reference number
ISO/IEC TR 5469:2024(en)

© ISO/IEC 2024

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 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

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

Contents

Page

Foreword ... **v**
Introduction ... **vi**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Abbreviated terms ... **4**
5 Overview of functional safety ... **4**
5.1 General ... 4
5.2 Functional safety ... 5
6 Use of AI technology in E/E/PE safety-related systems ... **6**
6.1 Problem description ... 6
6.2 AI technology in E/E/PE safety-related systems ... 6
7 AI technology elements and the three-stage realization principle ... **10**
7.1 Technology elements for AI model creation and execution ... 10
7.2 The three-stage realization principle of an AI system ... 12
7.3 Deriving acceptance criteria for the three-stage of the realization principle ... 12
8 Properties and related risk factors of AI systems ... **13**
8.1 Overview ... 13
8.1.1 General ... 13
8.1.2 Algorithms and models ... 13
8.2 Level of automation and control ... 14
8.3 Degree of transparency and explainability ... 15
8.4 Issues related to environments ... 17
8.4.1 Complexity of the environment and vague specifications ... 17
8.4.2 Issues related to environmental changes ... 17
8.4.3 Issues related to learning from environment ... 18
8.5 Resilience to adversarial and intentional malicious inputs ... 19
8.5.1 Overview ... 19
8.5.2 General mitigations ... 19
8.5.3 AI model attacks: adversarial machine learning ... 19
8.6 AI hardware issues ... 20
8.7 Maturity of the technology ... 21
9 Verification and validation techniques ... **21**
9.1 Overview ... 21
9.2 Problems related to verification and validation ... 22
9.2.1 Non-existence of an a priori specification ... 22
9.2.2 Non-separability of particular system behaviour ... 22
9.2.3 Limitation of test coverage ... 22
9.2.4 Non-predictable nature ... 22
9.2.5 Drifts and long-term risk mitigations ... 22
9.3 Possible solutions ... 23
9.3.1 General ... 23
9.3.2 Relationship between data distributions and HARA ... 23
9.3.3 Data preparation and model-level validation and verification ... 24
9.3.4 Choice of AI metrics ... 25
9.3.5 System-level testing ... 25
9.3.6 Mitigating techniques for data-size limitation ... 26
9.3.7 Notes and additional resources ... 26
9.4 Virtual and physical testing ... 26
9.4.1 General ... 26
9.4.2 Considerations on virtual testing ... 26

Copyrighted document, no reproduction or circulation
For review by PG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

9.4.3 Considerations on physical testing ... 28
9.4.4 Evaluation of vulnerability to hardware random failures ... 29
9.5 Monitoring and incident feedback ... 29
9.6 A note on explainable AI ... 29

10 Control and mitigation measures ... 30
10.1 Overview ... 30
10.2 AI subsystem architectural considerations ... 30
10.2.1 Overview ... 30
10.2.2 Detection mechanisms for switching ... 30
10.2.3 Use of a supervision function with constraints to control the behaviour of a system to within safe limits ... 33
10.2.4 Redundancy, ensemble concepts and diversity ... 34
10.2.5 AI system design with statistical evaluation ... 35
10.3 Increase the reliability of components containing AI technology ... 35
10.3.1 Overview of AI component methods ... 35
10.3.2 Use of robust learning ... 35
10.3.3 Optimization and compression technologies ... 36
10.3.4 Attention mechanisms ... 37
10.3.5 Protection of the data and parameters ... 37

11 Processes and methodologies ... 38
11.1 General ... 38
11.2 Relationship between AI life cycle and functional safety life cycle ... 38
11.3 AI phases ... 39
11.4 Documentation and functional safety artefacts ... 39
11.5 Methodologies ... 39
11.5.1 Overview ... 39
11.5.2 Fault models ... 39
11.5.3 PFMEA for offline training of AI technology ... 40

Annex A (informative) **Applicability of IEC 61508-3 to AI technology elements ... 41**
Annex B (informative) **Examples of applying the three-stage realization principle ... 54**
Annex C (informative) **Possible process and useful technology for verification and validation ... 59**
Annex D (informative) **Mapping between ISO/IEC 5338 and the IEC 61508 series ... 62**
Bibliography ... 65

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

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.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

Introduction

The use of artificial intelligence (AI) technology in industry has increased significantly in recent years and AI has been demonstrated to deliver benefits in certain applications. However, there is limited information on specification, design, and verification of functionally safe AI systems or on how to apply AI technology for functions that have safety-related effects. For functions realized with AI technology, such as machine learning (ML), it is difficult to explain why they behave in a particular manner and to guarantee their performance. Therefore, whenever AI technology is used in general and especially when it is used to realize safety-related systems, special considerations are likely to arise.

The availability of powerful computational and data storage technologies makes the prospect of large-scale deployment of ML possible. For more and more applications, adopting machine learning (as an AI technology) is enabling the rapid and successful development of functions that detect trends and patterns in data. This makes it possible to induce a function's behaviour from observation and to quickly extract the key parameters that determine its behaviour. Machine learning is also used to identify anomalous behaviour or to converge on an optimal solution within a specific environment. Successful ML applications are found in analysis of, for example, financial data, social networking applications and language recognition, image recognition (particularly face recognition), healthcare management and prognostics, digital assistants, manufacturing robotics, machine health monitoring and automated vehicles.

In addition to ML, other AI technologies are also gaining importance in engineering applications. Applied statistics, probability theory and estimation theory have, for example, enabled significant progress in the field of robotics and perception. As a result, AI technology and AI systems are starting to realize applications that affect safety.

Models play a central role in the implementation of AI technology. The properties of these models are used to demonstrate the compatibility of AI technology and AI systems with functional safety requirements. For instance, where there is an underlying known and understood scientific relationship between the key parameters that determine a function's behaviour, there is likely to be a strong correlation between the observed input data and the output data. This leads to a transparent and sufficiently complete model as the basis for AI technology. In this case, compatibility of the model with functional safety requirements is demonstrated. However, AI technology is often used in cases where physical phenomena are so complex or at such a small scale, or unobservable without influencing the experimental data, that consequently there is no scientific model of the underlying behaviour. In this case, the model of the AI technology is possibly neither transparent nor complete and the compatibility of the model with functional safety requirements is hard to demonstrate.

Machine learning is used to create models and thus to extend the understanding of the world. However, machine-learnt models are only as good as the information used to derive the model. If the training data does not cover important cases, then the derived models are incorrect. As more known instances are observed they are used to reinforce a model, but this biases the relative importance of observations, steering the function away from less frequent, but still real, behaviours. Continuous observation and reinforcement moves the model towards an optimum or it overemphasizes common data and overlook extreme, but critical, conditions.

