PD ISO/PAS 8800:2024

BSI Standards Publication

Road vehicles — Safety and artificial intelligence

Documents
Document Sharing Hub

Full document details: PD ISO/PAS 8800:2024

National foreword

This Published Document is the UK implementation of ISO/PAS 8800:2024.

The UK participation in its preparation was entrusted to Technical Committee AUE/32/1.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

This publication is not to be regarded as a British Standard.

© The British Standards Institution 2024
Published by BSI Standards Limited 2024

ISBN 978 0 539 19489 0

ICS 43.040.10; 43.040.15

Compliance with a Published Document cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 31 December 2024.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: PD ISO/PAS 8800:2024

Publicly Available Specification

ISO/PAS 8800

First edition
2024-12

Road vehicles — Safety and artificial intelligence

Véhicules routiers — Sécurité et intelligence artificielle

Reference number
ISO/PAS 8800:2024(en)

© ISO 2024

Documents
Document Sharing Hub

Full document details: PD ISO/PAS 8800:2024

COPYRIGHT PROTECTED DOCUMENT

© ISO 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: PD ISO/PAS 8800:2024

Contents

Page

Foreword ... **vi**

Introduction ... **vii**

1 Scope ... **1**

2 Normative references ... **1**

3 Terms and definitions ... **2**
3.1 General AI-related definitions ... 2
3.2 Data-related definitions ... 7
3.3 General safety-related definitions ... 9
3.4 Safety: Root cause-, error-and failure-related definitions ... 11
3.5 Miscellaneous definitions ... 12

4 Abbreviated terms ... **14**

5 Requirements for conformity ... **15**
5.1 Purpose ... 15
5.2 General requirements ... 15

6 AI within the context of road vehicles system safety engineering and basic concepts ... **16**
6.1 Application of the ISO 26262 series for the development of AI systems ... 16
6.2 Interactions with encompassing system-level safety activities ... 17
6.3 Mapping of abstraction layers between the ISO 26262 series, ISO/IEC 22989 and this document ... 20
6.4 Example architecture for an AI system ... 22
6.5 Types of AI models ... 23
6.6 AI technologies of a ML model ... 23
6.7 Error concepts, fault models and causal models ... 24
6.7.1 Cause-and-effect chain ... 24
6.7.2 Root cause classes ... 26
6.7.3 Error classification based on the safety impact ... 27

7 AI safety management ... **28**
7.1 Objectives ... 28
7.2 Prerequisites and supporting information ... 28
7.3 General requirements ... 28
7.4 Reference AI safety life cycle ... 31
7.5 Iterative development paradigms for AI systems ... 33
7.6 Work products ... 34

8 Assurance arguments for AI systems ... **35**
8.1 Objectives ... 35
8.2 Prerequisites and supporting information ... 35
8.3 General requirements ... 36
8.4 AI system-specific considerations in assurance arguments ... 36
8.5 Structuring assurance arguments for AI systems ... 37
8.5.1 Context of the assurance argument ... 37
8.5.2 Categories of evidence ... 38
8.6 The role of quantitative targets and qualitative arguments ... 39
8.7 Evaluation of the assurance argument ... 40
8.8 Work products ... 41

9 Derivation of AI safety requirements ... **41**
9.1 Objectives ... 41
9.2 Prerequisites and supporting information ... 42
9.3 General requirements ... 42
9.4 General workflow for deriving safety requirements ... 43
9.5 Deriving AI safety requirements on supervised machine learning ... 46
9.5.1 The need for refined AI safety requirements ... 46

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

9.5.2 Derivation of refined AI safety requirements to manage uncertainty 47
9.5.3 Refinement of the input space definition for AI safety lifecycle 50
9.5.4 Restricting the occurrence of AI output insufficiencies 50
9.5.5 Metrics, measurements and threshold design 54
9.5.6 Considerations for deriving safety requirements 55
9.6 Work products 56

10 Selection of AI technologies, architectural and development measures 56
10.1 Objectives 56
10.2 Prerequisites 56
10.3 General requirements 56
10.4 Architecture and development process design or refinement 57
10.5 Examples of architectural and development measures for AI systems 58
10.6 Work products 62

11 Data-related considerations 62
11.1 Objectives 62
11.2 Prerequisites and supporting information 62
11.3 General requirements 62
11.4 Dataset life cycle 63
11.4.1 Datasets and the AI safety lifecycle 63
11.4.2 Reference dataset lifecycle 64
11.4.3 Dataset safety analysis 65
11.4.4 Dataset requirements development 71
11.4.5 Dataset design 74
11.4.6 Dataset implementation 75
11.4.7 Dataset verification 75
11.4.8 Dataset validation 76
11.4.9 Dataset maintenance 77
11.5 Work products 77

12 Verification and validation of the AI system 78
12.1 Objectives 78
12.2 Prerequisites and supporting information 78
12.3 General requirements 78
12.4 AI/ML specific challenges to verification and validation 80
12.5 Verification and validation of the AI system 81
12.5.1 Scope of verification and validation of the AI system 81
12.5.2 AI component testing 84
12.5.3 Methods for testing the AI component 86
12.5.4 AI system integration and verification 88
12.5.5 Virtual testing vs physical testing 88
12.5.6 Evaluation of the safety-related performance of the AI system 89
12.5.7 AI system safety validation 90
12.6 Work products 91

13 Safety analysis of AI systems 91
13.1 Objectives 91
13.2 Prerequisites and supporting information 92
13.3 General requirements 92
13.4 Safety analysis of the AI system 93
13.4.1 Scope of the AI safety analysis 93
13.4.2 Safety analysis based on the results of testing 95
13.4.3 Safety analysis techniques 95
13.5 Work products 97

14 Measures during operation 97
14.1 Objectives 97
14.2 Prerequisites and supporting information 98
14.3 General requirements 98
14.4 Planning for operation and continuous assurance 99

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

14.4.1 Safety risk of the AI system during operation phase ... 99
14.4.2 Safety activities during the operation phase ... 99
14.5 Continual, periodic re-evaluation of the assurance argument ... 100
14.6 Measures to assure safety of the AI system during operation ... 101
14.6.1 General ... 101
14.6.2 Technical safety measures ... 101
14.6.3 Safe operation guidance and misuse prevention in the field ... 102
14.7 Field data collection ... 103
14.8 Evaluation and continuous development ... 104
14.8.1 Field risk evaluation ... 104
14.8.2 Countermeasures addressing field risk ... 105
14.8.3 AI re-training, re-validation, re-approval and re-deployment ... 105
14.9 Work products ... 106

15 Confidence in use of AI development frameworks and software tools used for AI model development ... 106
15.1 Objectives ... 106
15.2 Prerequisites and supporting information ... 107
15.3 General requirements ... 107
15.4 Confidence in the use of AI development frameworks ... 107
15.5 Confidence in the use of tools used to support the AI-safety lifecycle ... 109
15.6 Principles for data-driven AI model training and evaluation ... 110
15.7 Work products ... 110

Annex A (informative) **Overview and workflow of this document ... 111**

Annex B (informative) **Example assurance argument structure for an AI system ... 116**

Annex C (informative) **ISO 26262 gap analysis for ML ... 130**

Annex D (informative) **Detailed considerations on safety-related properties of AI systems ... 137**

Annex E (informative) **STAMP/STPA example ... 139**

Annex F (informative) **Identification of software units within NN-based systems ... 144**

Annex G (informative) **Architectural and development measures for AI systems ... 147**

Annex H (informative) **Typical performance metrics for machine learning ... 162**

Bibliography ... 167

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes 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 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. ISO 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.

This document was prepared by Technical Committee ISO/TC 22, *Road vehicles*, Subcommittee SC 32, *Electrical and electronic components and general system aspects*.

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.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Introduction

The purpose of this document is to provide industry-specific guidance on the use of AI systems in safety-related functions. It is not restricted to specific AI methods or specific vehicle functions.

This document defines a framework for managing AI safety that tailors or extends existing approaches currently defined in the ISO 26262 series and in ISO 21448.

Functional safety-related risks associated with malfunctioning behaviour of an AI system are addressed by tailoring or extending relevant clauses from ISO 26262-series.

The risks related to functional insufficiencies in the AI system are addressed by extending the concepts and guidance provided by ISO 21448. A causal model for understanding the sources of functional insufficiencies in the AI system is proposed. The model is used to derive a set of safety requirements on the AI system as well as a set of risk reduction measures.

NOTE 1 ISO 21448 is applicable to intended functionalities where proper situational awareness is essential to safety and where such situational awareness is derived from sensors and processing algorithms, especially functionalities of emergency intervention systems and systems with ISO/SAE PAS 22736 levels 1 to 5 for driving automation. It is therefore possible that systems utilize AI technologies that do not fall within the scope of ISO 21448.

EXAMPLE 1 ISO 21448 does not apply to the development of an engine control unit that uses AI to optimize its performance whereas this document does.

This document recognizes that due to the wide range of applications of AI and associated safety requirements, as well as the rapidly evolving state-of-the-art, it is not possible to provide detailed requirements on the process or product characteristics required to achieve an acceptably low level of residual risk associated with the use of AI systems. Therefore, in addition to providing guidance for tailoring or extending the ISO 26262 series and ISO 21448, this document focuses on the principles that support the creation of a project-specific assurance argument for the safety of the AI elements within on-board vehicle systems. This includes proposing risk reduction measures during the design and operation phases using an iterative approach to reducing risk as outlined in ISO/IEC Guide 51.

Hazard analysis and risk analysis are beyond the scope of this document. These are considered a part of the vehicle level systems safety engineering activities described in the ISO 26262 series and ISO 21448, or in application of specific standards such as ISO TS 5083.

ISO/IEC TR 5469 provides generic guidance for the application of AI technologies as part of safety functions, independent of specific industry sectors. Many of the concepts outlined in ISO/IEC TR 5469 can be applied in the context of road vehicles. There is therefore a close relationship to concepts described within this document and ISO/IEC TR 5469.

ISO/IEC TR 5469 provides classification schemes to determine the safety requirements on the AI/ML function. These include the usage level and AI technology class.

The usage level is related to the nature of the task being performed by the engineered AI system.

NOTE 2 The usage levels are described in ISO/IEC TR 5469:2024, 6.2.

The technology class is related to the problem complexity and the transferability of existing standards to demonstrating an adequate level of safety based on properties of the target function and the AI technology used.

NOTE 3 For the technology classes, see ISO/IEC TR 5469:2024, 6.2.

This document does not explicitly call out the classes and usage levels of ISO/IEC TR 5469.

EXAMPLE 2 For some AI technology, the application of ISO 26262 is deemed to be sufficient. This corresponds to Class I of ISO/IEC TR 5469.

The guidance outlined within this document is relevant for all usage of AI for which safety requirements can foreseeably be allocated either through:

a) the use of AI for the functionality itself;

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

b) the use of AI as a safety mechanism.

NOTE 4 These usages correspond to the usage levels A1, A2, C of ISO/IEC TR 5469. In all cases, the applicability of the guidance provided within this document can be determined by the allocation of safety requirements to the AI technology, whereas the usage levels of ISO/IEC TR 5469 can be used to support the requirements elicitation process.

This document is aligned with standards and documents developed by ISO/IEC JTC1/SC42. AI-specific definitions are used from ISO/IEC 22989, unless in conflict with safety-specific definitions.

Other documents developed within ISO/IEC JTC1/SC42 can be used to provide additional guidance on specific aspects of AI that are relevant to safety-related properties. Examples of such documents include ISO/IEC TR 24027 and ISO/IEC TR 24029-1.

This document harmonizes the concepts already described in ISO 21448:2022, Annex D.2 and ISO/TS 5083:20—[1], Annex B whilst extending these with specific guidance regarding the definition of safety requirements of machine learning (ML), ML safety analyses and the creation of associated safety evidence during the development and deployment lifecycle.

ISO/TS 5083:20—, Annex B is an application of this document to automated driving systems (ADS).

The relationship with the above-mentioned documents is summarized in Table 1–1.

Table 1–1 — How this document relates to other publications on AI safety

Publication	Relationship with this document
ISO/IEC 22989	AI-specific definitions are used from ISO/IEC 22989, unless in conflict with safety-specific definitions. Safety-related properties are a subset of generic AI properties described in ISO/IEC 22989.
ISO/IEC TR 5469	This document does not explicitly call out the classes and usage levels of ISO/IEC TR 5469. This document considers and adapts to road vehicles the general framework described in ISO/IEC TR 5469 on safety properties, virtual testing and physical testing, confidence in use of AI development frameworks and architectural redundancy patterns.
ISO 26262	This document is a tailoring or extension of ISO 26262 for AI elements of the system. See Clause 5 for details.
ISO 21448	This document is a tailoring or extension of ISO 21448 for AI elements of the system. See Clause 5 for details.
ISO TS 5083:20—	ISO TS 5083:20—, Annex B is an application of this document to automated driving systems (ADS).

This document adds the following contents with respect to the documents listed in Table 1-1:

— tailoring or extensions of ISO 26262 and ISO 21448 required specifically for AI elements of the system (referred to as AI systems);

— a conceptual model for reasoning about errors and their causes specific to AI systems;

— a reference AI safety lifecycle;

— the safety assurance argument for AI systems;

— a method for deriving AI safety requirements for AI systems;

— considerations for the design of safe AI systems;

— considerations on data management for the AI systems;

1) Under preparation. Stage at the time of publication: ISO/DTS 5083.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

— a verification and validation strategy for AI systems;

— a safety analysis approach for AI systems (focused on insufficiencies);

— activities during operation required to ensure the continuous AI safety.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Documents
Document Sharing Hub

Road vehicles — Safety and artificial intelligence

1 Scope

This document applies to safety-related systems that include one or more electrical and/or electronic (E/E) systems that use AI technology and that is installed in series production road vehicles, excluding mopeds. It does not address unique E/E systems in special vehicles, such as E/E systems designed for drivers with disabilities.

