FINAL DRAFT

TECHNICAL REPORT

ISO/IEC DTR 5469

ISO/IEC JTC **1**/SC **42**

Secretariat: **ANSI**

Voting begins on: **2023-10-06**

Voting terminates on: **2023-12-01**

Artificial intelligence — Functional safety and AI systems

Reference number
ISO/IEC DTR 5469:2023(E)

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

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, 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:—[1].

1) Under preparation. Stage at the time of publication: ISO/IEC FDIS 5338:2023.

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]

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]

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)

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 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.

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.

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;