In the case of continuous improvement of the model through the use of AI technology, the verification and validation activities in order to demonstrate its safety integrity are undermined as the function behaviour progressively moves away from the rigorously tested, ideally deterministic and repeatable behaviour.

The purpose of this document is to enable the developer of safety-related systems to appropriately apply AI technologies as part of safety functions by fostering awareness of the properties, functional safety risk factors, available functional safety methods and potential constraints of AI technologies. This document also provides information on the challenges and solution concepts related to the functional safety of AI systems.

Clause 5 provides an overview of functional safety and its relationship with AI technology and AI systems.

Clause 6 describes different classes of AI technology to show potential compliance with existing functional safety International Standards when AI technology forms part of a safety function. Clause 6 further introduces different usage levels of AI technology depending on their final impact on the system. Finally,

Copyrighted document, no reproduction or circulation
For review by PG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Clause 6 also provides a qualitative overview of the relative levels of functional safety risk associated with different combinations of AI technology class and usage level.

Clause 7 describes, based on ISO/IEC 22989, a three-stage realization principle for usage of AI technology in safety-related systems, where compliance with existing functional safety International Standards cannot be shown directly.

Clause 8 discusses properties and related functional safety risk factors of AI systems and presents challenges that such use raises, as well as properties that are considered when attempting to treat or mitigate them.

Clauses 9, 10 and 11 show possible solutions to these challenges from the field of verification and validation, control and mitigation measures, processes, and methodologies.

The annexes provide examples of application of this document and additional details. Annex A addresses how IEC 61508-3 is applied to AI technology elements, and Annex B provides examples to how to apply three-stage realization principles and define various properties. Annex C describes more detailed processes related to 9.3. Annex D shows the mapping between safety life cycle in IEC 61508-3 and AI system life cycle in ISO/IEC 5338.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

Documents
Document Sharing Hub

Artificial intelligence — Functional safety and AI systems

1 Scope

This document describes the properties, related risk factors, available methods and processes relating to:

— use of AI inside a safety related function to realize the functionality;

— use of non-AI safety related functions to ensure safety for an AI controlled equipment;

— use of AI systems to design and develop safety related functions.

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*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
safety
freedom from *risk* (3.3) which is not tolerable

[SOURCE: IEC 61508-4:2010, 3.1.11]

3.2
functional safety
part of the overall *safety* (3.1) 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 and other risk reduction measures

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

3.3
risk
functional safety risk
<functional safety> combination of the probability of occurrence of *harm* (3.5) and the severity of that *harm* (3.5)

Note 1 to entry: For more discussion on this concept, see Annex A of IEC 61508-5.

[SOURCE: IEC 61508-4:2010, 3.1.6, modified — Added < functional safety > domain]

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

3.4
risk
organizational risk
<organizational> effect of uncertainty on objectives

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

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

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

Note 4 to entry: This is the core definition of risk. As risks are specifically focused on *harm* (3.5) a discipline specific definition of *risk* (3.3) is used in this document in addition to the core risk definition.

[SOURCE: ISO 31000:2018, 3.1, modified — Added <organizational> domain and Note 4 to entry]

3.5
harm
physical injury or damage to the health of people, or damage to property or the environment

[SOURCE: IEC 61508-4:2010, 3.1.1]

3.6
hazard
potential source of *harm* (3.5)

[SOURCE: IEC 61508-4:2010, 3.1.2]

3.7
hazardous event
event that may result in *harm* (3.5)

[SOURCE: IEC 61508-4:2010, 3.1.4]

3.8
system
arrangement of parts or elements that together exhibit a stated behaviour or meaning that the individual constituents do not

[SOURCE: ISO/IEC/IEEE 15288:2023, 3.46, modified — Removed the three Notes to entry]

3.9
systematic failure
failure related in a deterministic way to a certain cause, which can only be eliminated by a modification of the design or of the manufacturing process, operational procedures, documentation or other relevant factors

[SOURCE: IEC 61508-4:2010, 3.6.6]

3.10
safety-related system
designated system that both

— implements the required safety functions necessary to achieve or maintain a safe state for the EUC; and

— is intended to achieve, on its own or with other E/E/PE safety-related systems and other risk reduction measures, the necessary safety integrity for the required safety functions

[SOURCE: IEC 61508-4:2010, 3.4.1]

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

3.11
safety function
function to be implemented by an E/E/PE safety-related system or other risk reduction measures, that is intended to achieve or maintain a safe state for the EUC, in respect of a specific *hazardous event* (3.7)

[SOURCE: IEC 61508-4:2010, 3.5.1]

3.12
equipment under control
EUC
equipment, machinery, apparatus or plant used for manufacturing, process, transportation, medical or other activities

Note 1 to entry: The EUC control system is separate and distinct from the EUC.

[SOURCE: IEC 61508-4:2010, 3.2.1]

3.13
programmable electronic
PE
based on computer technology which can be comprised of hardware, software and of input and/or output units

Note 1 to entry: This term covers microelectronic devices based on one or more central processing units (CPUs) together with associated memories, etc.

EXAMPLE The following are all programmable electronic devices:

— microprocessors;

— micro-controllers;

— programmable controllers;

— application specific integrated circuits (ASICs);

— programmable logic controllers (PLCs);

— other computer-based devices (e.g. smart sensors, transmitters, actuators).

[SOURCE: IEC 61508-4:2010, 3.2.12]

3.14
electrical/electronic/programmable electronic
E/E/PE
based on electrical (E) and/or electronic (E) and/or programmable electronic (PE) technology

Note 1 to entry: The term is intended to cover any and all devices or systems operating on electrical principles.

EXAMPLE Electrical/electronic/programmable electronic devices include:

— electro-mechanical devices (electrical);

— solid-state non-programmable electronic devices (electronic);

— electronic devices based on computer technology (programmable electronic).

[SOURCE: IEC 61508-4:2010, 3.2.13]

3.15
AI technology
technology used to implement an *AI model* (3.16)

Copyrighted document, no reproduction or circulation
For review by PG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

3.16
AI model
physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data

[SOURCE: ISO/IEC 22989:2022, 3.1.23, with the addition of AI]

3.17
test oracle
source of information for determining whether a test has passed or failed

[SOURCE: ISO/IEC/IEEE 29119-1:2022, 3.115]

4 Abbreviated terms

ALARP	as low as reasonably practicable
ANN	artificial neural network
CNN	convolutional neural network
CPU	central processing unit
CUDA	compute unified device architecture
DL	deep learning
DNN	deep neural network
GPU	graphics processing unit
EDDM	early drift detection method
E/E	electrical and/or electronic
E/E/PE	electrical/electronic/programmable electronic
EUC	equipment under control
FMEA	failure modes and effects analysis
GAMAB	globalement au moins aussi bon
HARA	hazard analysis and risk assessment
HAZOP	hazard and operability analysis
JPEG	joint photographic experts group
KPI	key performance indicator
MEM	minimum endogenous mortality
SVM	support vector machines

5 Overview of functional safety

5.1 General

The discipline of functional safety is focused on risks related to injury and damage to the health of people, or damage to the environment and, in some cases, mitigation against damage to product or equipment. The

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

definition of risk differs based on the domain tags as shown in Clause 3. Both definitions are valid concepts for the use of AI. All references to risk in this document from this point on are related to the definition from the functional safety domain.