This document addresses the risk of undesired safety-related behaviour at the vehicle level due to output insufficiencies, systematic errors and random hardware errors of AI elements within the vehicle. This includes interactions with AI elements that are not part of the vehicle itself but that can have a direct or indirect impact on vehicle safety.

EXAMPLE 1 Examples of AI elements within the vehicle include the trained AI model and AI system.

EXAMPLE 2 Direct impact on safety can be due to object detection by elements external to the vehicle.

EXAMPLE 3 Indirect impact on safety can be due to field monitoring by elements external to the vehicle.

The development of AI elements that are not part of the vehicle is not within the scope of this document. These elements can conform to domain-specific safety guidance. This document can be used as a reference where such domain-specific guidance does not exist.

This document describes safety-related properties of AI systems that can be used to construct a convincing safety assurance claim for the absence of unreasonable risk.

This document does not provide specific guidelines for software tools that use AI methods.

This document focuses primarily on a subclass of AI methods defined as machine learning (ML). Although it covers the principles of established and well-understood classes of ML, it does not focus on the details of any specific AI methods e.g. deep neural networks.

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 21448:2022, *Road vehicles — Safety of the intended functionality*

ISO 26262-1:2018, *Road vehicles — Functional safety — Part 1: Vocabulary*

ISO 26262-2:2018, *Road vehicles — Functional safety — Part 2: Management of functional safety*

ISO 26262-6:2018, *Road vehicles — Functional safety — Part 6: Product development at the software level*

ISO 26262-8:2018, *Road vehicles — Functional safety — Part 8: Supporting processes*

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 26262-1, ISO 21448, ISO/IEC 22989 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 General AI-related definitions

3.1.1
AI component
element of an *AI system* (3.1.17)

EXAMPLE 1 An *AI pre-processing* (3.1.11) component.

EXAMPLE 2 An *AI post-processing* (3.1.9) component.

EXAMPLE 3 An *AI model* (3.1.7).

EXAMPLE 4 A conventional software component inside an AI system.

Note 1 to entry: AI components that are not AI models or that do not contain AI models are not developed according to this document. The integration of these components with AI components that are AI models or that contain AI models is performed according to this document.

Note 2 to entry: See 6.3 for an elaboration of the relationship of the different abstraction layers of the ISO 26262 series, ISO/IEC 22989 and this document with each other.

[SOURCE: ISO/IEC 22989:2022, 3.1.2, modified to be consistent with ISO 26262-1 definitions —"Functional element" was replaced with "element", reworded to not use “construct”, examples and Notes to entry were added.]

3.1.2
AI controllability
ability of an external agent to control the *AI element* (3.1.3), its output or the behaviour of the item influenced by the AI output in order to prevent harm

EXAMPLE Before setting a pulse-width modulation (PWM) signal of an actor determined by an *AI model* (3.1.7), the PWM output is limited by a simple threshold or the consumer substitutes the PWM signal with an approximate physical model.

Note 1 to entry: An external agent is a person or an element not belonging to the *AI system* (3.1.17).

3.1.3
AI element
AI component (3.1.1) or *AI system* (3.1.17)

Note 1 to entry: An AI element can refer to a subset of *components* (3.5.2) within an AI system that provide related functionality.

Note 2 to entry: See 6.3 for an elaboration of the relationship of the different abstraction layers of the ISO 26262 series ISO/IEC 22989 and this document with each other.

3.1.4
AI explainability
property of an *AI system* (3.1.17) to express important factors influencing the AI system's outputs in a way that humans can understand

EXAMPLE The AI system can be explainable by natural language or by visualizing feature attribution methods like gradient-based heat/saliency maps.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.1.5
AI generalization
ability of an *AI model* (3.1.7) to adapt and perform well on previously unseen data during inference

3.1.6
AI method
type of *AI model* (3.1.7)

EXAMPLE 1 Deep neural network.

EXAMPLE 2 K-nearest neighbour.

EXAMPLE 3 Support vector machine.

3.1.7
AI model
construct containing logical operations, arithmetical operations or a combination of both to generate an inference or prediction based on input data or information without being completely defined by human knowledge

Note 1 to entry: Inference is using a model to understand the relation between predictors and a target. Prediction is using a model to generate a prediction (values close to the real seen or unseen targets) based on the inputs.

3.1.8
AI model validation
evaluation of the performance of different *AI model* (3.1.7) candidates through testing

Note 1 to entry: There are three terms, "AI model validation", "validation" and "safety validation", that are distinguished in this document. AI model validation originates from the validation data used by the AI community, validation originates from classic system development and safety validation originates from the ISO 26262 series.

Note 2 to entry: The AI model validation is executed using the AI validation dataset.

3.1.9
AI post-processing
any processing that is applied to the output of an *AI model* (3.1.7) for the purpose of mapping the raw output/s to a more contextually relevant and consumable format

EXAMPLE 1 A non-maximum suppression and thresholding for a bounding-box generation that serves to remove bounding boxes of low relevance and duplicates.

EXAMPLE 2 The outputs of a mixture density network are combined with a physical model (a hybrid model).

Note 1 to entry: AI post-processing also includes any data conversion that is used to bring the output into a common format for better comparability.

Note 2 to entry: AI post-processing can have a positive or a negative impact on the safety-related properties of the output of the *AI system* (3.1.17).

3.1.10
AI predictability
ability of the *AI system* (3.1.17) to produce trusted predictions

Note 1 to entry: Trusted predictions means that the predications are accurate and that this claim is supported by statistical evidence.

3.1.11
AI pre-processing
any processing that is applied to the input of an *AI model* (3.1.7)

3.1.12
AI reliability
ability of the *AI element* (3.1.3) to perform the *AI task* (3.1.18) without *AI error* (3.4.1) under stated conditions and for a specified period of time

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.1.13
AI resilience
ability of the *AI element* (3.1.3) to recover and continue performing the *AI task* (3.1.18) after the occurrence of an *AI error* (3.4.1).

3.1.14
AI robustness
ability to maintain an acceptable level of performance under the presence of semantically insignificant but reasonably expected changes to the input

EXAMPLE In image data these insignificant input changes can stem from naturally-induced image corruptions or sensor noise.

3.1.15
AI safety
absence of unreasonable *risk* (3.3.10) due to *AI errors* (3.4.1) caused by faults and functional insufficiencies

Note 1 to entry: This definition only applies in the context of this document. The term "AI safety" is commonly understood to have a broader meaning which includes ethics, value alignment, long-term considerations, etc.

3.1.16
AI safety requirement
safety requirement (3.3.14) of an *AI element* (3.1.3)

3.1.17
AI system
item or element that utilises one or more *AI models* (3.1.7)

EXAMPLE An AI system consisting of the *AI component* (3.1.1) “deep neural network for bounding box generation (AI model)” and of the AI component “non-maximum suppression algorithm (*AI post-processing* (3.1.9) AI component)”.

Note 1 to entry: The AI system can use various *AI methods* (3.1.6) and can utilize different *AI technologies* (3.1.19).

Note 2 to entry: The boundaries of the AI system are determined during the definition of AI system architecture.

Note 3 to entry: The AI system can contain one or more AI components.

Note 4 to entry: The term “AI system” serves in this document as the top level of abstraction of the content to be developed in conformity to the corresponding standard. As such it is possible in a distributed development that what one party considers to be an AI component, the other party considers to be an AI system, as for the latter it represents the top level of the content they develop.

Note 5 to entry: See 6.3 for an elaboration of the relationship of the different abstraction layers of the ISO 26262 series, ISO/IEC 22989 and this document with each other.

3.1.18
AI task
action required by the *AI element* (3.1.3) to achieve a specific goal

Note 1 to entry: Examples of AI tasks include classification, regression, ranking, clustering and dimensionality reduction.

Note 2 to entry: The AI task can be seen as a semantic description of the *AI model* (3.1.7).

[SOURCE: ISO/IEC 22989:2022, 3.1.35, modified — "task" has been replaced with "AI task", "by the AI element" has been added, "<artificial intelligence>" has been removed; and the Notes to entries have been modified.]

3.1.19
AI technology
any technology used within the lifecycle of an *AI system* (3.1.17) to design, develop, train, test, validate and implement the *AI model* (3.1.7)

EXAMPLE Examples of AI technologies are provided in 6.6

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.1.20
AI testing
testing the *AI system* (3.1.17) or *AI model* (3.1.7) to estimate the expected performance and generalization capability in the field

Note 1 to entry: The AI testing is executed using an AI test dataset.

Note 2 to entry: See also ISO 26262-1:2018, 3.169.

3.1.21
AI system safety validation
confirmation that an *AI safety requirement* (3.1.16) allocated to the *AI system* (3.1.17) is fulfilled

Note 1 to entry: In other standards, validation indicates that requirements are suitable for the intended use. In this document, the term is intentionally used in a different way that is common in the ML community, i.e. to verify the implementation of the requirement.

3.1.22
bias
undesired, systematic difference in the *AI systems* (3.1.17) predictions with respect to particular classes of inputs in comparison with others due to potential incorrect learning process

EXAMPLE The classes of inputs can refer to images of objects and people in the context of computer vision.

Note 1 to entry: Bias can arise from an undesired systematic difference within the dataset, from limitations within the training process, or from limitations within the *AI model* (3.1.7) capability itself to accurately reflect the dataset.

[SOURCE: ISO/IEC 22989:2022, 3.5.4, modified —the definition was adapted to the AI context, Note 1 to entry was replaced, the EXAMPLE was added.]

3.1.23
control element
element (3.5.4) controlling the execution of the *AI task* (3.1.18) by the *AI element* (3.1.3) and other AI element-related operations like updates

Note 1 to entry: The control element can control non-AI elements as well.

3.1.24
data pre-processing
part of the AI workflow that transforms raw data so they are usable as the input to create the *AI model* (3.1.7)

Note 1 to entry: Pre-processing can include reformatting, removal of outliers and duplicates, and ensuring the completeness of the dataset.

3.1.25
encompassing system
item which contains the *AI system* (3.1.17)

3.1.26
ground truth
set of dataset annotations that are taken to be correct

Note 1 to entry: Individual annotations are derived from information external to the dataset.

Note 2 to entry: Individual annotations may be refined as new information becomes available.

[SOURCE: ISO/IEC 2382-37:2022, 37.09.34]

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.1.27
hyperparameter
parameters of the used *AI technologies* (3.1.19) that affect both the performance of the *AI model* (3.1.7) and its learning process

Note 1 to entry: Hyperparameters are selected prior to training and can be used to help estimate *model parameters* (3.1.35).

Note 2 to entry: Examples of hyperparameters include the number of network layers, the width of each layer, the type of activation function, the optimization method, the learning rate for neural networks, the choice of kernel function in a support vector machine, the number of leaves or the depth of a tree, the number of clusters in K-means clustering, the maximum number of iterations of the expectation maximization algorithm and the number of Gaussians in a Gaussian mixture.

[SOURCE: ISO/IEC 22989:2022, 3.3.4, modified — The term has been redefined to be applicable to all kinds of AI methods, not only machine learning.]

3.1.28
input space
set of possible input values

Note 1 to entry: See *semantic input space* (3.1.34) and *syntactic input space* (3.1.36) for ways an input space can be specified.

3.1.29
machine learning
ML
process of optimizing model parameters through computational techniques, such that the model's behaviour aligns with data or experience and enables prediction beyond the training set

EXAMPLE Learning from experience can mean trying to represent non-static data like simulation, reinforcement learning environment, etc.

[SOURCE: ISO/IEC 22989:2022, 3.3.5, modified — "Reflects the data or experience" was replaced with "aligns with data or experience and enables prediction beyond the training set", the EXAMPLE was added.]

3.1.30
ML algorithm
algorithm to optimize parameters of a *ML model* (3.1.31) from data according to given criteria

EXAMPLE Consider solving a univariate linear function $y = \theta_0 + \theta_1 x$ where y is an output or result, x is an input, θ_0 is an intercept (the value of y where x=0) and θ_1 is a weight. In ML, the process of determining the intercept and weights for a linear function is known as linear regression.

[SOURCE: ISO/IEC 22989:2022, 3.3.6, modified — "determine" was replaced with "optimize".]

3.1.31
ML model
mathematical construct that generates an inference or prediction based on input data or information and comprises a functionality that is created by *machine learning* (3.1.29)

EXAMPLE If a univariate linear function ($y = \theta_0 + \theta_1 x$) has been trained using linear regression, the resulting model can be $y = 3 + 7x$.

Note 1 to entry: A ML model results from training based on a *ML algorithm* (3.1.30).

[SOURCE: ISO/IEC 22989:2022, 3.3.7, modified — "And comprises a functionality that is created by *machine learning*" was added to distinguish from other mathematical constructs.]

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.1.32
ML model training
iterative process to optimize a *ML model's* (3.1.31) input and output behaviour on a given training dataset with the intention to improve its quality (e.g. AI accuracy, *AI robustness* (3.1.14), generalization capability, run time), based on a *ML algorithm* (3.1.30) that can adapt ML model parameters, *hyperparameters* (3.1.27), cost function or the model structures itself

[SOURCE: ISO/IEC 22989:2022, 3.3.15, modified to elaborate the procedure and intention.]

