

DIN/TS 92004

DIN

ICS 35.240.01

**Künstliche Intelligenz –
Qualitätsanforderungen und -prozesse –
Risikoidentifikation und -analyse für KI-Systeme im gesamten
Lebenszyklus; Text Englisch**

Artificial intelligence –
Quality requirements and processes –
Risk identification and analysis for AI systems along the entire life cycle; Text in English

Intelligence artificielle –
Exigences et processus de qualité –
Schéma de risque pour les systèmes d'IA tout au long du cycle de vie; Texte en anglais

Gesamtumfang 38 Seiten

DIN-Normenausschuss Informationstechnik und Anwendungen (NIA)

Contents

	Page
Foreword	4
Introduction	5
1 Scope	6
2 Normative references	6
3 Terms and definitions	6
4 Introduction to AI risk assessment	7
4.1 Notion of risk	7
4.2 AI risk assessment along the lifecycle	7
5 Reliability specifications for AI systems	9
5.1 General	9
5.2 Risk identification	9
5.3 Risk analysis	10
5.3.1 General	10
5.3.2 Reliability in standard cases	11
5.3.3 Robustness	14
5.3.4 Intercepting errors at model level	16
5.3.5 Uncertainty estimation	20
6 Unwanted bias	20
6.1 General	20
6.2 Risk identification	20
6.3 Risk analysis	21
6.3.1 General	21
6.3.2 Identifying sensitive characteristics of (potentially disadvantaged) groups/individuals	22
6.3.3 Analysing and quantifying data bias	22
6.3.4 Checking for legal or regulatory relevance of the AI system’s potential biases	23
6.4 Best practices	23
7 Autonomy and control	23
7.1 General	23
7.2 Risk identification	23
7.3 Risk analysis	25
7.3.1 General	25
7.3.2 Appropriate and responsible task distribution between humans and AI application	26
7.3.3 Information and empowerment of users and affected persons	30
Annex A (informative) Performance metrics for Machine Learning tasks	35
A.1 Reliability metrics	35
A.2 Unwanted bias metrics	36
Annex B (informative) Begriffe	37
Bibliography	38

Figures

Figure 1 — Example AI system life cycle model with AI system-specific processes [1]	8
Figure 2 — Risk management process [2]	9

Tables

Table 1 — Reliability risk categorization	10
---	----

Table 2 — Unwanted bias risk categorization 21

Table 3 — Autonomy and control risk categorization 25

Foreword

This document has been prepared by Working Group NA 043-01-42-01 AK "KI-Qualität und Prüfverfahren" of Joint Working Committee NA 043-01-42 GA „DIN/DKE Gemeinschaftsarbeitsausschuss Künstliche Intelligenz“ of *DIN-Normenausschuss Informationstechnik und Anwendungen (NIA)* (DIN Standards Committee Information Technology and IT Applications).

Technical Specifications are not part of the body of German Standards.

A Technical Specification is the result of standards work which is not yet being published as a standard by DIN because there are certain reservations regarding its content, or because it was not prepared in accordance with the normal procedures for standards work.

No draft of this Technical Specification has been published.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. DIN shall not be held responsible for identifying any or all such patent rights.

Please tell us what you think of this Technical Specification:

- preferably in the form of a table, sent as a file by e-mail to nia@din.de. The template for this table is available for download from www.din.de/stellungnahme;
- or as hard copy to the DIN Standards Committee NIA, 10772 Berlin.

The document is prepared in English as it is intended to transfer it to ISO in midterm. The German definitions for the Terms mentioned in Clause 3 are given in Annex B.

For current information on this standard, please go to DIN's website (www.din.de) and search for the document number in question.

This DIN/TS is the result of „Zertifizierung von Standard-KI-Anwendungen (Zertifizierte KI)“ [en: „Certification of Standard-AI-Applications (Certified AI)“] funded by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia (funding code 005-2011-0053).