According to IEC 61508-1, control of risk is an iterative process of risk assessment and risk reduction. Risk assessment identifies sources of harm and evaluates the related risks for the intended use and the reasonably foreseeable misuse of the product or system. Risk reduction reduces risks until they become tolerable. Tolerable risk is a level of risk that is accepted in a given context based on the current state of the art.

The IEC 61508 series recognizes the following three-step (prioritised) approach as being good practice for risk reduction:

— Step 1: inherently functionally safe design;

— Step 2: guards and protective devices;

— Step 3: information for end users.

Risk reduction via the provision of functional safety is associated with Step 2.

This document focuses on the aspects of safety functions performed by a safety related system by making use of AI technology, either within the safety related system or during design and development of the safety related system (Step 2).

This document makes no provision of methodology for AI technology used for Steps 1 and 3.

5.2 Functional safety

IEC 61508-4[19] defines functional safety as that "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 and other risk reduction measures." The E/E/PE safety-related system is delivering a "safety function", which is defined in IEC 61508-4 as a "function to be implemented by an E/E/PE safety-related system or other risk reduction measures, that is intended to achieve or maintain a safe state for the EUC, in respect of a specific hazardous event." In other words, the safety functions control the risk associated with a hazard that leads to harm to people or the environment. The safety functions also reduce the risk of having serious economic implications.

As the term implies, functional safety - as defined in IEC 61508-4 - aims to achieve and maintain functionally safe system states of an EUC through the provision of safety functions. Based on the inclusion of "other risk reduction measures" in the definition of functional safety and safety functions, non-technical functions are explicitly included. The EUC is not limited to individual devices but it includes also systems.

Following these definitions, functional safety as a discipline is thus concerned with the proper engineering of these technical and non-technical safety functions for risk reduction or risk level containment of a particular equipment under control, from the component level up to the system level, including considering human factors, and under operational or environmental stress.

Functional safety focuses on safety functions for risk reduction and the properties of these functions required for risk reduction. While the functionality of a safety function is strongly use-case dependent, the properties required for risk reduction and the related measures are the main focus of functional safety standardization.

Prior to the advent of programmable systems, when safety functions were limited to implementation in hardware, the focus of functional safety was to reduce the consequences and the likelihood of random hardware failures. With software being increasingly used to implement safety functions, the focus shifted towards systematic failures introduced during design and development.

NOTE ISO 21448:2022[7] includes requirements on safety of the intended functionality including aspects such as performance limitation. Annex D describes implications for machine learning.

Copyrighted document, no reproduction or circulation
For review by FC on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub

Full document details: ISO_IEC TR 5469 WG_2024

There is a robust body of knowledge on how to avoid systematic failures in non-AI systems and in software development.[138] This document considers the use of AI technology in the context of safety functions. Functions containing AI technology, especially machine learning, typically follow a different development paradigm to that of non-AI systems. They are less specification-driven and more driven by observation of the data defining the system behaviour. For this reason, the catalogue of available measures for dealing with systematic failures is extended with respect to the specificities of AI technologies: Annex A provides an example of that extension. AI-specific risk reduction measures also differ from non-AI systems from a functional perspective. Functional safety puts a focus on systematic capabilities (IEC 61508-4:2010, 3.5.9) in addition to random hardware and systematic failures throughout the life cycle.

The relevance of AI technologies when used to realize a safety function is their potential to address new methods for risk reduction. This document examines the use of such technologies for this purpose, while maintaining existing risk reduction concepts, by introducing risk and classification considerations.

In general, achieving an acceptable risk level for increasingly complex and automated systems is likely to depend on advanced safety concepts. This includes the adequate implementation of both technical and non-technical risk reduction measures to achieve and maintain safe system states. Assuring the validity of such advanced safety concepts is a great challenge in functional safety. It also leads to an increase in the number of functional safety requirements. For all technical risk reduction measures, the distinction is made that hardware random faults and systematic faults are considered, which is done in basic International Standards like the IEC 61508 series or derived International Standards. However, for safety functions including AI technology, it is inevitable that there is additional focus on the assurance that systematic capabilities of systems that implement these functions are sufficient for the intended use environment.

6 Use of AI technology in E/E/PE safety-related systems

6.1 Problem description

The use of AI technology and AI systems is currently not treated in mature functional safety International Standards (indeed, in some International Standards their use is explicitly forbidden). International Standards that include AI-related contents include ISO 21448[7].

6.2 AI technology in E/E/PE safety-related systems

E/E/PE safety-related systems have a set of properties to ensure that they provide the intended safety mitigation measures in a dependable way. These properties are ideally generic and application independent. However, the data and the specifications vary based on application and technology domain. The process in which properties are selected is described in Figure 3 for each of the three stages of the three-stage realization principle. The properties are selected on a case-by-case basis, as relevant to a particular application or technology domain, their data and specifications. Properties are based on existing International Standards, including the IEC 61508 series[16]-[19], the ISO 26262 series[12]-[15], IEC 62061[21] and the ISO 13849 series.[5],[8] Clause 8 provides a list of typical properties.

Satisfying the selected properties is likely to place particular functional safety requirements on the realization, installation, validation, operation and maintenance of such systems. For example, IEC 61508-3[18] defines such requirements for the software part of E/E/PE systems. However, several AI technologies use different development approaches (e.g. learning-based) compared to the non-AI software engineering life cycles targeted by IEC 61508-3.

To address the difference between traditional development processes and the approach typical of AI technologies, this clause provides a general classification scheme for the applicability of AI technology in E/E/PE safety-related systems, based on various contexts of the application of AI technology.

An example of a classification scheme is, summarized in Table 1 and the related flowchart represented in Figure 1. The scheme is intended to provide insight on how an AI technology is addressed in the context of functional safety for a specific application.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

The classification scheme (see Table 1) is organized along two axes:

— AI Application and Usage Level. This axis considers the application of the AI technology and includes, among other things, the way in which it is used. It is classified from A to D, with two intermediate levels for A and B.