3.1.33
safety-related AI element
AI element (3.1.3) that contributes to the achievement of an *AI safety requirement* (3.1.16) allocated to the AI system or can contribute to the violation of an *AI safety requirement* allocated to the AI system or can contribute to both

3.1.34
semantic input space
set of possible input values on a semantic level

EXAMPLE The semantic input space acquired by a camera sensor can be described as consisting of street images containing lane markers of different colours, orientation and degradations that appear in different lighting and weather conditions.

Note 1 to entry: The semantic values correspond and conform to abstract semantic concepts expected within the input space.

3.1.35
semantic output space
set of possible output values on a semantic level

3.1.36
syntactic input space
set of possible input values on a syntactic level

EXAMPLE The syntactic input space acquired by a camera sensor can be described as an RGB image array of integers.

Note 1 to entry: The syntactic values can correspond and conform to the output values from a low-level sensor.

3.1.37
syntactic output space
set of possible output values on a syntactic level

3.1.38
trained ML model
ML model (3.1.31) with a set of model parameters as result of *ML model training* (3.1.32)

[SOURCE: ISO/IEC 22989:2022, 3.3.14, modified — "ML model with a set of model parameters as" added as part of the term.]

3.2 Data-related definitions

3.2.1
AI test dataset
dataset used to estimate the performance and generalization capability of an AI model or an AI system

Note 1 to entry: See Clause 11 for more details.

3.2.2
AI validation dataset
dataset used to compare the performance of different candidate *AI models* (3.1.7)

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.2.3
dataset insufficiency
insufficiency of the dataset regarding data-related safety properties under consideration

Note 1 to entry: Dataset insufficiency includes data integrity errors and data distribution errors.

3.2.4
field monitoring dataset
dataset collected after the release of the *AI system* (3.1.17) while the product is in operation and used specifically for field monitoring of the performance of the AI system

3.2.5
hybrid dataset
dataset comprising data elements that are both real-world data elements and synthetic data elements

3.2.6
in distribution data
data whose features relevant to the *AI task* (3.1.18) are present and sufficiently well represented in the training dataset

Note 1 to entry: In distribution, input does not guarantee correctness of AI model output.

3.2.7
metadata
data that provides additional information about the data element or dataset but is usually not directly involved in the training process

Note 1 to entry: Some metadata (e.g. ground truth) is also used for training.

3.2.8
out of distribution data
data containing features relevant for the *AI task* (3.1.18), either absent or not sufficiently well represented in the *training dataset* (3.2.12), that can result in an *AI error* (3.4.1)

Note 1 to entry: Out of distribution (OOD) refers to data or inputs that fall outside the scope of what an AI or ML model was trained on or is designed to handle. When an AI system encounters OOD data, it can struggle to make accurate predictions or decisions because it lacks the necessary knowledge and experience to handle such inputs effectively. OOD data can lead to unexpected or unreliable model behaviour.

3.2.9
real-world dataset
dataset comprising data elements that have been created by real world acquisitions

3.2.10
safety-related KPI
key performance indicator relevant for the achievement of *AI safety* (3.1.15)

3.2.11
synthetic dataset
dataset comprising data elements that have been created artificially

Note 1 to entry: "Created artificially" implies that the data was not directly collected from something that happened in the real world. Additionally, the data does not necessarily represent something that already happened in the real world.

3.2.12
training dataset
dataset used to train an *ML model* (3.1.31)

[SOURCE: ISO/IEC 22989:2022, 3.3.16, modified — definition reworked to contain the term "dataset".]

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.3 General safety-related definitions

3.3.1
assurance
grounds for justified confidence that a claim has been or will be achieved

[SOURCE: ISO/IEC/IEEE 15026-1:2019]

3.3.2
assurance argument
reasoned, auditable artefact created supporting the contention that its top-level claim (or set of claims) is satisfied, including systematic arguments, its underlying evidence and explicit assumptions that support the claim(s)

Note 1 to entry: An assurance argument contains the following and their relationships:

— one or more claims about properties;

— arguments that logically link the evidence and any assumptions to the claim(s);

— a body of evidence and possible assumptions supporting these arguments for the claim(s);

— justification of the choice of top-level claim and the method of reasoning.

[SOURCE: ISO/IEC/IEEE 15026-1:2019, modified — "argumentation" replaced with "arguments".]

3.3.3
claim
true-false statement about the limitations on the values of an unambiguously defined property — called the claim's property — and limitations on the uncertainty of the property's values falling within these limitations during the claim's duration of applicability under stated conditions

Note 1 to entry: Uncertainties may also be associated with the duration of applicability and the stated conditions.

Note 2 to entry: A claim can contain the following:

— property of the system-of-interest;

— limitations on the value of the property associated with the claim (e.g. on its range);

— limitations on the uncertainty of the property value meeting its limitations;

— limitations on the duration of the claim's applicability;

— duration-related uncertainty;

— limitations on conditions associated with the claim;

— condition-related uncertainty.

Note 3 to entry: The term "limitations" is used to fit the many situations that can exist. Values can be a single value or multiple single values, a range of values or multiple ranges of values, and can be multi-dimensional. The boundaries of these limitations are sometimes not sharp, e.g. they can involve probability distributions and can be incremental.

[SOURCE: ISO/IEC/IEEE 15026-1:2019]

3.3.4
undesired safety-related behaviour at the vehicle level
hazardous behaviour, *RFIM prevention issue* (3.3.9) or malfunctioning behaviour that can cause a hazard

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.3.5
hazard
potential source of harm

[SOURCE: ISO 26262-1:2018, 3.75, modified — "caused by malfunctioning behaviour of the item" and Note 1 to entry deleted.]

3.3.6
influencing factor
factor contributing to the achievement or the absence of a *safety-related property* (3.3.13)

3.3.7
misuse
usage in a way not intended by the manufacturer or the service provider

[SOURCE: ISO 21448:2022, 3.17 modified — The Note to entry and EXAMPLEs were deleted.]

3.3.8
reasonably foreseeable indirect misuse
RFIM
reasonably foreseeable misuse which leads to a reduced controllability of the hazardous behaviour, to a potentially increased severity of an occurring accident or a combination of both

[SOURCE: ISO 21448:2022, 3.17 modified — The term was taken from Note 5 to entry of 3.17 and is now explicitly defined.]

3.3.9
RFIM prevention issue
inability to prevent or detect and mitigate a *reasonably foreseeable indirect misuse* (3.3.8)

3.3.10
risk
combination of the probability of occurrence of harm and the severity of that harm

Note 1 to entry: Other forms of risk definitions exist, e.g. risk for other topics like the risk of a project to fail, etc. This document focuses on the risk regarding safety. Hence this definition was chosen.

Note 2 to entry: The resulting risk evaluation of an error of an AI component is typically equivalent to the evaluation of the potential to lead to a violation of a safety requirement allocated to the AI system. The evaluation can be quantitative as well as qualitative, depending on the safety requirement.

[SOURCE: ISO 26262-1:2018, 3.128, modified — Notes 1 and 2 to entry were added.]

3.3.11
AI safety measure
activity or technical solution to avoid, detect or control *AI errors* (3.4.1), to mitigate their harmful effects or a combination thereof

EXAMPLE AI safety analysis.

Note 1 to entry: Safety measures include architectural measures.

Note 2 to entry: The AI safety measures include safety measures of *AI elements* (3.1.3) as defined in the ISO 26262 series as well as measures to address functional insufficiencies in conformity to ISO 21448 (e.g. functional modifications addressing SOTIF-related risks).

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.3.12
safety validation
assurance, based on examination and tests, that the safety goals are adequate and have been achieved with a sufficient level of integrity

Note 1 to entry: There are three terms, "AI model validation", "validation" and "safety validation", that are distinguished in this document. AI model validation originates from the validation data used by the AI community, validation originates from classic system development and safety validation originates from the ISO 26262 series. The three validation meanings are not the same.

[SOURCE: ISO 26262-1:2018, modified — Note 1 to entry was replaced.]

3.3.13
safety-related property
property impacting safety

3.3.14
safety requirement
requirement related to safety

EXAMPLE 1 SOTIF requirement.

EXAMPLE 2 Functional safety requirement.

EXAMPLE 3 Technical safety requirement.

Note 1 to entry: This includes, but is not limited to, safety requirements motivated by functional safety as well as SOTIF.

3.3.15
work product
work product of the safety lifecycle that can be used as evidence within a safety assurance argument

3.4 Safety: Root cause-, error-and failure-related definitions

3.4.1
AI error
one or more discrepancies between computed, observed or measured values or conditions of the *AI element* (3.1.3) and the true, specified or theoretically correct values or conditions of the AI element

Note 1 to entry: An AI error can be a single discrepancy or a sequence of discrepancies.

Note 2 to entry: An AI error can be an error caused by a fault. Faults are typically addressed by the ISO 26262 series.

Note 3 to entry: An AI error can be an output insufficiency caused by a functional insufficiency.

3.4.2
AI triggering condition
specific conditions of a scenario that serve as an initiator for a subsequent *AI error* (3.4.1)

Note 1 to entry: Functional insufficiencies or faults are themselves not AI triggering conditions but are potentially activated by them thus leading to the occurrence of an AI error.

3.4.3
contributing AI error
AI error (3.4.1) which can lead to a violation of an *AI safety requirement* (3.1.16) allocated to the *AI system* (3.1.17), either by itself or in combination with one or more other AI errors

3.4.4
AI error rate
probability density of *AI error* (3.4.1) occurrence divided by probability of no AI error occurring until the measuring point

Note 1 to entry: Measurement units can include errors per h, errors per km, etc.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Note 2 to entry: This is an analogue definition to the failure rate.

3.4.5
functional insufficiency
insufficiency of specification (3.4.6) or performance insufficiency

Note 1 to entry: A functional insufficiency activated by a triggering condition leads, per definition, to either an output insufficiency, a hazardous behaviour, a RFIM prevention issue or a combination of these.

[SOURCE: ISO 21448:2022, 3.8, modified – The EXAMPLEs, Figures and Notes to entry have been removed. Note to entry 1 has been added.]

3.4.6
insufficiency of specification
specification, possibly incomplete, contributing to either a hazardous behaviour or an *RFIM prevention issue* (3.3.9) when activated by one or more triggering conditions

Note 1 to entry: An insufficiency of specification activated by a triggering condition leads per definition to either an output insufficiency, a hazardous behaviour, an RFIM prevention issue or a combination of these.

Note 2 to entry: More details can be found in 6.7.1.

[SOURCE: ISO 21448:2022, 3.12, modified – The EXAMPLEs and Notes to entry have been removed. A new Note to entry has been added for clarification.]

3.4.7
output insufficiency
incorrect output of an element as a result of a triggering condition activating a *functional insufficiency* (3.4.5) of the element, contributing to either a hazardous behaviour, a *RFIM prevention issue* (3.3.9) or both

[SOURCE: ISO 21448:2022, 3.8, modified – The term was taken from Note 6 to entry and explicitly defined.]

3.4.8
safety-related AI error
AI error (3.4.1) of a *safety-related AI element* (3.1.33)

3.4.9
safety-related fault
fault of a safety-related *AI element* (3.1.3)

3.4.10
systematic error
error due to a systematic fault

3.5 Miscellaneous definitions

3.5.1
architectural measure
technical solution implemented by the *AI element* (3.1.3) to detect and mitigate or tolerate *AI errors* (3.4.1) in order to uphold the ability to execute the *AI task* (3.1.18) in a safe manner or to achieve or maintain a dedicated operating mode in case of AI errors without unreasonable *risk* (3.3.10)

EXAMPLE 1 Addition of output layers in the AI model for classification. AI models can make incorrect predictions that can lead to hazardous behaviour. Therefore, it would be beneficial for a model to be cautious in situations where it is uncertain about its predictions. One way to accomplish this is to design AI models by adding output layer(s) to represent reject class or reject option. Such models assess their confidence in each prediction and have the option to abstain from making a prediction when they are likely to make incorrect predictions.

EXAMPLE 2 Addition of redundant *AI components* (3.1.1).

Note 1 to entry: Architectural measures have a tangible impact on the *AI system* (3.1.17) or AI component and can enhance or modify the architecture of AI system or AI component.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Note 2 to entry: If an architectural artefact is used during development (e.g. using a saliency map to argue the explainability of the system) but is removed for system deployment, this is not considered an architectural measure.

3.5.2
component
non-system level element that is logically or technically separable and is comprised of more than one hardware part or one or more software units or a combination of hardware part(s) and software unit(s)

EXAMPLE A microcontroller.

Note 1 to entry: A component is a part of a system.

[SOURCE: ISO 26262-1:2018, modified — "Or a combination of hardware part(s) and software unit(s)" was added.]

3.5.3
development measure
appropriate process step(s) (activity) for the development of an *AI system* (3.1.17) or an *AI component* (3.1.1) that facilitates fulfilling *AI safety requirements* (3.1.16) and/or enhancing the AI properties

Note 1 to entry: When analysing an AI system or AI component, specific activity used during or after the training of the AI component can be a development measure. See 10.4 for more details.

3.5.4
element
system, components (system, hardware or software), hardware parts or software units

[SOURCE: ISO 26262-1:2018, modified — "System" added to the components and both Note 1 and Note 2 to entry removed.]