Introduction

The use of artificial intelligence (AI) systems is sometimes associated with certain risks of material damage as well as immaterial harm (e.g., in the areas of safety, security, and violation of personal rights through unjustified discrimination). Potential risks of an AI application should therefore be assessed in a systematic manner. In particular, the European regulation calls for AI risk management systems to be introduced in relation to 'high-risk' AI systems, which comprises the identification, analysis, and evaluation of risks as a continuous iterative process. In addition, the regulation includes system-related requirements regarding 'high-risk' AI systems covering aspects such as reliability and robustness, bias examination, and transparency and provision of information, which can be addressed most effectively when the respective risk sources for the specific application are identified.

The purpose of this document is to provide a basis for the structured risk identification and analysis of an AI system in its specific application context with respect to the following three objectives:

- reliability;
- avoidance of unwanted bias; and
- autonomy and control.

Notably, the focus of the document is on AI systems which are primarily realized through Machine Learning (ML). Furthermore, only those risks are considered which are directly linked to the functioning of the software-related aspects of the AI system. Thus, risks regarding distributive justice or environmental protection, for example, are not considered in this document.

Overall, this document is intended to support the identification of risks emerging from an AI system and their analysis. Especially, it can be used by organizations to inform the implementation of their respective risk management processes in accordance with ISO/IEC 23894 (based on DIN ISO 31000:2018-10). It is noteworthy that, contrary to ISO/IEC 23894, this document emphasizes a system perspective on AI risks, and the guidance provided can serve as a technical foundation complementing ISO/IEC 23894.

1 Scope

This document specifies an AI risk scheme covering the objectives

- reliability,
- avoidance of unwanted bias, and
- autonomy and control

along the entire lifecycle of artificial intelligence (AI) systems that include Machine Learning (ML) components.

It aims to serve as a basis for the identification of the AI risks which are present in a specific AI system and their analysis from a product perspective. Thus, it aims to inform the parts of the risk management processes within an organization responsible for the identification and analysis of risks providing specific technical guidance for the mentioned objectives.

The scheme maps risk sources to each objective and gives guidance on risk identification and risk analysis. This document does not specify the risk mitigation.

This document is intended to be applicable to all developers, providers, and operators of AI systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 23894:2023¹, *Information technology — Artificial intelligence — Guidance on risk management*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

DIN and DKE maintain terminological databases for use in standardization at the following addresses:

- DIN-TERMinologieportal: available at <https://www.din.de/go/din-term>
- DKE-IEV: available at <https://www.dke.de/DKE-IEV>

3.1

risk

effect of uncertainty on objectives

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

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

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

[SOURCE: DIN ISO 31000:2018-10, 3.1]

¹ A German version of this document is currently under preparation. Stage at the time of publication: E DIN EN ISO/IEC 23894:2023-11.

3.2**autonomy****autonomous**

characteristic of a system governed by its own rules

[SOURCE: ISO/IEC TR 24028:2020, 3.7 – modified: the definition has been shortened and Note 1 to entry has been deleted]

3.3**reliability**

property of consistent intended behaviour and results

[SOURCE: DIN EN ISO/IEC 27000:2020-06, 3.55]

4 Introduction to AI risk assessment**4.1 Notion of risk**

The term “risk” is typically used to describe the uncertainty about the (future) occurrence of events. In this respect, risk is a measure of both the (adverse) impacts that may arise if the event occurs and the likelihood of occurrence.

In addition to the uncertainty of the occurrence of events that might be fully specifiable in a stochastic model (quantitatively known unknowns), for the identification and analysis of AI risks, it is noteworthy that other types of uncertainty exist. In particular, there is uncertainty as to whether the correct ML model has been chosen for the given task, also called „model risk“ (qualitatively known unknowns). Furthermore, there is “severe uncertainty” as to whether the true model is even close to the set of considered models (unknown unknowns).

When identifying and analysing AI risks, it is thus important that all three types of uncertainty are being considered. For instance, risks may arise if distorted data is fed into an AI application (quantitatively known unknowns) but they also do exist with respect to normal (expected) input data e.g., if the model is not „true“ for the given task.

The following clauses cover the objectives:

- reliability;
- avoidance of unwanted bias; and
- autonomy and control.

