

ICS 35.080; 35.240.01

**Artificial Intelligence –
Life Cycle Processes and Quality Requirements –
Part 2: Robustness**

Künstliche Intelligenz –
Life Cycle Prozesse und Qualitätsanforderungen –
Teil 2: Robustheit

Intelligence artificielle –
Processus du cycle de vie et exigences pour la qualité –
Partie 2: Exigences techniques et organisationnelles

Gesamtumfang 43 Seiten

This document has been developed and approved by the authors named in the Foreword.

Contents

	Page
Foreword	3
Introduction.....	5
1 Scope	6
1.1 Field of Application	6
1.2 Limitations.....	6
2 Normative References.....	7
3 Terms and Definitions	7
3.1 General Terminology	7
3.2 Terminology – Adversarial Robustness	7
3.3 Terminology – Corruption Robustness.....	9
3.3.1 Corruption Robustness – General terminology.....	9
3.3.2 Distributional Shift/ Dataset Shift.....	10
3.3.3 Sample Selection Bias.....	10
4 AI Quality Metamodel.....	11
5 Robustness.....	13
5.1 Introduction to AI Robustness.....	13
5.2 Requirements and Guidelines on Risk Management.....	14
5.2.1 Overview	14
5.2.2 Scope, Context, and Criteria.....	17
5.2.3 General Goals and Objectives.....	19
5.3 Requirements specific to Adversarial Robustness.....	20
5.3.1 Scope, Context, and Criteria.....	20
5.3.2 Threat Model Analysis	20
5.3.3 Likelihood & Impact Analysis.....	22
5.3.4 Robustness Evaluation	27
5.3.5 Mitigations	29
5.4 Requirements specific to Corruption Robustness.....	32
5.4.1 Scope, Context, and Criteria.....	32
5.4.2 Threat Model Analysis	32
5.4.3 Likelihood & Impact Analysis.....	34
5.4.4 Robustness Evaluation	37
5.4.5 Mitigations	39
6 Implementation Guidelines	41
Bibliography	43

Foreword

This DIN SPEC has been developed according to the PAS procedure. The development of a DIN SPEC according to the PAS procedure is carried out in DIN SPEC (PAS)-consortiums and does not require the participation of all stakeholders.

This document has been developed and adopted by the initiator(s) and authors named below:

- Microsoft Deutschland GmbH
Deussen, Peter, Dr. (Workshop-Leiter)
- neurocat GmbH
Assion, Felix (stellv. Workshop-Leiter)
- Robert Bosch GmbH
Abrecht, Stephanie
- neurocat GmbH
Besold, Tarek R., Dr.
- Hochschule für Angewandte Wissenschaften (HAW)
Dai, Zhen Ru, Prof. Dr.-Ing.
- TÜV Süd Auto Service GmbH
Eicher, Matthis, Dipl.-Ing.
- neurocat GmbH
Firnkorn, Jörg, Dr.
- neurocat GmbH
Günther, Wiebke
- Fraunhofer — Institut für Offene Kommunikationssysteme FOKUS
Hackel, Sascha
- neurocat GmbH
Hinze, Stephan (Initiator)
- IABG mbH
Höhndorf, Lukas, Dr.-Ing.
- neurocat GmbH
Liegl, Christian
- Hochschule für Technik und Wirtschaft Berlin (HTW) FB4 Wirtschaftswissenschaften
de Meer, Jan, Dipl.-Inform., Dipl.-Ing.
- Fraunhofer — Institut für Offene Kommunikationssysteme FOKUS
Rennoch, Axel

- neurocat GmbH
Schwerdtner, Paul
- Robert Bosch GmbH
Sudholt, Sebastian

At present, there are no standards covering this topic in the body of German Standards.

DIN SPEC (PAS)s are not part of the body of German Standards.

A draft of this DIN SPEC (PAS) has not been published.

Despite great efforts to ensure the accuracy, reliability and precision of technical and non-technical information, the DIN SPEC (PAS)-consortium cannot give any explicit or implicit assurance or warranty in respect of the accuracy of the document. Users of this document are hereby made aware that the consortium cannot be held liable for any damage or loss. The application of this DIN SPEC (PAS) does not release users from the responsibility for their own actions and is applied at their own risk.

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.

Provision of this document free of charge as a PDF via the Beuth WebShop has been financed in advance.