3.5.5
item
system or combination of systems, to which ISO 26262 is applied, that implements a function or part of a function at the vehicle level

[SOURCE: ISO 26262-1:2018, modified — Note 1 to entry removed.]

3.5.6
off-board
property indicating that a given task is external to the vehicle system

3.5.7
on-board
property indicating that a given task is internal to the vehicle system

3.5.8
testing
process of planning, preparing and operating or exercising an item or element to verify that it satisfies specified requirements, to detect safety anomalies, to validate that requirements are suitable in the given context and to create confidence in its behaviour

Note 1 to entry: "To create confidence in its behaviour" also includes evaluating the performance of the element or item.

[SOURCE: ISO 26262-1:2018, modified — Note 1 to entry was added]

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

3.5.9
validation
confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled

Note 1 to entry: There are three terms, "AI model validation", "validation" and "safety validation", that are distinguished in this document. AI model validation originates from the validation data used by the AI community, validation originates from classic system development and safety validation originates from ISO 26262. The three validation meanings are not the same.

[SOURCE: ISO/IEC 22989:2022, 3.5.18, modified — Note 1 to entry was added]

3.5.10
verification
confirmation, through the provision of objective evidence, that specified requirements have been fulfilled

EXAMPLE The typical verification activities can be classified as follows:

— verification review, walk-through, inspection;

— verification testing;

— simulation;

— prototyping;

— analysis (safety analysis, control flow analysis, data flow analysis, etc.)

Note 1 to entry: Verification only provides assurance that a product conforms to its specification.

[SOURCE: ISO/IEC 22989:2022, 3.5.18, modified — EXAMPLE was added]

4 Abbreviated terms

ACP assurance claim point

ADS automated driving system

AI artificial intelligence

ASIL automotive safety integrity level

DFA dependent failure analysis

DLC dataset lifecycle

DNN deep neural network

E/E electrical/electronic

FMEA failure mode and effects analysis

FN false negative

FP false positive

FPS frames per second

GSN goal structuring notation

HARA hazard analysis and risk assessment

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

HAZOP	hazard and operability study
HMI	human machine interface
KPI	key performance indicator
ID	in distribution
ML	machine learning
NN	neural network
ODD	operational design domain
OOD	out of distribution
OTA	over the air
PFD	probability of failure on demand
PWM	pulse width modulation
RFDM	reasonably foreseeable direct misuse
RFIM	reasonably foreseeable indirect misuse
SOTIF	safety of the intended functionality
TN	true negative
TP	true positive

5 Requirements for conformity

5.1 Purpose

This clause describes how to:

a) achieve conformity to this document;

b) interpret the applicability of each clause;

c) interpret the tables and figures used in this document.

5.2 General requirements

When claiming conformity to this document, each requirement shall be met, unless one of the following applies:

a) tailoring of the safety activities as defined in Clause 6 or in accordance with ISO 26262-2 has been performed that shows that the requirement does not apply;

b) a rationale is available that the non-conformity is acceptable, and the rationale has been evaluated in accordance with this document and ISO 26262-2, when applicable.

The results of safety activities are given as work products. “Prerequisites” are information which shall be available as work products of a previous phase or from an external source. Given that certain requirements of a clause depend on the automotive safety integrity level (ASIL) or may be tailored, certain work products

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

may not be needed as prerequisites. A summary of the normative parts of this document is provided in Annex A.

NOTE External sources are used in this document to refer to work products that are the result of activities outside of the defined AI safety lifecycle, such as input received from the development process of the encompassing system and interpretations of tables and figures.

Tables and figures can be normative or informative depending on their context. Tables and figures that are referenced by normative requirements are considered normative unless it is explicitly specified otherwise. Other tables and figures are only informative.

If it is possible to fulfil a requirement with a different combination of methods, a rationale is provided that the chosen combination of methods fulfil the requirement.

6 AI within the context of road vehicles system safety engineering and basic concepts

6.1 Application of the ISO 26262 series for the development of AI systems

This document is intended to be applied in combination with the ISO 26262 series to specifically address the safety of AI systems.

— For AI components that are not AI models, or do not contain AI models, the ISO 26262 series can be applied by itself.

— For AI components that are AI models, or that contain AI models, the ISO 26262 series can be tailored and applied in combination with this document (see Figure 6–1).

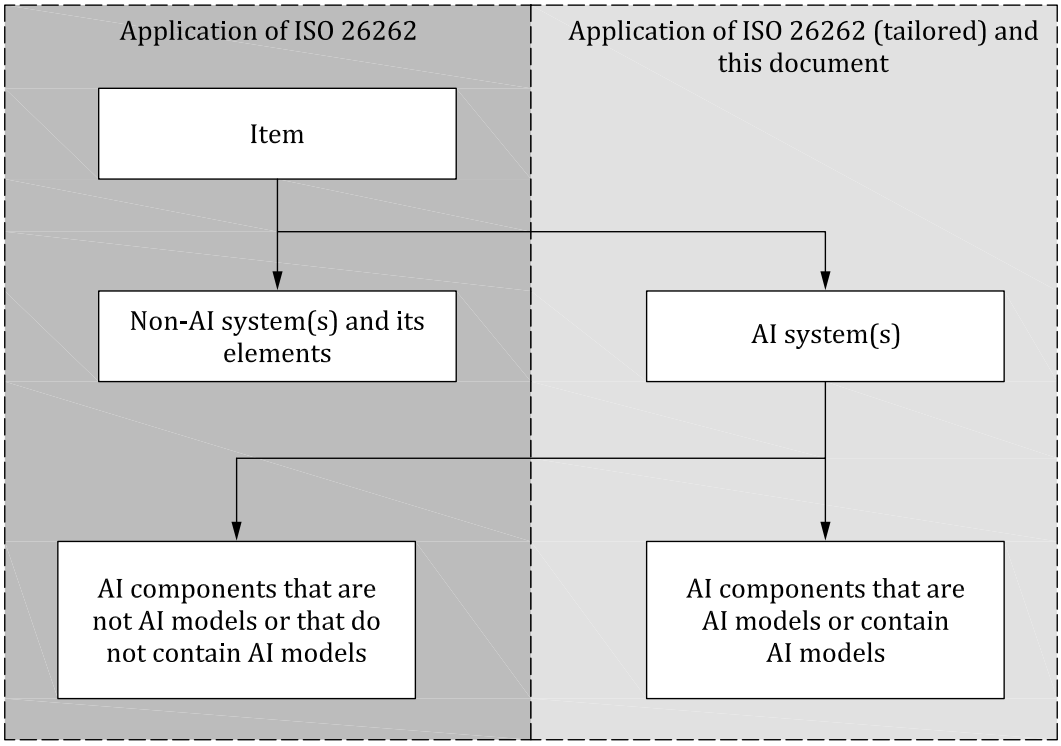


Figure 6–1 — Visualization of the applicability of the ISO 26262 series and this document to the item and its elements

NOTE See Annex C for a possible tailoring of ISO 26262-4 and ISO 26262-6 for ML.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

6.2 Interactions with encompassing system-level safety activities

An example interaction of AI system development with the encompassing system development based on the ISO 26262 series and ISO 21448 can be found in Table 6–1 and in Table 6–2. These tables also include remarks regarding the interaction and the applicability of the corresponding standards with this document and the development of the AI elements.

NOTE 1 ISO 21448 is applicable to intended functionalities where proper situational awareness is essential to safety and where such situational awareness is derived from complex sensors and processing algorithms. This applies in particular to functionalities of emergency intervention systems and systems that have SAE levels of driving automation from 1 to 5. However, this document applies to all AI systems whose errors can impact safety independent of the vehicle-level functionality. Systems can utilize AI technologies that do not fall within the scope of ISO 21448 but that do fall within the scope of this document.

NOTE 2 Although this document and ISO 21448 both focus on functional insufficiencies, conformity to one does not automatically imply conformity to the other.

During the encompassing system architecture design phase, encompassing system requirements are decomposed and allocated to the AI systems as well as other elements of the encompassing system.

Table 6–1 — Example interaction of the AI element development in conformity to this document and with the ISO 26262 series

ISO 26262:2018	Interaction of the AI system with the encompassing system (the AI system is not an item)	AI system: System component consisting of hardware and software components	AI component: Conventional [a] hardware component	AI component: AI model implemented by software component(s)
Part 2, Clause 5: Overall safety management	-	Adapted to also address the management of AI safety	Directly applicable	Adapted to also address the management of AI safety
Part 2, Clause 6: Project dependent safety management	The management of the AI safety is part of the safety management of the encompassing system	Adapted to also address the management of AI safety	Directly applicable	Adapted to also address the management of AI safety
Part 2, Clause 7: Safety management regarding production, operation, service and decommissioning	The management of the AI safety is part of the safety management of the encompassing system	Adapted to also address the management of AI safety	Directly applicable	Adapted to also address the management of AI safety
Part 3, Clause 5: Item definition	Potential source of input space specification	-	-	-
Part 3, Clause 6: Hazard analysis and risk assessment	-	-	-	-
Part 3, Clause 7: Functional safety concept	Potential source of safety requirements allocated to the AI system	-	-	-
Part 4, Clause 6: Technical safety concept	Potential source of safety requirements allocated to the AI system	Applicable (tailoring can be necessary)	Hardware safety requirements are derived from technical safety requirements allocated to the AI system	Software safety requirements are derived from technical safety requirements allocated to the AI system

[a] Conventional hardware is hardware that is not specifically designed to implement an AI model, e.g. CPUs, GPUs or FPGAs.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Table 6–1 *(continued)*

ISO 26262:2018	Interaction of the AI system with the encompassing system (the AI system is not an item)	AI system: System component consisting of hardware and software components	AI component: Conventional [a] hardware component	AI component: AI model implemented by software component(s)
Part 4, Clause 7: System and item integration and testing	AI system as a system component to be integrated into the encompassing system	Integration of the hardware and software components of the AI system (tailoring can be necessary)	-	-
Part 4, Clause 8: Safety validation	Potential source of additional validation strategies and requirements	-	-	-
Part 5: Product development at the hardware level	Potential source of hardware safety requirements allocated to the AI elements	Applicable (tailoring can be necessary)	Applicable	Refinement of hardware-software interface
Part 6: Product development at the software level	Potential source of software safety requirements allocated to the AI elements	Applicable (tailoring can be necessary)	Refinement of hardware-software interface	Applicable (tailoring can be necessary)
Part 7: Production, operation, service and decommissioning	AI elements can be part of the production process	Potential source of requirements and work products relevant for production, operation, service and decommissioning	Potential source of requirements and work products relevant for production, operation, service and decommissioning	Potential source of requirements and work products relevant for production, operation, service and decommissioning

[a] Conventional hardware is hardware that is not specifically designed to implement an AI model, e.g. CPUs, GPUs or FPGAs.

Table 6–2 — Example interactions with ISO 21448

ISO 21448:2022, Clause	Interaction of the AI system development with encompassing system activities, motivated by ISO 21448:2022
Clause 5: Specification and design	Clause 5 activities: — provide the interfaces of the AI system with the encompassing system; — determine the semantic input space; — provide the functionality required by the AI system; — provide safety requirements allocated to the AI system, including, but not limited to, safety-related KPIs. Activities of this document: — provide triggering conditions and functional insufficiencies of the AI system; — provide achieved safety-related KPIs; — provide a description of deployment measures required to support the AI and data lifecycles.
Clause 6: Identification and evaluation of hazards	Clause 6 is a potential source of safety requirements, including, but not limited to, safety-related KPIs, allocated to the AI system.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Table 6–2 *(continued)*

ISO 21448:2022, Clause	Interaction of the AI system development with encompassing system activities, motivated by ISO 21448:2022
Clause 7: Identification and evaluation of potential functional insufficiencies and potential triggering conditions	Clause 7 activities: — a potential source of safety requirements allocated to the AI system; — a potential source of AI triggering conditions. Activities of this document: — provide potential triggering conditions of the AI system.
Clause 8: Functional modifications addressing SOTIF-related risks	Activities of this document: — modification request to the encompassing system in case safety requirements allocated to the AI system cannot be fulfilled.
Clause 9: Definition of the verification and validation strategy	Clause 9 is a potential source of safety-related KPIs, allocated to the AI system.
Clause 10: Evaluation of known scenarios	Activities of this document: — provide triggering conditions of the AI system and the associated AI error modes, error patterns or a combination of both; — provide achieved safety-related KPIs.
Clause 11: Evaluation of unknown scenarios	This document provides means to achieve safety-related KPIs.
Clause 12: Evaluation of the achievement of the SOTIF	This document provides a safety assurance argument for the AI system which can be used within the evaluation of the achievement of the SOTIF.
Clause 13: Operation phase activities	This document provides the AI system requirements to the encompassing system regarding the operation phase.

During development and as part of continuous assurance activities during operation, it can become necessary to adjust the safety requirements allocated to the AI system leading to an iterative feedback cycle to the encompassing system safety concept and safety requirements. Iterations of the requirements are triggered for example if:

— an AI system that is capable of fulfilling its assigned safety requirements and associated safety-related properties cannot be feasibly developed (e.g. due to inherent performance limitations in the ML algorithm used);

— suitable training data and test data cannot be found;

— evidence to demonstrate that the safety requirements and associated safety-related properties are fulfilled cannot be collected with sufficient confidence.

In each of these cases, changes to the encompassing system safety concept can be defined, leading to a set of updated, realisable requirements on the AI system.