NOTE 1 The factors identified in Clause 8 are of high relevance in the context of the classification. These factors are described further in Clause 8 and include the level of automation and control (see 8.2), the degree of decision transparency and explainability (see 8.3), the complexity of the environment and vague specifications (see 8.4), security (see 8.5), system hardware issues see (8.6) and the maturity of the technology (see 8.7).

An example of a classification of Usage Level is as follows:

— Usage Level A1 is assigned when the AI technology is used in a safety-relevant E/E/PE system and where automated decision-making of the system function using AI technology is possible;

— Usage Level A2 is assigned when the AI technology is used in a safety-relevant E/E/PE system and where no automated decision-making of the system function using AI technology is possible (e.g. AI technology is used for diagnostic functionality within the E/E/PE system);

NOTE 2 The evaluation can change depending on the role of the diagnostic function, such as whether the diagnostic is critical to maintaining the functional safety of the system or is merely a minor contributor to functional safety amongst many others.

— Usage Level B1 is assigned when the AI technology is used only during the development of the safety-relevant E/E/PE system (e.g. an offline support tool) and where automated decision-making of the function developed using AI technology is possible;

— Usage Level B2 is assigned when the AI technology is used only during the development of the safety-relevant E/E/PE system (e.g. an offline support tool) and where no automated decision-making of the function is possible:

— Usage Level C is assigned when the AI technology is not part of a functional safety function in the E/E/PE system, but can have an indirect impact on the function:

NOTE 3 The Usage Level C includes AI techniques clearly providing additional risk reduction and whose failure is not critical to the level of acceptable risk.

EXAMPLE An AI technique that increases or decreases the demand rate placed on a safety system.

— Usage Level D is assigned if the AI technology is not part of a safety function in the E/E/PE system and has no impact on the safety function due to sufficient segregation and behaviour control.

NOTE 4 An example is separation through a “sandbox” or ”hypervisor” in such a way that it does not affect the safety functionality.

— AI Technology Class. This axis considers the level of fulfilment of AI technology in satisfying the identified set of properties, in which:

  — Class I is assigned if AI technology is developed and reviewed using existing functional safety International Standards, for example, if the properties and the set of methods and techniques leading to achievement of the properties are identified using existing functional safety International Standards;

  — Class II is assigned if AI technology cannot be fully developed and reviewed using existing functional safety International Standards, but it is still possible (as shown in Figure 4) to identify additional complementary requirements, methods and techniques for development, design, verification and validation of the desired safety properties to achieve the necessary risk reduction;

  — Class III is assigned if AI technology cannot be developed and reviewed using existing functional safety International Standards and it is also not possible to satisfy the set of identified properties with related methods and techniques.

Copyrighted document, no reproduction or circulation
For review by FC on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Components containing AI technology are composed of various technology elements (see 7.1). Each element belongs to a different AI technology class. For example, the lower levels of abstraction of a neural network are implemented using C++ libraries, which by themselves are systematically reviewed (e.g. according to the requirements in IEC 61508-3[18], see the example in Annex A). As such, they are classified as Class I, though the higher level of abstractions (e.g. deep learning models) are more likely classified as Class II or Class III.

For AI Technology Class I components, application of existing functional safety International Standards is possible and desired in general. For AI Technology Class II components, use of existing functional safety International Standards leads to partial achievement of the properties required for functional safety. Then the residual part is addressed using a set of complementary methods and techniques such as additional verification and validation (as detailed in Clause 9). Effectiveness of complementary methods and techniques are different between applications and Usage Levels. Components in AI Technology Class III, at the time of publication of this document, lack known methods and techniques to identify a set of properties to achieve sufficiently reduction of risk.

Table 1 — Example of AI classification table

<table>
<tr><th>AI Technology Class = > AI application and usage level</th><th>AI technology Class I</th><th>AI technology Class II</th><th>AI technology Class III</th></tr>
<tr><td>Usage Level A1 [a]</td><td rowspan="5">Application of risk reduction concepts of existing functional safety International Standards possible</td><td>Appropriate set of requirements [c]</td><td rowspan="5">At the time of publication of this document no appropriate set of properties with related methods and techniques is known to achieve sufficiently reduction of risk</td></tr>
<tr><td>Usage Level A2 [a]</td><td>Appropriate set of requirements [c]</td></tr>
<tr><td>Usage Level B1 [a]</td><td>Appropriate set of requirements [c]</td></tr>
<tr><td>Usage Level B2 [a]</td><td>Appropriate set of requirements [c]</td></tr>
<tr><td>Usage Level C [a]</td><td>Appropriate set of requirements [c]</td></tr>
<tr><td>Usage Level D [b]</td><td colspan="3">Application of risk reduction concepts of existing functional safety International Standards</td></tr>
</table>

[a] Static (offline) (during development) teaching or learning only.

[b] Dynamic (online) teaching or learning possible.

[c] The appropriate set of requirements for each usage level is established by application of risk reduction concepts of existing functional safety International Standards and additional considerations based on the literature review performed in Clauses 8, 9, 10 and 11. Examples are provided in Annex B.

Copyrighted document, no reproduction or circulation
For review by EG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

