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

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) 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.

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;

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.

— 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.

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*

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.

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

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