NOTE 3 Measures on the encompassing system safety concept to reduce the safety load of an AI system towards better feasibility can be (see also 10.5):

— restrictions in the ODD;

— implementation of diversity such as different processing algorithms or sensing modalities;

— implementation of redundancy such as multiple hardware components in parallel;

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

— a combination of the aforementioned measures.

Once the AI system has reached an adequate level of performance in relation to the safety requirements, the AI system can be integrated into the encompassing system including evidence to support the achievement of the safety requirements. This can result in the need for further iterations of the AI safety life cycle, for example, due to the following conditions:

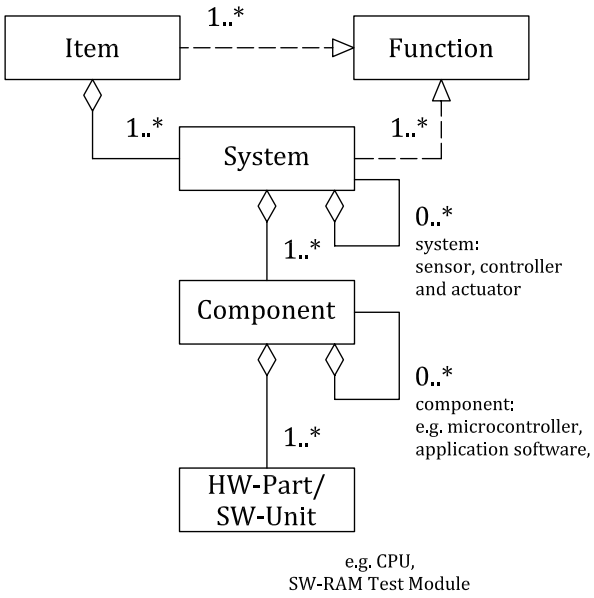
— Integration tests of the encompassing system reveal previously undiscovered faults or functional insufficiencies in the AI system that require additional development cycles.

— The encompassing system assurance case requires additional evidence to support safety claims related to the AI system that require additional effort to collect the required evidence.

Collected field data and observations made during operation related to the performance of the AI system (e.g. increased number of false positive errors under certain traffic conditions or an increased rate of out-of-distribution inputs) can indicate changes in the input space. It may not be possible to address these changes by refining the safety requirements on the AI system or through additional development activities. Changes may need to be made at the encompassing system level. This can lead to changes in the safety requirements assigned to the AI system and a repetition of the safety life cycle.

6.3 Mapping of abstraction layers between the ISO 26262 series, ISO/IEC 22989 and this document

The ISO 26262 series uses the following levels of abstraction: item, system, component, software unit and hardware part. The relationship between these is visualized in Figure 6–2 (ISO 26262-10:2018, Figure 3). It is typically used when a given requirement can be applied on different levels of abstraction, e.g. on hardware components as well as on hardware parts. An example item composition is shown in Figure 6–3 (ISO 26262-10:2018, Figure 4).



Key

realization: one instance is realized by another instance (e.g. a function or part of a function is realized by an item)

aggregation: one instance has a set of other instances (e.g. a system has a set of components)

NOTE 1 Depending on the context, the term “element” can apply to the entities “system”, “component”, “hardware part” and “software unit” in this chart, according to ISO 26262-1:2018, 3.41.

NOTE 2 "*" means N elements are possible, where N is a positive integer number.

Figure 6–2 — Relationship of item, system, component, hardware part and software unit

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

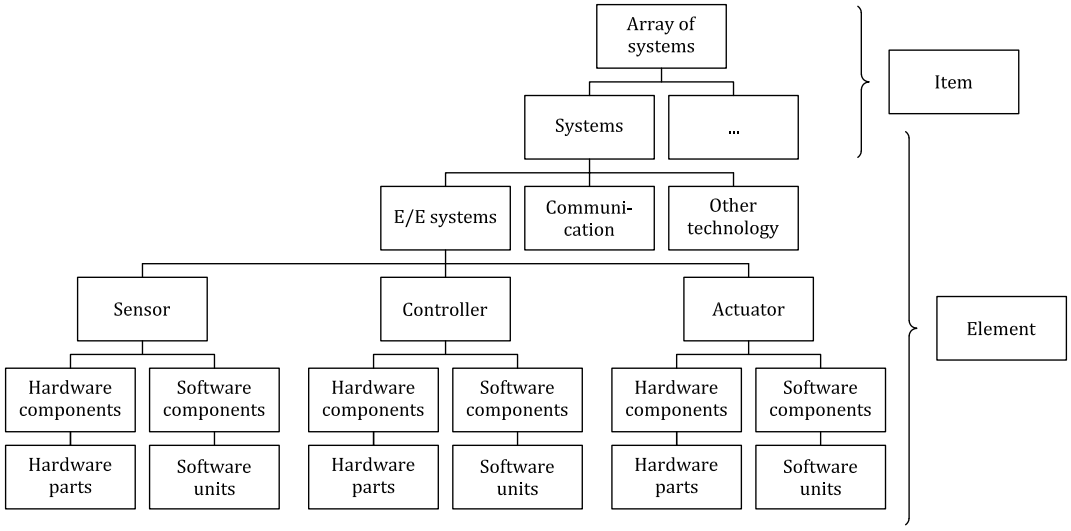
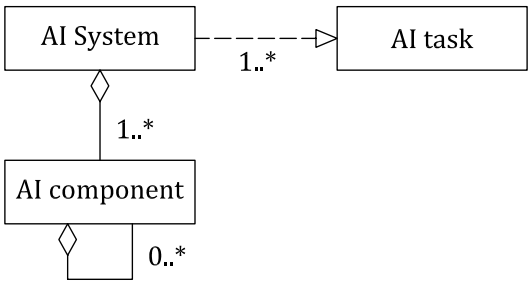


Figure 6–3 — Example item composition

ISO/IEC 22989 uses the following abstraction layers: AI system and AI components, where the AI system consists of AI components. ISO/IEC 22989 does not explicitly state if a given AI component can itself consist of AI components. In this document this is possible. The AI component can be an AI model, a conventional element, i.e. an element not considered to be an AI model, or a combination of both. The AI system contains at least one AI model and realizes the AI task. Figure 6–4 uses the same notation as Figure 6–2 to visualise the relationship between AI system and AI components.



Key

realization: one instance is realized by another instance

aggregation: one instance has a set of other instances

NOTE 1 Depending on the context, the term “AI element” can apply to “AI system” and “AI component”.

NOTE 2 "*" means N elements are possible, where N is a positive integer number.

Figure 6–4 — Relationship of AI system and AI component

This document uses terms from both ISO/IEC 22989 and ISO 26262-1. The terms from the different standards do not map one-to-one. So, depending on the context, multiple mappings are possible as shown in Table 6–3.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Table 6–3 — Possible mappings between ISO/IEC 22989 and ISO 26262-1 terms, depending on the context

ISO/IEC 22989 terminology	ISO 26262-1 terminology
AI system	Item, system, component, software unit or hardware part Element
AI component	System, component, software unit or hardware part Element

In this document, the term "AI system" is the top level of abstraction of the content to be developed. As such it is possible in a distributed development that what one party considers to be an AI component, the other party considers to be an AI system, as for them it represents the top level of the content they develop.

Figure 6–5 provides an example of an item decomposed into its elements. In this example, the item consists of two systems: system A and system B. For the sake of simplicity, system B is not further decomposed. System A is composed out of the system components A.1 and A.2. System component A.1 does not contain an AI model. As such it cannot be an AI system. System component A.2 contains an AI model and is declared to be the AI system in this example.

NOTE It would have been possible to declare system A as the AI system, as it too contains at least one AI model. The decision regarding the scope of the AI system is made as part of negotiations between the development organisations responsible for the encompassing system.

The system components themselves consist of hardware and software components. System component A.2 is considered to be an AI component as it composes the AI system. A further breakdown into software units and hardware parts of the components has been omitted for the sake of simplicity.

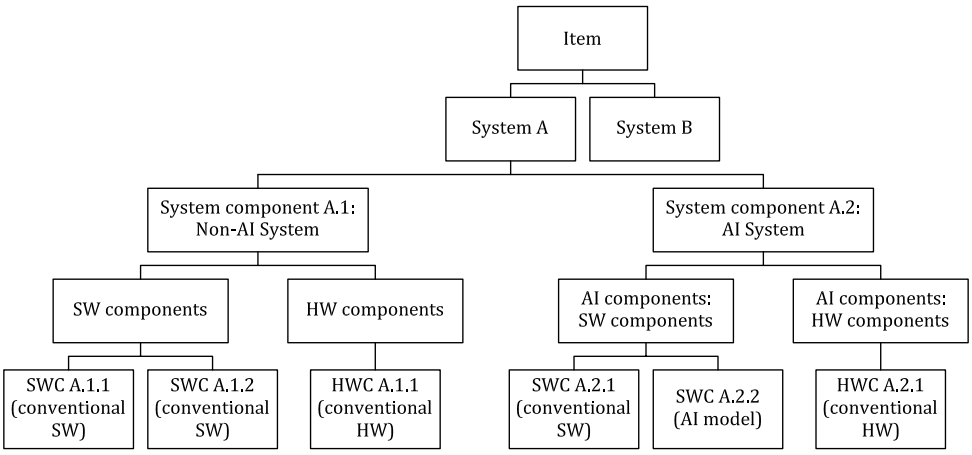


Figure 6–5 — Example of a hierarchical decomposition of an item into its elements down to the component level - decomposition tree view

6.4 Example architecture for an AI system

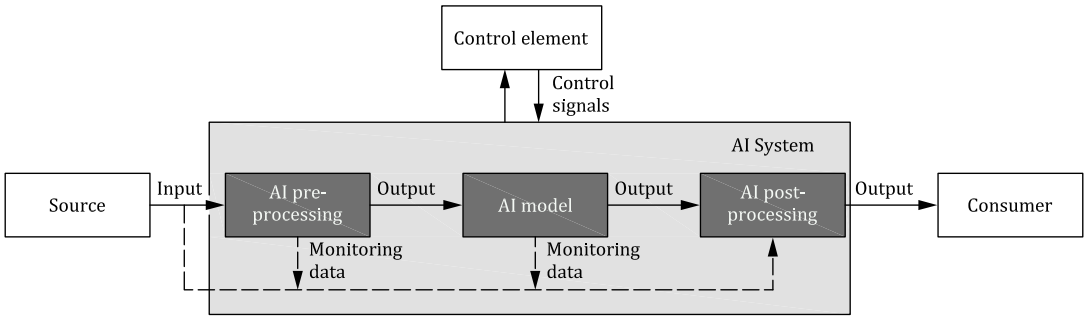
This document uses the architecture shown in Figure 6–6 as an example architecture. The AI system receives its input from the source, executes its task based on the input and the control signals and then provides its output to the consumer. The AI system itself consists of the AI components AI pre-processing, AI model and AI post-processing. The AI post-processing uses data provided from the previous process steps (i.e. AI pre-processing and the AI model) in combination with the original input data for monitoring purposes.

NOTE 1 This architecture is just an example and has no claim of representing all possible architectures of AI systems.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

NOTE 2 If the AI system is implemented for a real-time task, the execution of the AI system can be triggered synchronously or asynchronously depending on the behaviour of the control element, sources of the input streams and the consumers of the outputs. These considerations can be relevant to the definition of the safety requirements on the AI system.



Key

- AI system
- AI component
- element not belonging to this AI system

Figure 6–6 — Example architecture of an AI system

6.5 Types of AI models

Examples for types of AI models include, but are not limited to, deep neural networks, k-nearest neighbours, support vector machines, decision trees, symbolic AI and fuzzy logic. These can be clustered in different categories. For example, deep neural networks, k-nearest neighbours, support vector machines and decision trees can be categorized as ML models as shown in Figure 6–7.