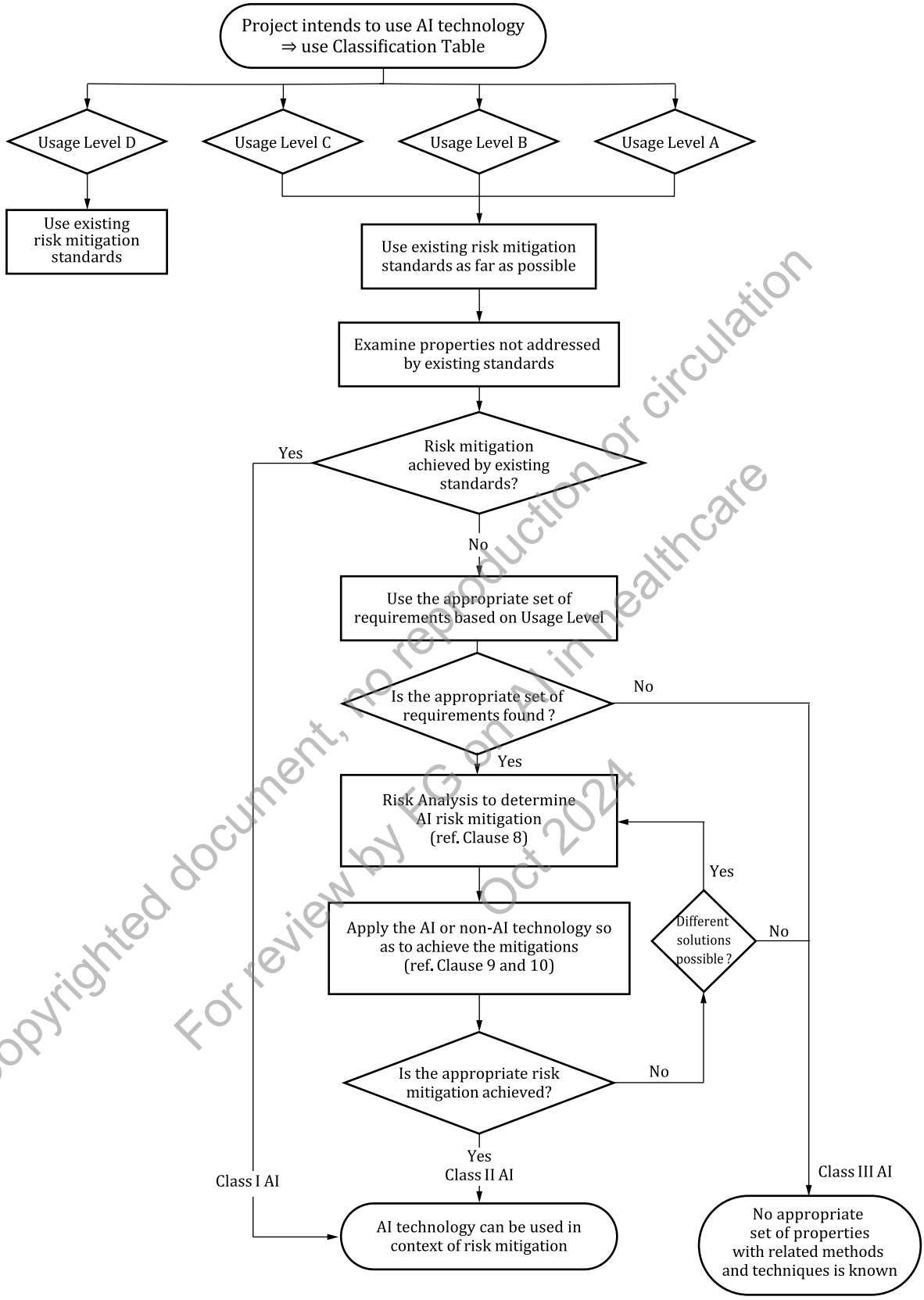


Figure 1 — Flowchart for determining classification of AI technology

Copyrighted document, no reproduction or circulation
For review by ECG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

7 AI technology elements and the three-stage realization principle

7.1 Technology elements for AI model creation and execution

The creation and the execution of an AI model involves different AI technology elements. Table 2 provides a high-level description of the AI landscape and the typical technology elements involved, based on the functional layers of an AI ecosystem as described in ISO/IEC 22989:2022, Figure 6. Table 3 is an example of those technology elements for the specific case of machine learning.

Table 2 — AI technology elements (from ISO/IEC 22989:2022)

AI technology element
AI services
Machine learning — Model development and use — Tools — Data for machine learning
Engineering — Knowledge based on domain experience — Tools
Cloud and edge computing and big data and data sources
Resource pool-compute, storage, network
Resource management-resource provisioning

Table 3 — Example technology elements involved in model creation and execution for ML

Technology element (machine learning example)	Examples (not exhaustive)[b]
Application graph	Graph eXchange Format (GXF) graph in YAML Ain't Markup Language (YAML), recently qualified teacher (rqt) graph in Robot Operating System (ROS™/ROS2™)
Machine learning framework[a]	TensorFlow®, PyTorch®, Keras, mxnet, Microsoft Cognitive Toolkit (CNTK), Caffe, Theano
Machine learning model language	Open Neural Network Exchange (ONNX®), Neural Network Exchange Format (NNEF™) PolyML
Machine learning graph compiler	TensorRT™, Glow, Multi-Level Intermediate Representation (MLIR), nGraph, Tensor Virtual Machine (TVM), PlaidML, Accelerated Linear Algebra (XLA)
Machine code compiler	Gcc, nvcc, clang/llvm, SYCL, dpc++, OpenCL™, openVX™

NOTE This table does not distinguish between elements used for training and those used for inference.

[a] A machine learning framework is an end-to-end machine learning platform including tools, libraries and community resources.

[b] The trade name(s) or trademark(s) in this table are examples of suitable products available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC of these products.

Copyrighted document - no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Table 3 *(continued)*

Technology element (machine learning example)	Examples (not exhaustive)[b]
Libraries	CUDA® C++, CUTLASS, NumPy, SciPy, Pandas, Matplotlib, CUDA Deep Neural Network (cuDNN™), SYCL DNN, oneAPI Deep Neural Network (OneDNN), Math Kernel Library (MKL)
Executable machine code	executable machine code compiled for the following architectures: aarch64, PTX, RISC-V®, AMD64, x86/64, PowerPC™
Computational hardware	GPU, CPU, FPGA, ASIC, accelerators

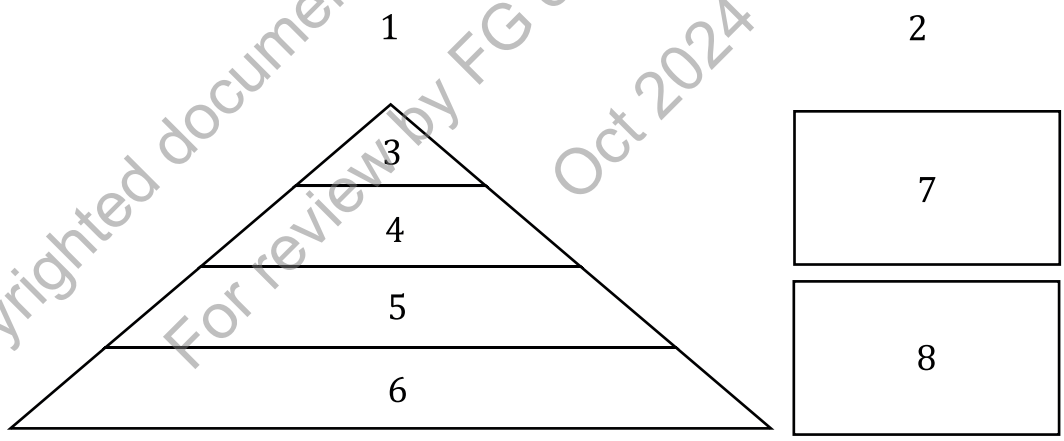
NOTE This table does not distinguish between elements used for training and those used for inference.

[a] A machine learning framework is an end-to-end machine learning platform including tools, libraries and community resources.