For each objective, guidance is provided on the implementation of AI risk identification and analysis. While the objectives are described separately, it should be noted that they are not orthogonal to each other. Risk sources can affect several objectives equally and risks in relation to different objectives can influence each other.

NOTE 1 This document focuses primarily on risks arising from the data-driven nature of the ML component(s) included in an AI system, especially the training of the ML component(s). Additionally, it covers risks that are increased by the use of ML to such extent that ML-specific measures might be required to mitigate them.

NOTE 2 This document addresses the entire lifecycle of an AI system, including AI risks existing during the design and development phase, as well as AI risks arising during the system’s operation.

4.2 AI risk assessment along the lifecycle

Figure 1 shows an example for an AI system life cycle model with AI system-specific processes.

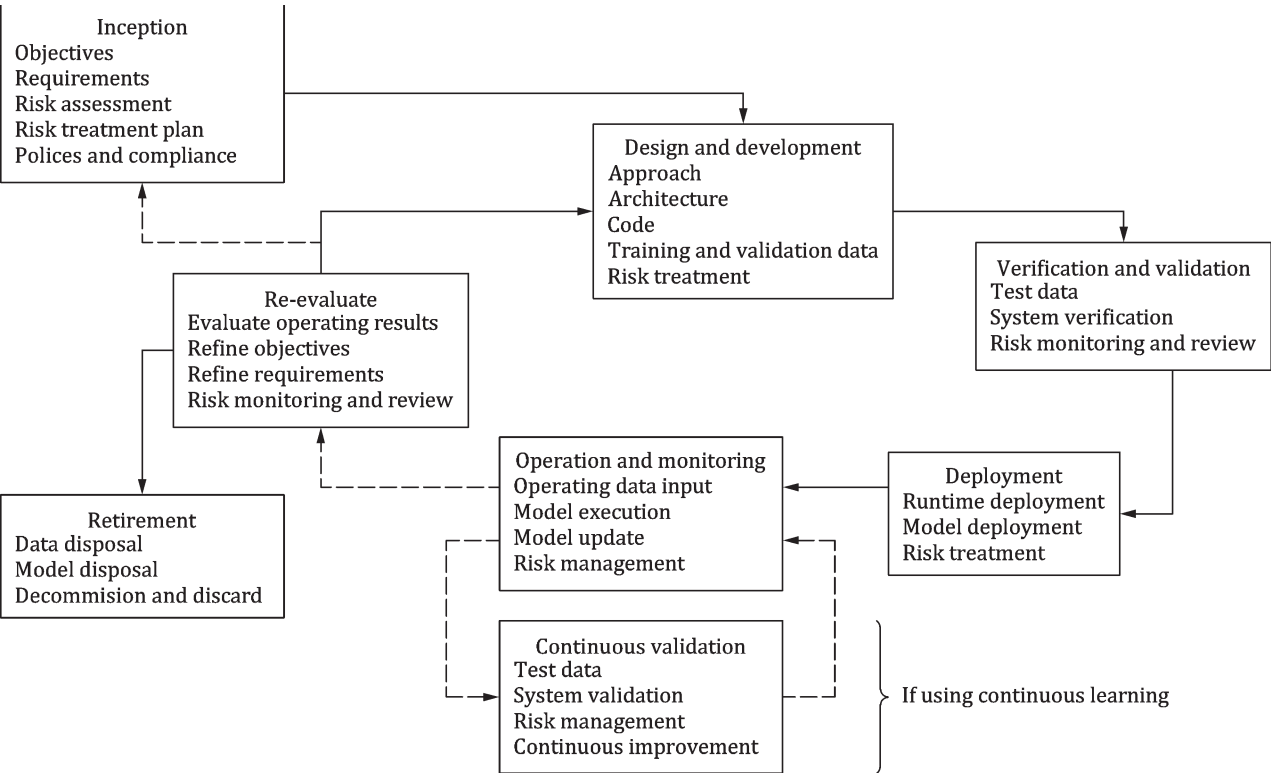


Figure 1 — Example AI system life cycle model with AI system-specific processes [1]

Part of the AI system life cycle model is the inception, which contains the risk assessment. The risk assessment consists of risk identification, risk analysis and risk evaluation, see Figure 1 and Figure 2.