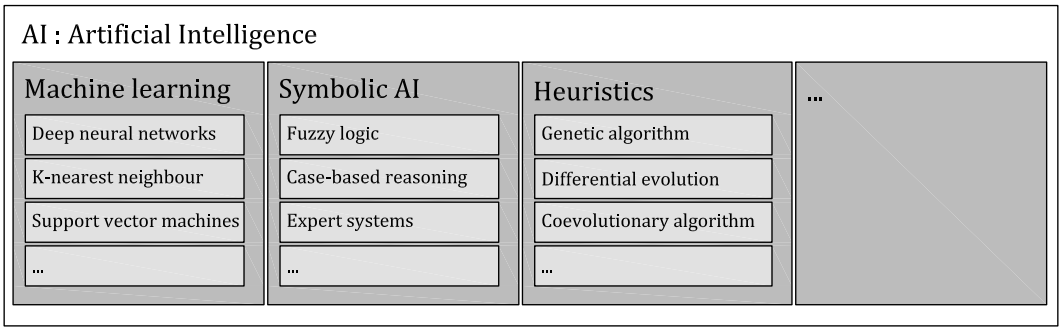


Figure 6–7 — Example of different types of AI models

6.6 AI technologies of a ML model

Figure 6–8 and Figure 6–9 show an example of possible AI technologies utilized for an ML model that is implemented in hardware and software. In addition to the AI method itself, the AI technology also contains the tools and procedures to generate the AI model.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

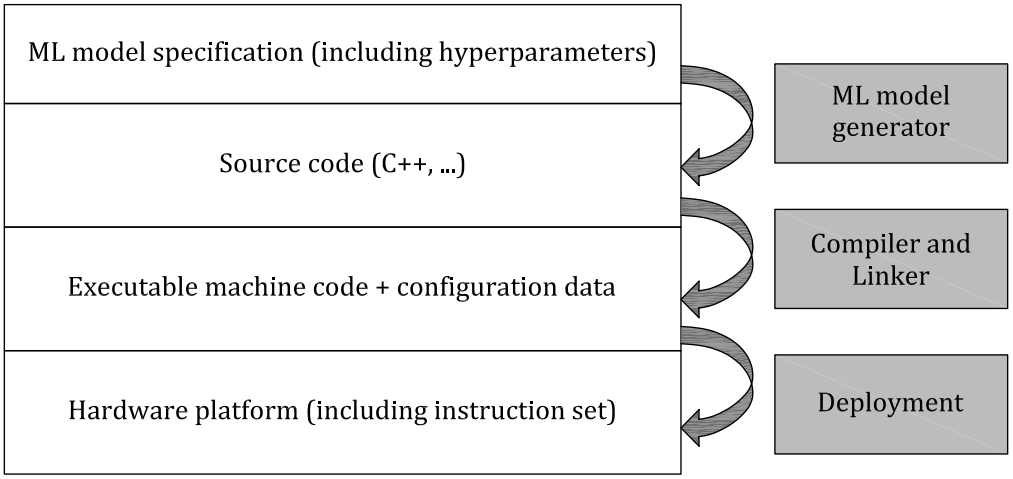


Figure 6–8 — AI technologies to create an executable ML model (application of ISO 26262)

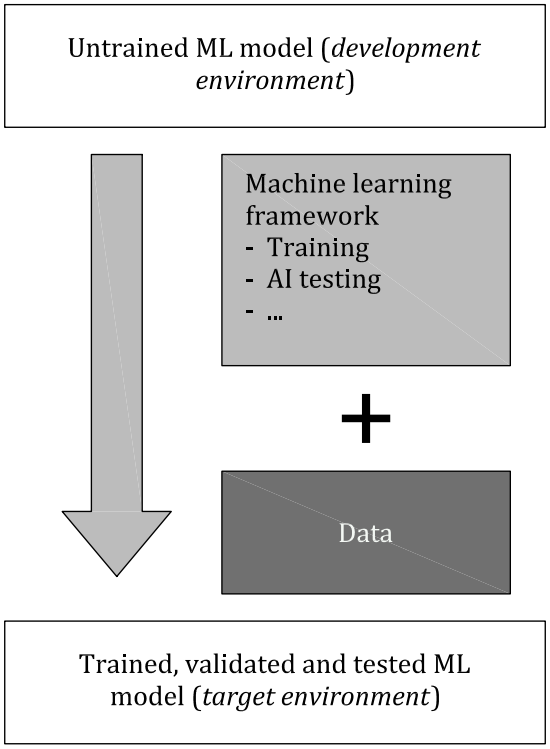


Figure 6–9 — AI technologies to create a trained ML model (application of this document)

The AI technologies listed in Figure 6–8 are not considered to be relevant sources for functional insufficiencies, e.g. compiler and linker are well known technologies already utilized by non-AI systems. For these technologies, it is enough to apply the ISO 26262 series to achieve AI safety. The AI technologies listed in Figure 6–9 are considered relevant sources of functional insufficiencies for which the application of the ISO 26262 series alone is not considered to be sufficient to achieve AI safety. For these technologies, the remaining clauses of this document are applied.

6.7 Error concepts, fault models and causal models

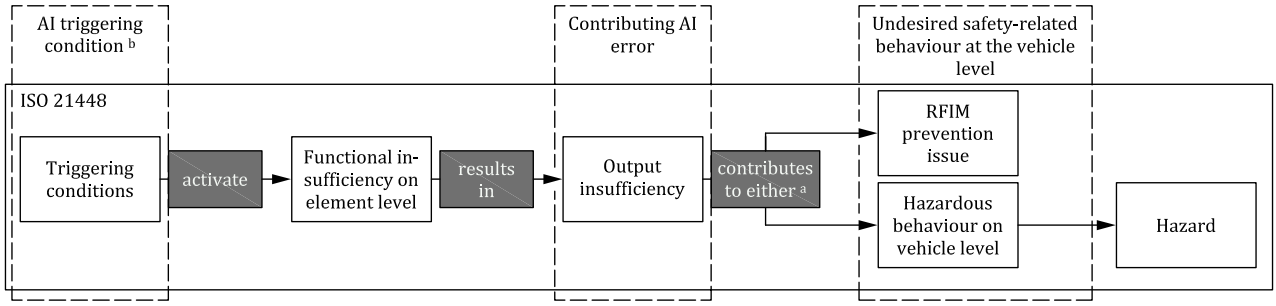
6.7.1 Cause-and-effect chain

This document utilizes the concept of AI triggering conditions, faults, functional insufficiencies, AI errors and the undesired safety-related behaviour at the vehicle level. The mapping of the terms of this document

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

to the cause-and-effect chain used by ISO 21448 can be found in Figure 6–10. The mapping of the terms of this document to the cause-and-effect chain used by ISO 26262 series can be found in Figure 6–11. The AI triggering condition activates a fault or a functional insufficiency, resulting in an AI error in case of a fault and a contributing AI error in case of a functional insufficiency.



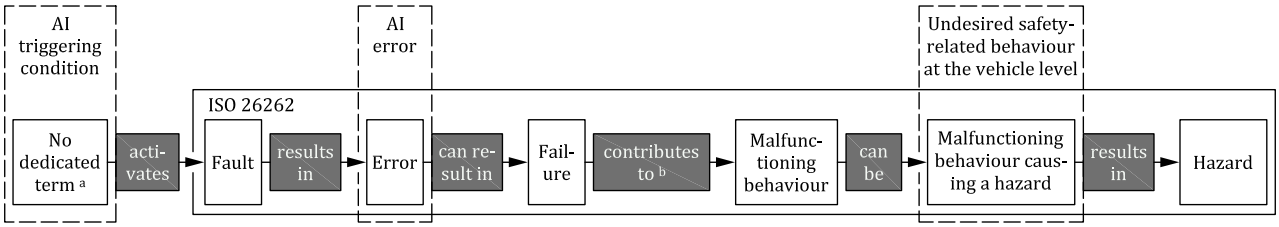
a An output insufficiency, either by itself or in combination with one or more output insufficiencies of other elements, contributes to either a hazardous behaviour at the vehicle level or an inability to prevent or detect and mitigate a reasonably foreseeable indirect misuse.

b Since a triggering condition of ISO 21448 results in a contributing AI error in the context of this document, they represent a subset of all AI triggering conditions.

Figure 6–10 — Mapping of the cause-and-effect chain of ISO 21448 to the terms of this document

EXAMPLE An insufficiency of specification can be a missing object in the AI training, AI validation and AI test dataset of an AI system utilizing an ML model for object classification. In this case, encountering this object during operation in the field is the AI triggering condition that activates this insufficiency of specification, resulting in the occurrence of a contributing error. The contributing AI error would be the incorrect classification of the object by the ML model and consequently by the AI system.

An AI error of an AI component can propagate through the AI system and can result in an AI error of the AI system. The AI error of the AI system can propagate through the encompassing system and can contribute either by itself or in combination with one or more other errors or output insufficiencies of the elements of the encompassing system to an undesired safety-related behaviour at the vehicle level.



a In the ISO 26262 series, there is neither a dedicated term for the condition that activates a fault nor is this concept explicitly utilized.

b A failure, either by itself or in combination with one or more failures of other elements, contributes to a malfunctioning behaviour.

Figure 6–11 — Mapping of the cause-and-effect chain of the ISO 26262 series to the terms of this document

The undesired safety-related behaviour at the vehicle level is used as an umbrella term for the corresponding terms of the ISO 26262 series (i.e. the malfunctioning behaviour at the vehicle level which can cause hazards) and ISO 21448 (i.e. the hazardous behaviour at the vehicle level and RFIM prevention issue).

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

6.7.2 Root cause classes

Different error classes can be distinguished depending on the root cause. The correlation between contributing AI errors and their different root causes is shown in Figure 6–12.

NOTE 1 A contributing AI error of the AI element can lead to the undesired safety-related behaviour at the vehicle level by itself or in combination with one or more other AI errors.

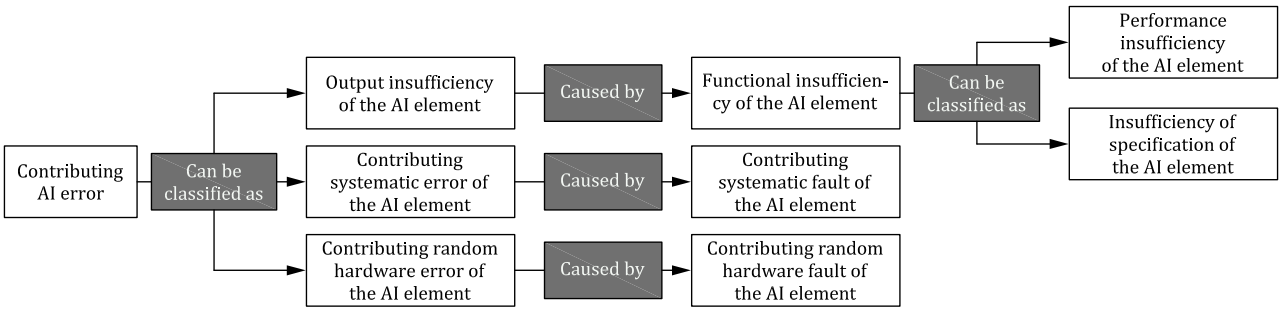


Figure 6–12 — Correlation of safety-related errors with their different classes of root causes

The root causes for the different kinds of AI errors are:

— insufficiency of the specification

EXAMPLE 1 An insufficiency of the specification can be missing datasets in the AI training, AI validation or AI test dataset. The resulting output insufficiency can be a misclassification when exposed to the missing datasets.

EXAMPLE 2 Specification of a neural network model with insufficient complexity.

EXAMPLE 3 An inadequate training loss function.

EXAMPLE 4 Inadequate labelling specification.

— performance insufficiency

EXAMPLE 5 A performance insufficiency can be an insufficient range of a sensor in case of certain environmental conditions. The resulting output insufficiency can be a false negative detection of an obstacle in the trajectory.

NOTE 2 In the case of ML models, performance insufficiencies can be specifically caused by training and test dataset related issues, e.g. insufficiencies in the coverage of the respective input space. These data-related issues are in turn considered to be insufficiencies of specification, or more precisely as insufficiency of specification of the data.

— contributing systematic fault

EXAMPLE 6 A contributing systematic fault can be to divide by zero in the software or to use incorrect variable names.

EXAMPLE 7 Overfitting the DNN resulting in wrong high-confidence classification outputs of corner cases can be regarded to be a contributing systematic fault in the training procedure.

NOTE 3 Sometimes the classification of a given issue in either a systematic fault or a functional insufficiency can be ambiguous. Independent of the classification, a safety assurance argument is provided to argue that this issue does not represent an unreasonable risk. As long as this safety assurance argument is available, the exact classification is not relevant.

— contributing random hardware fault

EXAMPLE 8 Physical defect causing a short to ground.

When evaluating the effectiveness of safety mechanisms, it can be necessary to distinguish the different classes of errors.

EXAMPLE 9 Homogenous redundancy can be effective in detecting random hardware errors in one of the redundant elements, but it is not effective in detecting systematic errors.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

The different fault classes can also be used within the safety assurance argument by addressing each root cause class with a dedicated set of safety measures.

EXAMPLE 10 Coding guidelines are a measure to avoid systematic faults in SW. In combination with other fault avoidance measures, e.g. as specified in ISO 26262-6, the absence of unreasonable risk due to systematic faults can be argued.

It can also be useful to classify the AI triggering conditions in different categories.

EXAMPLE 11 For ML based AI models, the AI triggering conditions can be distinguished as:

— cases that are similar to known training data samples ("in-distribution" cases), such as: cases at decision boundaries ("hard" cases), and undecided cases (high aleatoric uncertainty, e.g. due to label noise);

— unseen cases ("out-of-distribution" cases), such as: novel objects (semantic shift), or novel image styles (covariate shift);

— addition of small changes to a sample which causes no error (e.g. addition of adversarial crafted perturbation);

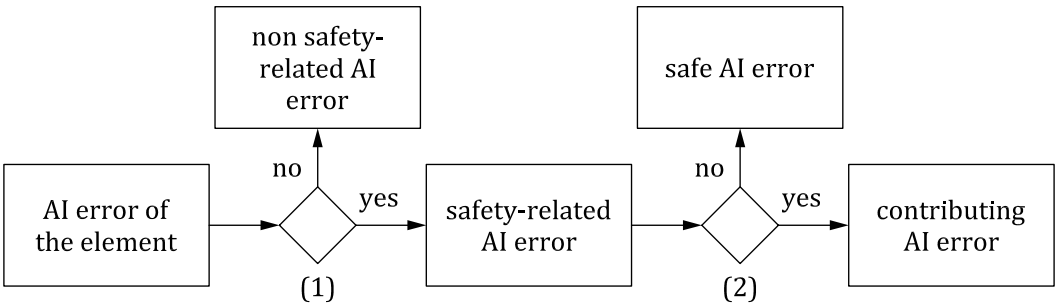
— applying a maliciously inserted trigger pattern (e.g. inserted via data or model poisoning).

Statistically obtained ML models like deep neural networks exhibit for any output an inherent uncertainty about their correctness. Therefore, besides classification of AI errors, root causes, and AI triggering conditions, a classification of uncertainty types and their associated sources can also be helpful in practice.