[b] The trade name(s) or trademark(s) in this table are examples of suitable products available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO or IEC of these products.

Some technology elements are addressed with existing concepts of functional safety (e.g. the software translating the model to an executable code). However, all technology elements involved in the model creation and execution are part of the safety considerations, including those that are handled with existing concepts of functional safety and those for which new concepts are defined. Annex A includes an example of how existing concepts of functional safety are applied to AI technology via assessment of the applicability of IEC 61508-3[18]. Annex B includes an example of how specific properties, such as the ones described in Clause 8, are applied to AI technology for which existing concepts of functional safety cannot be applied.

As shown in Figure 2, elements containing AI technology are used at different levels of a system or application: for the higher-level elements (e.g. application graph or ML model, and related tools) specific properties, such as the ones described in Clause 8, are applicable; the lower-level elements (e.g. executable machine code, and related tools) are handled with non-AI properties as described in this clause, such as the properties defined in the IEC 61508 series[16]-[19], the ISO 26262 series[12]-[15] and other International Standards.



Key

1 code/models
2 tools/libraries
3 application graph
4 ML model
5 executable machine code
6 computational hardware
7 ML framework, code/graph/ML compilers
8 libraries

Figure 2 — Hierarchy of technology elements (ML example)

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

7.2 The three-stage realization principle of an AI system

Summarizing ISO/IEC 22989:2022, Figure 5, an AI system is represented (see Figure 3) by a realization principle comprising three main stages:

— data acquisition;

— knowledge induction from data and human knowledge;

— processing and generation of outputs.

NOTE 1 With respect to ISO/IEC 22989:2022 Figure 5, the first stage is mapped to the inputs task, the second stage is mapped to learning task and the third stage to the processing task.

NOTE 2 In this context, human knowledge is derived from a range of different expertise, both in the relevant domain as well as in AI systems.

NOTE 3 The described realization principle is general. Specific more detailed examples for AI system are the Monitor-Analyse-Plan-Execute (MAPE) or Sense-Understand-Decide-Act (SUDA).

NOTE 4 The intent of the three-stage realization principle is not to describe a life cycle (that is described in Clause 11 and includes all stages from concept development and maturation through to development of requirements) but mainly to show that AI includes another point of view (the data).

NOTE 5 Figure 3 does not show feedback loops that apply for AI systems that are tightly bound into decision loops or that change the real-world situation.

NOTE 6 Knowledge induction includes training, while processing includes inference.

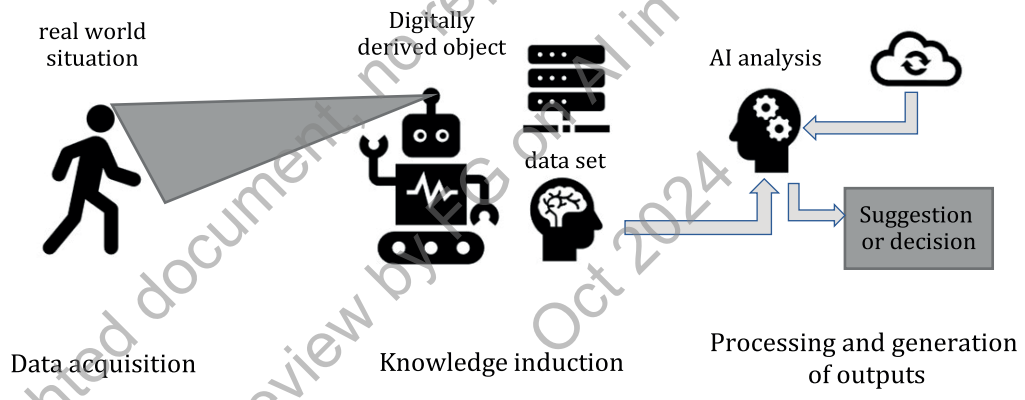


Figure 3 — Three-stage realization principle

Copyrighted document, no reproduction or circulation
For review by IEC on AI in healthcare
Oct 2024

7.3 Deriving acceptance criteria for the three-stage of the realization principle

Typically (see for example in ISO/IEC 22989:2022), a three-stage realization principle (see Figure 4) is used to derive acceptance criteria:

— Desirable properties are defined for each of the three stages.

— The properties are related to topics and eventually to detailed methods and techniques addressing those topics.

— Acceptance criteria are identified from the set of the detailed methods and techniques. As shown in Figure 4, the criteria for accepting selected methods and techniques and possibly metrics or thresholds like limits for estimated uncertainties are embedded in an overall acceptance argument. The overall acceptance argument demonstrates that the selected acceptance criteria harmonize with technology-independent risk acceptance criteria such as ALARP, MEM, GAMAB, or PRB. The overall argument is, for example, represented in an assurance case.

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

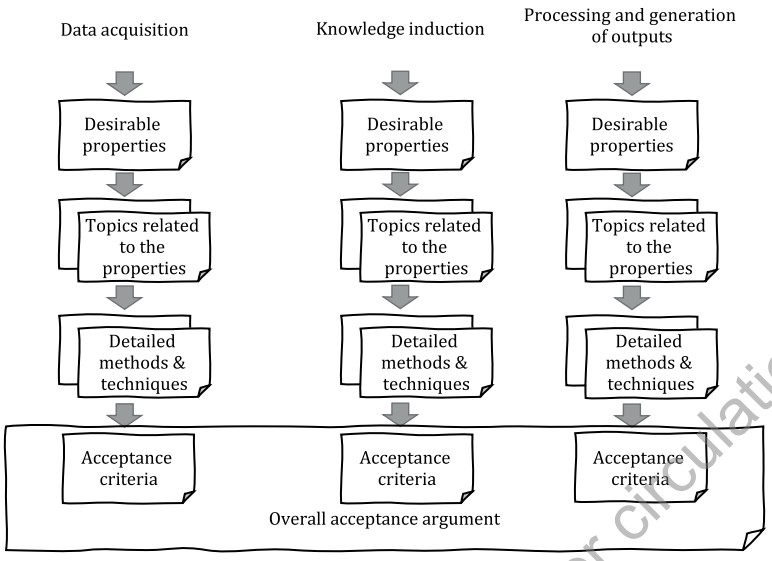


Figure 4 — Processes in each stage

NOTE 1 The properties are defined on a case-by-case basis or derived from properties listed in existing International Standards, based on the level of the elements containing AI technology. Refer to Clause 8 for the list of considered properties.

NOTE 2 The "Desirable properties", "Topics related to the properties" and "Detailed methods & techniques" are specific to each stage or common, depending on the specific application.

NOTE 3 In this context, the acceptance criteria are intended as something that is identified and confirmed during development.