This document covers the risk identification and risk analysis with respect to the objectives listed in 4.1. In addition, further requirements may arise from the application of an AI system, such as monitoring of risks and re-assessment during operation (as part of risk management and re-evaluation). These aspects are application-dependent and are therefore not specified in this document.

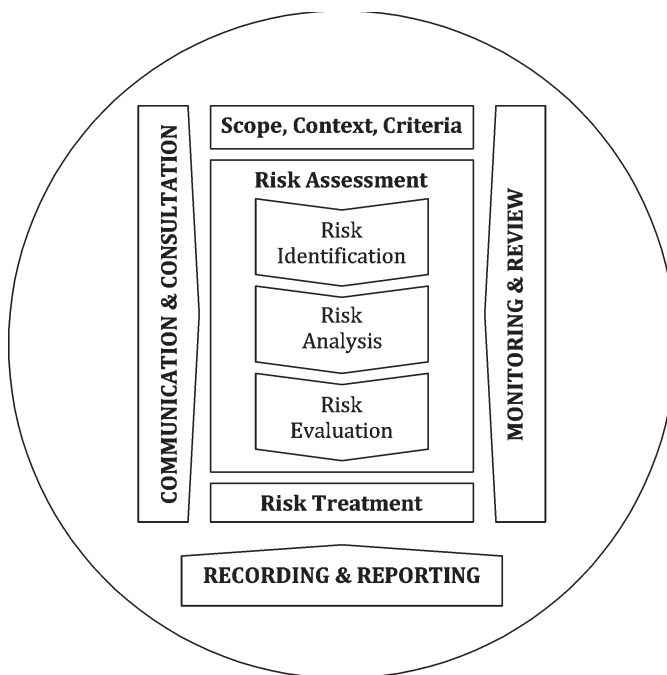


Figure 2 — Risk management process [2]

5 Reliability specifications for AI systems

5.1 General

Malfunction or failure of the ML component(s) included in the AI system can facilitate or even cause (if not mitigated by other system components) personal injury, property, and financial damage. Correspondingly, the objective of reliability refers to the ML component's ability to perform its required function under stated conditions during its operation stage. Notably, reliability is affected by the ML component's robustness i.e., its ability to generalize well to inputs dissimilar to those on which it has been trained and to maintain a consistent level of performance even on atypical data (see ISO/IEC 22989:2022, 5.15.2).

In particular, the reliability of an AI application includes appropriate handling of deviations from standard cases (e.g., in direct human-machine interaction, new error modes that are unexpected or unknown to humans can lead to potentially critical situations that have not been dealt with previously). Thus, not only model performance (e.g., accuracy) is object of consideration in relation to this objective, but also the extent to which potential errors are already prevented by detection strategies at the AI component level (supplementing the traditional methods for fault tolerance and fail-safe operation at the embedding level).

The following subclauses provide guidance on the risk identification and analysis regarding the different aspects of reliability of AI systems.

5.2 Risk identification

From a technical perspective, the reliability of an AI application is an umbrella term that involves different aspects of the quality of its AI component: the correctness of the outputs, the assessment of the ML model uncertainty, robustness to corrupted or manipulated inputs and unexpected situations, and intercepting errors.

An initial risk assessment with respect to the reliability objective should be based on the maximum amount of damage that (incorrect) outputs of the AI application can cause without deliberate subsequent human revision.

EXAMPLE There is generally a damage potential for AI applications embedded in autonomous robotics systems or in decision-making systems that determine aspects such as the distribution of goods or the medical treatment of individuals.