EXAMPLE 12 The two major types used to model the uncertainty of the ML model are epistemic and aleatoric uncertainty. While epistemic uncertainty stems from uncertainty of having the right model for the given sample and can often be fixed with more data, aleatoric uncertainty stems from intrinsic noise in the training data.

6.7.3 Error classification based on the safety impact

Compared to the criticality classification of random hardware faults as described in ISO 26262-5, this document uses a simplified scheme as shown in Figure 6–13.



Key

(1) Is the element a safety-related element?

(2) Can the error significantly contribute to the occurrence of a safety-related undesired behaviour at the vehicle level?

Figure 6–13 — Error classification scheme based on the potential to lead to an undesired safety-related behaviour at the vehicle level

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

7 AI safety management

7.1 Objectives

The objectives of this clause are:

a) to define an AI safety lifecycle and its activities to ensure that contributing errors of the AI system do not lead to unreasonable risk of undesired safety-related behaviour at the vehicle-level. The AI safety lifecycle includes:

   1) a definition of activities necessary to develop the AI system, to provide the assurance and the evidence that the AI system is safe and to ensure the AI safety during operation;

   2) in case of the utilization of ML based AI technologies: a data-driven, iterative approach for the development, evaluation and continuous assurance of AI system within the context of a encompassing system's safety lifecycle;

b) to ensure that overall and project specific safety management processes and activities are appropriate to ensure the safety of the AI system;

   NOTE ISO 26262-2 provides suitable guidance on overall safety management and project-specific safety management. This guidance can require extensions based on recommendations in this document.

c) to plan, initiate and conduct the AI safety activities.

7.2 Prerequisites and supporting information

The following information shall be available (from external sources, e.g. the encompassing system development):

a) the AI system definition, including:

   1) the AI system functionality;

   2) the interfaces of the AI system with the encompassing system, including if applicable, the ASIL capability of the inputs to the AI system;

   3) the safety requirements allocated to the AI system, including if applicable:

      i) the ASIL rating of the safety requirements;

      ii) the acceptance criteria or validation targets derived in conformity to ISO 21448:2022, Clause 6 or 9.

NOTE 1 The safety requirements allocated to the AI system from external sources are typically requirements regarding the avoidance or control of safety-related faults, the allowed maximum error occurrence rate of contributing AI errors, the identification of AI triggering conditions, and the robustness against certain environmental conditions.

NOTE 2 For an elaboration of the fault model, the causal model, and the error concepts used by this document, see 6.7.

7.3 General requirements

7.3.1 An AI safety lifecycle shall be defined that specifies the activities necessary to develop the AI system, to provide the assurance and the evidence that the AI system is safe and to ensure the maintenance of AI safety during operation. It can be based on the reference lifecycle (Figure 7–2) and can be tailored according to project-specific needs. The tailoring is supported by a rationale for why the tailored AI safety lifecycle is appropriate to achieve AI safety.

7.3.2 At each phase within the AI safety life cycle, work products shall be defined to support the safety assurance claims of the AI system.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

7.3.3 The activities of the AI safety lifecycle shall be coordinated with the safety lifecycle activities of the encompassing system as defined by ISO 26262-2 and, if applicable, ISO 21448.

NOTE The AI system can be developed as a safety element out of context. The concept of a safety element out of context is described in ISO 26262-10:2018, Clause 9.

7.3.4 The activities described in ISO 26262-2 shall be adapted in order to address the management of AI safety, including:

a) the integration of the AI safety lifecycle into the ISO 26262 series safety lifecycle (see ISO 26262-2:2018, Figure 2);

b) the enhancement of "functional safety" to "AI safety";

c) measures to ensure that a sufficient level of cross domain competences regarding safety and AI are available, in conformity to ISO 26262-2:2018, 5.4.4.1;

d) adding this document as a relevant standard of ISO 26262-2;

e) the use of a safety assurance argument as part of the safety case of ISO 26262-2;

f) extending the safety plan of ISO 26262-2 to include the safety activities of this document;

g) tailoring ISO 26262-2:2018, Table 1, to address the work products of this document (see Table 7–1).

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Table 7–1 — Required confirmation measures, including the required level of independence

Confirmation measure	Level of independence [a] applies to					Scope
	QM	ASIL A	ASIL B	ASIL C	ASIL D	
Confirmation review of the safety plan Independence with regard to the developers of the item, [b] project management and the authors of the work product	-	I1	I1	I2	I3	Applies to the highest ASIL among the safety requirements
Confirmation review of the AI system validation report Independence with regard to the developers of the item, [b] project management and the authors of the work product	-	I0	I1	I2	I2	Applies to the highest ASIL among the safety requirements
Confirmation review of the AI safety analyses Independence with regard to the developers of the item, [b] project management and the authors of the work product	-	I1	I1	I2	I3	Applies to the highest ASIL among the safety requirements

[a] The notations are defined as follows:

— -: no requirement and no recommendation for or against regarding this confirmation measure;

— I0: the confirmation measure should be performed; if the confirmation measure is performed, it shall be performed by a different person in relation to the person(s) responsible for the creation of the considered work product(s);

— I1: the confirmation measure shall be performed, by a different person in relation to the person(s) responsible for the creation of the considered work product(s);

— I2: the confirmation measure shall be performed, by a person who is independent from the team that is responsible for the creation of the considered work product(s), i.e. by a person not reporting to the same direct superior;

— I3: the confirmation measure shall be performed by a person who is independent, regarding management, resources and release authority, from the department responsible for the creation of the considered work product(s).

[b] The developers of the item include the developers of the AI system.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

Table 7–1 *(continued)*

Confirmation measure	Level of independence [a] applies to					Scope
	QM	ASIL A	ASIL B	ASIL C	ASIL D	
Confirmation review of the safety assurance argument Independence with regard to the authors of the safety assurance argument	-	I1	I1	I2	I3	Applies to the highest ASIL among the safety requirements
AI safety audit Independence with regard to the developers of the item [b] and project management	-	-	I0	I2	I3	Applies to the highest ASIL among the safety requirements
AI safety assessment Independence with regard to the developers of the item[b] and project management	-	-	I0	I2	I3	Applies to the highest ASIL among the safety requirements

[a] The notations are defined as follows:

— -: no requirement and no recommendation for or against regarding this confirmation measure;

— I0: the confirmation measure should be performed; if the confirmation measure is performed, it shall be performed by a different person in relation to the person(s) responsible for the creation of the considered work product(s);

— I1: the confirmation measure shall be performed, by a different person in relation to the person(s) responsible for the creation of the considered work product(s);

— I2: the confirmation measure shall be performed, by a person who is independent from the team that is responsible for the creation of the considered work product(s), i.e. by a person not reporting to the same direct superior;

— I3: the confirmation measure shall be performed by a person who is independent, regarding management, resources and release authority, from the department responsible for the creation of the considered work product(s).

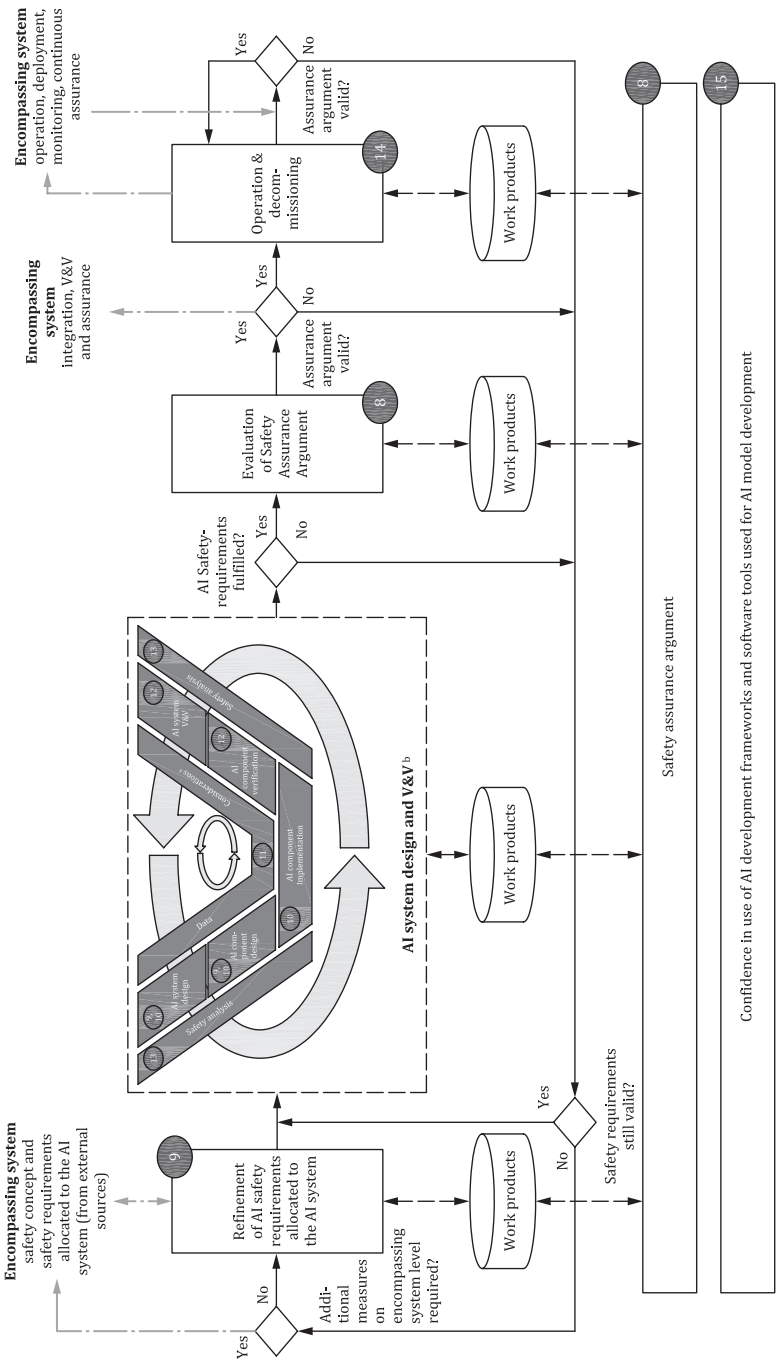
[b] The developers of the item include the developers of the AI system.

7.4 Reference AI safety life cycle

The reference AI safety life cycle described in this clause covers the activities at the different phases of AI system development, deployment and operation: safety-related requirements derivation, AI system design, verification and validation, deployment and operation. The AI safety life cycle is summarised in Figure 7–1 and is used to structure the remainder of this document. A detailed view of the AI system design and V&V phase is shown in Figure 7–2. Clause 8 through Clause 15 (as indicated by the numbered black circles in Figure 7–1 and Figure 7–2) are used to describe the activities within the safety life cycle in more detail.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024



Key

X, Y clause number(s)

process flow to/from the development and operation of the encompassing system

process flow

relation

represents the iterative nature, in particular of the AI component design and verification

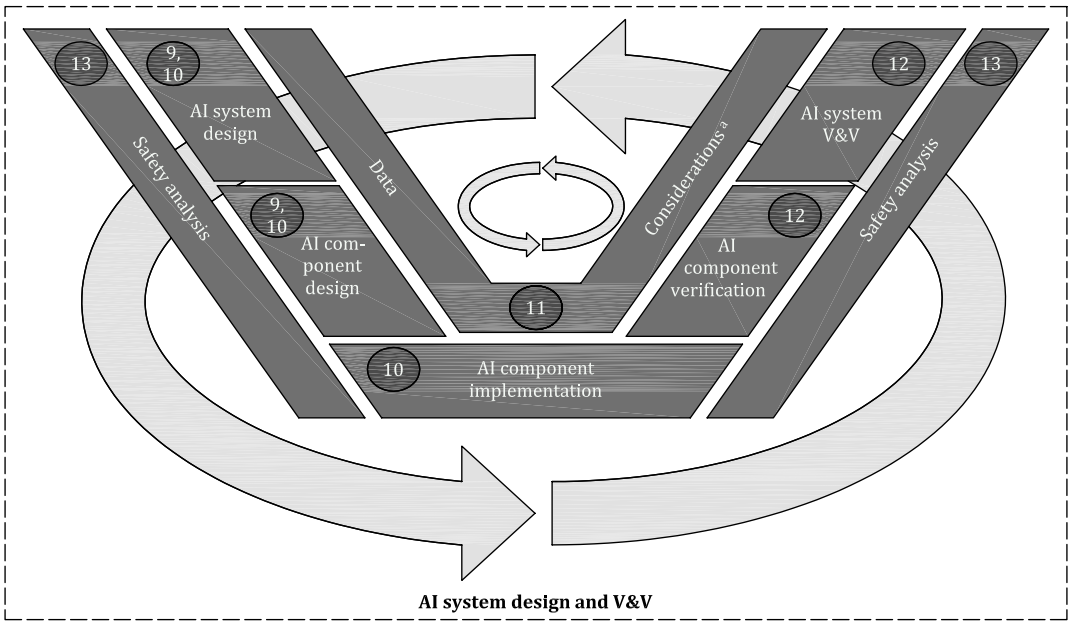
a Specific to ML-based AI technologies.

b See Figure 7-2.

Figure 7–1 — Reference AI safety lifecycle

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024



Key

X, Y — clause number(s)

represents the iterative nature, in particular of the AI component design and verification

a Specific to ML-based AI technologies.

Figure 7–2 — Detailed view of the AI system design and V&V phase of the reference AI safety lifecycle

7.5 Iterative development paradigms for AI systems

The AI system development phase of the AI safety life cycle covers all activities required to design, implement, verify and validate the AI system.