8 Properties and related risk factors of AI systems

8.1 Overview

8.1.1 General

Clause 7 describes how the definition of desirable properties is the first step of the three-stage realization principle. The properties are related to topics and eventually to detailed methods and techniques addressing those topics. Acceptance criteria are then identified from the set of the detailed methods and techniques.

This clause provides a literature review on the properties that characterize systems using AI technology and their related risk factors. Such properties and risk factors include degree of automation and control (see 8.2), degree of decision transparency and explainability (see 8.3), environmental complexity and vagueness of their defining specifications (see 8.4), resilience to adversarial inputs (see 8.5), system hardware considerations (see 8.6) and technological maturity (see 8.7).

Details of the properties and risk factors of systems using AI technology, and their related topics and challenges, are discussed in this clause.

8.1.2 Algorithms and models

On a technological level, the capability of AI is often achieved by the combination of an AI model and parameters of the model. The parameters, which are generated by an AI algorithm during training, typically represents information that achieves the application's purpose, (e.g. knowledge about how various inputs are distinguished and recognized), while the model computes information from parameters and inputs, (e.g. to make a prediction). This means the functional safety of applications using AI technology depends on both.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Example types of AI models include linear functions, logical calculi, dynamic Bayesian networks and artificial neural networks (ANN). The parameters of the models are either handcrafted by an engineer, or are synthesized from data by machine learning algorithms that themselves use a systematic analysis process. The models are usually implemented as an executable representation, such as machine code (in the case of software), or special hardware, such as field programmable gate arrays (FPGAs).

Usually, models itself without parameters contain only a limited amount of knowledge or implications about the application's goals. This is quite similar to the role of foundational software libraries or programming environments (e.g. compilers) in non-AI software. As described by the IEC 61508 series[16]-[19], achieving functional safety includes demonstrating the correct functioning of the model. In this way, the integrity of model implementations in AI technology is often handled with existing principles of functional safety as specified in the IEC 61508 series[16]-[19], similar to that of non-AI software components. The same holds for the logic involved in the translation of the models and the parameters.

By contrast, model parameters often contain knowledge related to the objective of the systems involving functional safety. There are several different ways of constructing parameters and different approaches are used for assessing the completion of risk reduction measures to ensure functional safety.

For example, when model parameters are created manually by engineers, the models likely reflect the engineers' knowledge about the application, which is assessed during the management processes used within functional safety life cycles. In these cases, the life cycle of existing functional safety International Standards is followed (AI technology Class I as described in 6.2). It is often feasible to create parameters manually for simple models such as simple linear functions or logical calculi.

In some cases, parameters derived from data by machine learning algorithms are analysed and verified after their creation. Alternatively, parameters derived by machine learning algorithms are analysed, the underlying parameters extracted and used to extend general engineering knowledge, that, in turn, are used to develop further AI technologies. With the application of validated engineering knowledge, the life cycle of existing functional safety International Standards are applied (e.g. treating these models as AI technology Class I as described in 6.2).

In other cases, model parameters derived from data by machine learning algorithms are too complex to be understood, analysed and verified. This is often the case for complex types of models, such as neural networks, because representations of models in these types do not necessarily reflect human understanding or reasoning. The use of different approaches for assessing the risk reduction for functional safety is appropriate in these cases, which constitutes a major challenge for the use of AI technology in implementation of functional safety systems.

8.2 Level of automation and control

The level of automation (sometimes called "levels of autonomy" in the literature) describes the extent to which an AI system functions independently of human supervision and control. It thus determines not only how much information about the behaviour of the system is available to the operator, but also defines the control and intervention options of the human.

Dimensions of this topic include how high the level of automation is for the respective application, as well as the extent to which the user's control options are restricted. Systems using AI technology with a high-level of automation can exhibit unexpected behaviour that can be difficult to detect and control. Highly automated systems can thus pose risks in terms of their performance characteristics such as reliability and safety.

Several aspects are relevant in considering whether functional safety is achieved, such as the responsiveness of the AI system and the presence or absence of a "supervisor" (as discussed in 10.2.3). In this context, a "supervisor" serves to validate or approve automated decisions of the system. Such a "supervisor" is achieved by technical control functions, though in some situations such a supervisory function is not feasible, e.g. highly complex decisions or ML systems that have learned new behaviours. For example, a second safety instrumented system for critical controls (Usage Level C or D as described in 6.2) is added and assigned to a safety function for redundant components, as in functional safety International Standards like IEC 61508-1[16]. See 10.2.3 for further details.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

Another way for adding a "supervisor" is to use a human whose task it is to intervene in critical situations or to acknowledge system decisions. One known way this is addressed is by adding a system (at Usage Level C or D as described in 6.2) to aid the human supervisor by detecting possible outcomes of the decision. An example is a simulation system that gives "what if" information for different decisions and checks for outcomes. However, even if humans are in the loop and oversee the actions of a system, sometimes this does not reduce such risks to a suitable level and can introduce additional risks.

EXAMPLE A human driver is not aware of a "black ice" road condition and takes control of a vehicle because they don't understand why the autonomous system is driving so carefully.

Furthermore, the adaptability of the AI system is a consideration, and in particular whether a system changes its own behaviour over time, as is the case in some machine learning systems. These systems adapt to changing environmental conditions (e.g. via feedback loops or an evaluation function) and even acquire entirely new functions over time. A disadvantage of such learning systems, however, is that they can deviate from the initial specification and are difficult to validate. Therefore, such systems are considered very carefully before being used in systems of higher usage levels A-C as described in 6.2.

Table 4 (derived from ISO/IEC 22989:2022, Table 1) introduces several terminology and concepts related to the level of automation.

Table 4 — Relationship among autonomy, heteronomy and automation (derived from ISO/IEC 22989:2022, Table 1)

		Level of automation	Comments
Automated system	Autonomous	Autonomy	The system is capable of modifying its operating domain or its goals without external intervention, control or oversight.
	Heteronomous	Full automation	The system is capable of performing its entire mission without external intervention.
		High automation	The system performs parts of its mission without external intervention.
		Conditional automation	Sustained and specific performance by a system, with an external agent being ready to take over when necessary.
		Partial automation	Some sub-functions of the system are fully automated while the system remains under the control of an external agent.
		Assistance	The system assists an operator.
		No automation	The operator fully controls the system.

NOTE The dividing into levels applies to the control automation functions in any implementation of an automated AI system and taking into account the functions of the components of this system (for example, onboard equipment, floor equipment and control room equipment in railway systems).

8.3 Degree of transparency and explainability

Often, aspects of transparency and explainability are summarized under the term "transparency". Literature described in the following clearly distinguish these terms.

