Technical Report

ISO/IEC TR 5469

Artificial intelligence — Functional safety and AI systems

Intelligence artificielle — Sécurité fonctionnelle et systèmes d'intelligence artificielle

First edition
2024-01

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

© ISO/IEC 2024



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2024

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

Published in Switzerland

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 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

© ISO/IEC 2024 – All rights reserved

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

© ISO/IEC 2024 – All rights reserved

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

© ISO/IEC 2024 – All rights reserved

Introduction

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,

© ISO/IEC 2024 – All rights reserved

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.

© ISO/IEC 2024 – All rights reserved

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]

© ISO/IEC 2024 – All rights reserved