The AI system development activities described in Clause 8 through Clause 13 are iteratively performed until a sufficient level of performance with respect to the safety requirements and the associated safety properties can be demonstrated and associated work products are generated.

An essential characteristic of this development process is the analysis to identify potential functional insufficiencies, their root causes, and their impact on safety. This analysis is used to derive appropriate measures to reduce these functional insufficiencies during design (including the selection of training data in case of ML) as well as to reduce the impact of contributing errors (through architectural measures).

NOTE 1 During the AI system development, the property of "independence" is considered wherever appropriate (e.g. independence between training datasets and test datasets). An analysis similar to the dependent failure analysis (DFA) described in ISO 26262-9:2018, Clause 7 can be used.

The development model described here reflects the iterative development model typically used in the area of ML with a focus on activities to identify, analyse, reduce and mitigate functional insufficiencies in the trained model and the cumulative collection of evidence to support claims regarding safety requirements allocated to the function. Furthermore, for machine learning, the specification and collection of suitable training and test data is one of the most influential factors for the performance of the function. Therefore, the specification, planning, collection, acquisition, preparation and labelling of data related to AI component implementation and verification is treated as a safety-related development activity in this document, with specific objectives and associated safety artefacts (see Clause 11).

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

The iterative development of the AI system is guided by a set of performance indicators, including safety-related properties associated with the safety requirements allocated to the AI system.

NOTE 2 "Iteration" in the context of AI system development can be defined as a single repetition of one or more AI system safety lifecycle phases. It is applied until conditions defined by a set of KPIs or other target parameters are either fulfilled or demonstrated to be unachievable.

NOTE 3 Considering that there can be multiple, potentially conflicting target KPIs, the presence of contributing errors in the AI system can be inevitable.

Despite the inherently iterative nature of AI system development, different iteration cycles of the development can focus on different sets of performance indicators, depending on the maturity of the development.

Examples of such distinct iteration cycles can include:

— Proof of concept: In this phase of development, AI models are designed, implemented and tested against a set of initially defined safety requirements. The objective of this phase is to evaluate the potential of the chosen AI technologies to fulfil the safety requirements, as well as defining a set of measures required to minimise the number or the impact of functional insufficiencies in the function (e.g. optimisation of model parameters and defining a data collection strategy).

— Series development of the AI system: In this phase of development, the AI system is iteratively developed against all safety requirements. Errors of the AI system or the AI model are analysed with respect to their potential root causes. Measures are defined to reduce functional insufficiencies through design and data selection or to minimise their impact through architectural measures. This phase of development can use a host platform to implement and test the AI model and continue until an adequate level of performance for all safety-related properties has been met.

— Deployment to the target hardware platform: In this phase of development, the AI system is transferred to the target hardware platform and target software platform. This can include, for example a change in numerical precision used to calculate the results as well as the consideration of constraints such as timing, memory limitations (e.g. resulting in the need to prune the computational graph) and robustness against potential random hardware faults. The focus of this development phase is to ensure that the safety requirements are met, despite any limitations of the target hardware platform and the target software platform.

— Further improvement after deployment: In this phase of development, the AI system is iteratively updated based on observations made during operation in the field and new requirements mandated by the developer. This can include compensating for previously unknown triggering conditions (e.g. concept drift) and distributional shift in the environment (e.g. domain shift). Performance indicators within this phase of development are monitored to ensure a monotonic safety improvement with respect to previous iterations of the AI model. This phase of development can also include development versions of the function running in "shadow" mode within an operational environment in order to collect suitable data and evaluate the potential performance under realistic conditions.

7.6 Work products

7.6.1 AI safety lifecycle resulting from 7.3.1 to 7.3.4.

7.6.2 Work products of ISO 26262-2:2018, 5.5, resulting from 7.3.4.

7.6.3 Work products of ISO 26262-2:2018, 6.5, resulting from 7.3.3 and 7.3.4, in particular the safety plan.

7.6.4 Work products of ISO 26262-2:2018, 7.5, resulting from 7.3.4.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

8 Assurance arguments for AI systems

8.1 Objectives

The objectives of this clause are:

a) to develop an assurance argument demonstrating that the safety requirements allocated to the AI system are fulfilled;

NOTE 1 The assurance argument for the AI system contributes to the safety assurance argument of the encompassing system.

NOTE 2 The assurance argument can be developed independently from the encompassing system as a safety element out of context (SEooC) development activity (see ISO 26262-10:2018, Clause 9). In such cases, the assurance argument documents all necessary assumptions on the encompassing system for arguing that the safety requirements allocated to the AI system are fulfilled.

b) to evaluate whether the assurance argument reflects the actual residual risk of the AI system violating its safety requirements.

8.2 Prerequisites and supporting information

The following information shall be available at the initiation of these activities:

a) the AI system definition (from external sources), e.g. the encompassing system development, including:

1) a specification of the safety requirements allocated to the AI system;

2) a definition of the technical context within the encompassing system (e.g. definition of interfaces to and from the encompassing system and, if applicable, the environment, conditions under which the AI system functionality is triggered, etc.);

NOTE 1 This includes the ASIL capability and noise to signal ratio of the input signals provided by the source, if applicable.

3) a specification of the input space;

b) requirements on the assurance argument and work products for the AI system (from external sources). These requirements can be derived from the assurance argument of the encompassing system as well as safety management procedures from Clause 7;

The following information shall be available for the finalization of these activities:

c) the work products of the AI safety lifecycle;

NOTE 2 This body of evidence can be cumulatively collected as part of the iterative development process phases or produced within dedicated development process cycles.

The following information can be considered for the finalization of this phase:

d) required properties of the encompassing system to achieve AI safety;

e) relevant properties of the input space (e.g. distribution of critical events, physical constraints on changes of input values over time);

f) evidence of organization-specific rules and processes for AI safety, evidence of competence management and evidence of a quality management system, from 7.6.2;

g) evidence that the organization-specific rules and processes have been followed and that the work products have the required maturity and quality (see 7.6.3 and 7.6.4).

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

8.3 General requirements

8.3.1 An assurance argument for the fulfilment of the safety requirements allocated to the AI system shall be provided.

NOTE 1 The assurance argument can be part of the encompassing system's safety case in accordance with ISO 26262-2:2018, 6.4.8.

NOTE 2 The assurance argument can be constructed at the level of the encompassing system. In this case, the development of the AI system-specific contributions and supporting evidence are considered part of the AI safety life cycle and therefore within the scope of this document.

8.3.2 The assurance argument shall use the relevant work products generated during the AI safety lifecycle to support the assurance claims.

NOTE Changes to the work products and their impact on the assurance argument are considered as part of change management throughout the AI safety life cycle.

8.3.3 Confirmation measures of 7.3.4 shall be applied to the assurance argument.

NOTE The evaluation of the validity of the assurance argument can be performed as part of confirmation measures of the encompassing system (see ISO 26262-2:2018, 6.4.9 and ISO 21448:2022, 12.3), or as an independent activity, for example in the case of a SEooC (see ISO 26262-10:2018, Clause 9) or as part of a distributed development.

8.4 AI system-specific considerations in assurance arguments

A number of AI system-specific considerations impact the creation of the assurance argument.

a) The formulation of the AI safety requirements (see Clause 9):

- These include quantitative properties expressed in the form of probabilities (e.g. proportion of false positive classifications).
- Arguments are expressed that demonstrate that these properties have been achieved with a level of statistical confidence appropriate to the quantitative targets associated with the requirement's acceptance criteria.
- This can lead to additional requirements on the nature of evidence to support the claim and how the validity of this evidence is evaluated.

b) Statistical arguments related to aggregated performance metrics:

- These might not be sufficient to argue a suitable level of safety in rare but critical situations (e.g. edge cases, sensor defects or adversarial perturbations).
- Arguments can be required to demonstrate that such input conditions nevertheless lead to a suitable level of AI safety.
- The probability of unknown triggering conditions leading to a violation of safety requirements can depend on:
  - features of the input space not directly related to the function (e.g. due to spurious correlations in the training data);
  - predictions based on past inputs (e.g. the accuracy of previous detections of dynamic objects can impact the future behaviour of a planning task).

c) Verification of the AI system:

- Direct introspective approaches of AI models might not be effective.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

— Alternative means of arguing the correct behaviour of the AI model or an increased reliance on indirect verification (e.g. test) can be required.

EXAMPLE Due to the lack of transparency with respect to the individual contributions of the large number of parameters used in some ML models, introspective approaches to verification can have limited applicability.

d) Reliance on training and test data:

— In machine learning-based AI methods, the behaviour of the AI system as well as its verification and validation are predominantly reliant on the selection of suitable training and test datasets as well as the training procedures themselves.

— Dedicated assurance arguments for demonstrating how the data selection process supports the achievement of AI safety can be required (see Clause 11).

— These arguments consider the training process and associated tools (see Clause 15).

e) Conditions during operation:

— Conditions can occur during operation that invalidate the assurance argument due to the complex nature of the environment in which vehicles containing AI systems are deployed.

— These conditions might include distributional shift of the input space (e.g. new types of road vehicles, changes in road infrastructure), changes to the technical system (e.g. replacement or upgrade of sensors) or previously undiscovered unknown triggering conditions.

— A continual, periodic re-evaluation and adaptation of the assurance argument is therefore performed, including an impact analysis of which parts of the assurance argument and associated evidence are to be re-evaluated (see Clause 14).

NOTE The degree to which continual, periodic re-evaluation is required depends on the properties of the input space and operating environment. If the input space, its distribution and change of distribution over time (e.g. due to ageing) are well known, re-evaluation can be performed within regular software update activities.

8.5 Structuring assurance arguments for AI systems

8.5.1 Context of the assurance argument

Within the scope of this document, assurance relates to the claim that the AI system achieves AI safety. An assurance argument communicates the relationship between evidence and the AI safety requirements.

The level of confidence in the assurance argument should be appropriate to the required level of integrity (functional safety) and acceptance criteria assigned to the AI system within the context of the encompassing system.

NOTE A model-based graphical representation of the assurance argument can aid the communication and evaluation of the assurance argument. Examples of graphical notations for assurance arguments include the goal structuring notation (GSN)[22] and claims argument evidence (CAE)[23] based on the structured assurance case metamodel (SACM)[24].

The structure of the assurance argument can appeal to one or a combination of the following perspectives:

— features of the implemented item (product argument);

— features of the development measures and assessment process (process argument);

— factors impacting the residual risk associated with the AI system (e.g. potential causes of insufficiencies and failure modes).

EXAMPLE 1 Process-focused aspects of the assurance argument for the AI system can include an argument for the appropriate tailoring of the AI safety life cycle and the effectiveness with which each activity has been performed, based on an evaluation of the work products developed in each phase.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024

EXAMPLE 2 A risk-oriented assurance structure can include an argument that all possible causes of contributing AI errors are identified (e.g. via safety analyses), and suitable countermeasures for each cause have been identified and implemented, either through specific development measures or dedicated architectural measures.

The assurance argument for the AI system begins with a claim that the safety requirements allocated to the AI system are achieved. This can include statements related to a reasonable level of residual AI errors with respect to the AI safety requirements and the target functionality of the AI system.

An explicit definition of the context, as well as relevant assumptions, increases the transparency of the assurance argument and limits the scope of the argument to the specific AI system, its technical context within the encompassing system and its operating conditions.

EXAMPLE 3 Examples of context information that can be referenced by the assurance argument include:

— definition of the technical system context of the encompassing system;

— definition of the set of environmental conditions and operating context for which the assurance argument is valid;

— potential causes of contributing AI errors considered as part of the assurance argument.

EXAMPLE 4 Examples of assumptions that might be discharged as separate arguments can include:

— assumptions on the usage and operational profile of the AI system;

— assumptions on the reliability of inputs to the AI system;

— assumptions on the fundamental performance potential of the chosen AI technology.

An example of an assurance argument for an AI system structured according to a strategy that addresses possible sources of insufficiencies can be found in Annex B.

8.5.2 Categories of evidence

The following categories of evidence in the form of work products created during the AI safety life cycle can be considered for use within the assurance argument.

a) Addressing insufficiencies in the specification of the AI safety requirements:

   — Evaluation of the completeness of the definition of the environmental conditions and operating context (input space). This is used to confirm completeness requirements on training and test datasets (see Clause 9).

   — Evaluation of the validity of the AI safety requirements derived from the safety requirements allocated to the AI system (from external sources). This includes traceability to safety requirements allocated to the AI system and a review of the completeness and consistency of the safety-related properties used to define the AI safety requirements (see Clause 9).

b) Addressing performance insufficiencies in the design of the AI system:

   — Justification for the selection of the chosen AI methods, AI technologies and AI system architecture. This can include references to performance benchmarks and analysis indicative of the fundamental potential of the chosen technology and AI system architecture to meet the safety requirements (see Clause 10).

   — Evaluation of the effectiveness of architectural and development measures. This can include an evaluation of the ability of architectural and development measures (see Clause 10) to limit the impact of contributing AI errors in the AI model.

     NOTE 1 These measures can include hyperparameter optimization (a development measure) as well as monitoring components (an architectural measure) that detect inconsistencies in the outputs and trigger a dedicated AI error reaction to ensure AI safety. This can include components that ensure a continuous availability of the functionality through redundancy and voting, and dynamic adaptation of vehicle behaviour based on the evaluated performance of the AI system.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/PAS 8800:2024