ISO/IEC 22989:2022, 5.15.6 defines explainability as the property of an AI system to express important factors influencing the results of the AI system in a way that is understandable to humans, whereas transparency is defined as the property of a system that appropriate information about the internal processes of an AI system is made available to relevant stakeholders-see also ISO/IEC TR 24028[11].

Information about the model underlying the decision-making process is likely to be relevant. Systems with a low degree of transparency can pose risks in terms of their fairness, security and accountability. Furthermore, such systems can complicate the assessment of their quality. On the other hand, a high degree of transparency can lead to confusion due to information overload, or can conflict with privacy, security, confidentiality requirements and intellectual properties. According to Reference [11], a desirable

Copyrighted document - no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

level of explainability is often achieved without a high-level of transparency. Finding an appropriate level of transparency provides developers with opportunities for error identification and correction, as well as ensures that a user can trust the AI system.

In non-AI software, the intention and knowledge of the engineer is generally encoded into the system using a logical process, making it possible to trace through the code to determine how and why a certain decision has been made by the software. This is done by backtracking through and debugging the software or by reverse engineering the software. By contrast, decisions made by AI models, especially by models of high complexity, or models derived from machine learning algorithms, are more difficult to understand for humans. The way knowledge is encoded in the structure of the model and the way decisions are made, are often different from how humans reason about their own decision-making processes[130], [131].

A high-level of explainability protects against unpredictable behaviour of the system but is sometimes accompanied by lower overall performance in terms of the quality of decisions, due to the limitation of current explainability technology (which limits the amount of information contained in the model to create explanations of reasonable size). Here, a trade-off is often made between explainability and the performance of a system. In addition, the relevance of the information about an AI system's decision-making process is likely an important factor. It is possible that a system provides clear and coherent information about its decision-making process, but that this information is inaccurate or incomplete.

Explainable AI is also used to assist with post-incident analysis when the input data, which are sometimes transient, are recorded and reproduced.

Consequently, transparency and explainability are included in the general evaluation of the AI system. Considerations include:

— whether sufficient information about the system is available;

— whether it is understandable or at least delivers comprehensible information (possibly indirectly) to the intended recipient (the intended recipient of an explanation varies depending on the context. It is, for example, the system developer, first responders of the system in use, or bystanders in some cases);

— whether it delivers correct, complete and reproducible results consistently;

— whether the deliverable is adequate enough to the required confidence.

Several evaluation concepts and strategies exist to judge the transparency and explainability of an AI system, such as those reported in Reference [2] and in References [39] and [40]. Additionally, empirical assessments of the decision-making process of complex models are carried out, for example, by inspecting a convolutional neural network through visualization of components of its internal layers.[41] The goal is to make the network's decision process more explainable by determining how input features affect the model output. Reviewing the output of a convolutional neural network by having its internal state inspected by a human expert is an approach that is extended in related work such as References [42], [43] and [44]. Even when access to internal model states is completely unavailable, approaches such as randomized input sampling for explanation (RISE)[45] still provide insights for certain network types.

Even systems traditionally believed to be reasonably explainable with regards to inspection, (e.g. decision trees), quickly reach a complexity that defies understanding when deployed in real-world applications. In situations where an interpretable result is desired, tools such as optimal classification trees[46] or born-again tree ensembles[47] are applied to reduce complexity and allow for human expert review. See Reference [48] for a further discussion on the relation between AI model types and their interpretability.

Generally speaking, even when fully "explainable AI" is not immediately achievable (such as a Class II AI technology), a methodical and formally documented evaluation of model interpretability is employed in regards to risk, subject to careful consideration of the consequences on functional safety risk that arise from inappropriate decisions. This aids in comparability and model selection and provides insights during an "after the event" failure analysis.

Copyrighted document: no reproduction or circulation
For review by FC on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024

8.4 Issues related to environments

8.4.1 Complexity of the environment and vague specifications

AI systems are often used under environmental conditions whose complexity is difficult for humans to fully analyse and describe. AI technology automatically generates rules, or apply judgement, without reliance on human-generated representations of analytic, detailed and complex environmental conditions. Further, the development life cycle for AI systems has the potential to begin with vague specifications or vague goals. Such specifications are given in the form of either functional or non-functional requirements.

Vagueness of specifications leads to difficulty during assurance of functional safety-related properties. The complexity of the environment only worsens the situation. Even the definition of a tolerable level of functional safety is likely to be undermined by a vague specification, because the definition of ”safety function” depends on the given specification.

Functional safety applications involve some minimal explanation for the functional completeness (as defined in ISO/IEC 25010[148]). Functional completeness is addressed by seeking an extremely detailed specification or by seeking to cover the whole complex environment (e.g. by training data), or by the combination of the two. See 9.3 for additional information.

Another feature of AI systems complexity is that, although their models are often deterministic, their output seem to be probabilistic. For example, given a very complex environment that is represented by a large state space and is subject to constant change and expansion, it is difficult to ensure that a model which generalizes the behaviour well under finite state conditions reacts appropriately to every possible state of the environment.

The effect of operating in complex, not completely defined environments, results in a new type of uncertainty beyond the scope of current functional safety assessments.

The extent to which the adequateness of models for the intended application is considered. Additionally, uncertainty of the model caused by possible incorrect predictions and misclassifications is considered in terms of behaviour planning and functionality.

To address the stochastic concept to address the operational environment, expanding stochastic assumptions generally used in the functional safety discipline for the random failures of the hardware and “proven in use” software is a relatively novel approach that is relevant to AI technologies, see 10.2.5.

For a more comprehensive list of emerging issues, see References [92] and [93].

8.4.2 Issues related to environmental changes

8.4.2.1 Data drift

Data drift is a phenomenon that distribution of runtime input data departs from those used in training phase which causes degradation of performance including safety. Data drift is often tied to an incomplete representation of the input domain during training. This is due, for example, to failure to account for seasonal changes in input data, unforeseen input by operators, or the addition of new sensors that become available as input features.

Components containing AI technology are inspected for sources of data drift in the context of a risk analysis and adequate measures are planned where appropriate.

Some examples of data drift are attributed to failure to apply best practices in model engineering. Common examples include picking inappropriate training data, data whose distribution does not reflect the actual distribution encountered in the application context or omitting important examples in the training data. These problem instances are fixed by improved input data modelling and retraining.

Data drift is caused by external factors, such as seasonal change or a change in process that induces data drift. Examples include replacement of a sensor with a new variant featuring a different bias voltage, or the sensor encountering different lighting conditions in training and previously unseen data. In some cases, the

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

© ISO/IEC 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: ISO_IEC TR 5469 WG_2024