It can be expected that risks with respect to the reliability objective are not classified as low, since otherwise the outputs of the AI application could produce only little or no damage in the event of a malfunction. In particular, it would mean that no bodily injury would occur in the event of incorrect outputs, and potential financial damage, such as loss of earnings or reputational damage, would either not be a threat or would only be a threat to a minor extent. In that case, the AI application would be completely non-critical and could equally be replaced by a non-AI based system that generates random outputs. Table 1 can be used for an initial risk categorization into “high” or “medium”, based on the criticality of potential misprediction and errors of the AI application.

Table 1 — Reliability risk categorization

Risk category	Description	Example
High	An incorrect prediction of the AI application can lead to bodily injury or major financial damage.	Pedestrian detection in autonomous vehicles, automated credit lending decisions, medical treatment recommendations
Medium	An incorrect prediction of the AI application can at worst lead to medium financial damage.	A poor-quality route planner causes increased energy consumption and a longer time of travel, a faulty forecast of machine utilization in a manufacturing facility causes delays, a defective obstacle detection of a robot vacuum cleaner causes property damage
Low	Does not apply.	—
NOTE The amount of damage in all of the stated examples depends on the particular circumstances and does not necessarily come under the category chosen here.		

Focusing on risks emerging from the “unpredictable data-driven nature” of ML-based systems (see ISO/IEC 25059), different categories of data (inside or outside of the application domain) should be considered for identifying potential risk sources. Especially, risks may exist in

- erroneous predictions by the ML component on regular (expected) input data (for guidance on analysis, see 5.3.2),
- erroneous predictions on corrupted or manipulated input data, for which acceptable processing by the ML component is intended (for guidance on analysis, see 5.3.3), to
- the handling of input data that are out of the application domain and for which correct processing by the ML component cannot be expected (for guidance on analysis, see 5.3.4).

In addition, risks with respect to reliability may exist in the uncertainty of ML models, given their typically complex and dynamic application environments. Therefore, risks should also be identified in relation to

- an unrealistic, unusable, or absent uncertainty estimation (for guidance on analysis, see 5.3.5), and
- the ML model’s capability to maintain its level of performance over time, which is particularly related to (unintended) model drifts or changes in the application environment i.e., concept drift.

5.3 Risk analysis

5.3.1 General

In order to systematically analyse risks related to different reliability aspects or requirements, along with the associated threats, the input space needs to be well described (see also ISO/IEC 24029-2). An essential part of the set of possible input data is the “application domain”, which designates the input data expected during operation and that the AI application is meant to process correctly.

In order to analyse the model quality with respect to such regular/standard input data, the application domain should be specified and formalized as precisely as possible (5.3.2). In most cases, corruption in the input data that may occur during regular operation and are part of the application domain are also intrinsically represented in the database. At the same time, for some application contexts, it may make sense to single out such corruptions to better understand how consistently the AI system performs at the boundary of the application domain (5.3.3).

Furthermore, input data that is well outside the defined application domain might pose risks, as it cannot be expected to be processed correctly by the AI component. Depending on the application, such data may better be detected, and potential errors intercepted (5.3.4).

EXAMPLE An uncertainty estimation can be used to identify and respond to weaknesses in the ML model. It can also provide an indication of whether inputs are on the boundary or outside the application domain. However, absent or inaccurate statements about the confidence of the model outputs pose risks regarding the use or further processing of the confidence scores (5.3.5).

The following subclauses provide guidance on how to analyse the risks specific to the AI application for the risk sources mentioned in the first four key points in 5.2.

5.3.2 Reliability in standard cases

5.3.2.1 General

This subclause provides guidance for a universal analysis regarding incorrect predictions of the AI application which can potentially occur on input data from the application domain and lead to damage (for an in-depth analysis of malfunction in the boundary of the application domain, see 5.3.3 and 5.3.5).

As a first step, the (permissible) application domain shall be specified. Based on this defined domain, a comprehensive assessment is made of the type and amount of possible damage that can be caused directly or indirectly by using the AI application for its intended purpose.

5.3.2.2 Specifying the application domain

The following points should be addressed in a qualitative and conceptual manner:

- Application domain: The application domain designates the input data to be expected during operation that the AI application is meant to process correctly. This shall be described in detail. In addition, the application domain should be roughly defined using examples (faulty input data or limiting cases of the application domain, including non-semantic corruptions such as noise, are specified in more detail in 5.3.3.2).
- If possible, the application domain defined here should be distinctly separate from related or supposedly similar application domains for which the AI application was not developed and the reliability for which is thus not assessed.

5.3.2.3 Analysis of potential damage and typical threats

An assessment is made of the risks created by insufficient reliability of the AI application when using regular input data. To this end, the potential damage resulting from reliability insufficiency is described. Moreover, to analyse and quantify different performance aspects of the AI component, various mathematical-statistical measures or metrics (see 5.3.2.4) can be used in tests (under different scenarios, if possible, see 5.3.2.7, 5.3.2.8, 5.3.2.9). To avoid overfitting to a specific metric, i.e., one-sided optimization of the AI component with respect to a single target aspect, as many different performance metrics as possible should be considered to analyse reliability. In addition, when analysing the likelihood of low prediction quality of the model, the following AI-specific threats should be considered:

- training, test and validation data do not cover the application domain adequately (for guidance on quantification, see 5.3.2.5);

- low quality of training and test data (for guidance different data quality aspects, see 5.3.2.6);
- choice of model components (training algorithm, loss function, etc.) is inadequate for the given reliability requirements;
- inappropriate feature selection, design or architecture of the AI component for its application area.

5.3.2.4 Quantification of reliability

When investigating reliability aspects of the AI component, it is important to draw a distinction between general performance metrics and the loss function. The latter is a crucial part of the training of many ML models. In the case of supervised Machine Learning in particular, learning algorithms such as stochastic gradient descent are generally used to adjust the model at each training step in a way that the loss function of the training data set is gradually optimized. The value of the loss function (also referred to as “loss”) with respect to the training data, compared to the loss with respect to the test data, provides information about the quality of the training and can indicate aspects such as possible overfitting.

However, a performance metric is the general term for a measure or metric that assesses how good/bad a model or the AI application based on it is at solving a given task (see also ISO/IEC TS 4213:2022 for further information on metrics for assessing ML classification performance). The performance metric used to assess the quality of a model can differ from its loss function (if there is one). In particular, it is not part of the training process, but is actually determined on the basis of validation or test data to measure the performance of the AI component in relation to a given task.

When using performance metric(s) to assess the reliability of the AI application in its application domain, justification should be provided as to why the chosen metrics are appropriate for evaluating the fulfillment of a task by the AI application and adequately reflect the specific quality requirements of the model that, for example, stem from the business context of the AI application. If the model is trained with a loss function, it should also be investigated whether the choice of the loss function is reasonable and adequate for the given task.

Annex A provides an overview of key performance metrics for different (also unsupervised) Machine Learning tasks. Various other metrics also exist.

5.3.2.5 Quantification of the application domain coverage

Adequate (quantitative) coverage of the application domain by training, validation and test data is a key factor for ensuring that the ML model considers all possible application scenarios and therefore learns optimal decision rules. Coverage of the application domain by these data sets should be formalized and captured using a quantitative metric, if possible.

EXAMPLE If the input space can be formalized as a low-dimensional vector space and the application domain as a subset of it, then the simplest metric to check is whether each cell of a grid in that area contains data points. It is often possible to make use of additional criteria to evaluate the “newness” of created data points.

If adequate coverage of the application domain justifiably cannot be demonstrated using quantitative means, qualitative reasoning may be used instead. Particularly in contexts such as open-world applications in which this type of review can never be conclusive, there should be a strategy to consider (e.g., logging) new scenarios and input data during operation.

5.3.2.6 Quality of training and test data

In order to ensure reliable performance of the AI component in its application domain, qualitative requirements should be set for the data in addition to data coverage, as these can have an equal effect on the quality of the ML model. The following aspects should be considered as a minimum:

- The truthfulness of the data and the accuracy of annotations/labels, which is essential to ensure that the AI component draws the right conclusions from the learned context. Data processed and stored shall be accurate where possible, particularly with regard to personal data.