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

Introduction

Artificial Intelligence (AI) is an interdisciplinary and complex field. It provides a tool to cover tasks that are usually associated with human intelligence. Despite the fact that the operating principles governing some AI methodologies are still an active field of research, the performance of AI modules that is often superior to traditional software has led to increasing deployment of AI-based technologies.

The transfer from research field to economically and societally valuable technology, however, necessitates that a quality and reliability concept is established. In AI, not only traditional software quality issues need to be considered but novelties such as the lack of insight into the underlying logic of an AI module which is characteristic for the subfield Machine Learning (ML) also need to be addressed. Furthermore, some AI modules change during operation and thus must be continuously monitored so that AI validation and verification cannot be considered complete after development.

Therefore, this document defines requirements that go beyond traditional software quality requirements as defined in [1]. DIN SPEC 92001-1 establishes an AI quality metamodel and lifecycle for AI modules to highlight quality characteristics of AI modules. In DIN SPEC 92001-2, this document, specific requirements that ensure AI quality with respect to robustness are provided. Specifically, the AI quality pillar "robustness" is explained further and specific requirements for this pillar are listed. Furthermore, each requirement within the pillars is mapped to a set of lifecycle stages to facilitate the temporal classification of the requirements. The robustness requirements contained in this document have been developed with the aim of being comprehensive regarding both adversarial and corruption robustness in order to support the development and deployment of safe and secure AI innovations.

1 Scope

1.1 Field of Application

This DIN SPEC applies to all life cycle stages of an AI module – concept, development, deployment, operation, and retirement – and addresses a variety of different life cycle processes. Due to the fact that AI technologies are used for a wide range of different tasks, this DIN SPEC is not only aimed at one particular field but also applies to companies and AI products across all sectors.

This DIN SPEC applies to all kinds of AI modules including ML and expert systems.

1.2 Limitations

This document does not define or list algorithms, methods, or technologies which are part of AI. Hence, the user of the DIN SPEC is required to evaluate whether the given AI quality metamodel and the associated AI quality requirements apply.

While requiring an evaluation of ethical conduct in the development of an AI module, this DIN SPEC does not provide any specific requirements that define ethical conduct.

The software life cycle considerations in this document are compatible with ISO/IEC/IEEE 12207:2017, *Systems and software engineering - Software life cycle processes* [1].

The DIN SPEC proposes the separation between AI modules with high and low risk concerning safety, security, privacy, and ethical relevance. It also provides relevant aspects in the context of risk evaluation. Neither a precise risk evaluation process, nor an ethical design framework is established in this document. It is, however, constrained by the existing legal framework and each institution's ethical code of conduct. Compliance with regulations is presumed. Furthermore, the stakeholders of this document are required to form an expert group to assess their AI module's risk profile.

The quality requirements listed in DIN SPEC 92001-2 are not domain-specific. There is room for the requirements given in this document to be extended to include specific fields of application of AI in a subsequent standardization step.

2 Normative References

There are no normative references in this document.

3 Terms and Definitions

For the purposes of this document, the following terms and definitions apply. ISO, IEC, and IEEE maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at <http://www.electropedia.org/>
- ISO Online browsing platform: available at <http://www.iso.org/obp>

3.1 General Terminology

This document is based on the terms and definitions of [1].

3.1.1

artificial intelligence (AI)

interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning [2]

3.1.2

artificial intelligence (AI) module

software module consisting of AI algorithms

3.2 Terminology – Adversarial Robustness

3.2.1

adversarial attack

an algorithm developed by an adversary which returns adversarial perturbations or adversarial examples

3.2.2

adversarial capabilities

the various actions, information, techniques, or attack vectors available to an attacker on a threat surface [3]

3.2.3

adversarial example

input to an AI module that an attacker has intentionally designed to cause the model to make a mistake [4]

3.2.4

adversarial example transferability

the property that adversarial examples crafted to induce malperformance in one model are likely to induce malperformance in a different model [3]

3.2.5

adversarial perturbation

a carefully crafted perturbation which is designed to cause malperformance when added to or combined with one or more input data points of the AI model

3.2.6

adversarial robustness

the ability of an AI module to cope with adversarial examples (or adversarial perturbations)

3.2.7

adversary

a malicious individual who wishes to attack the AI module either during learning by tampering with the training data, or during inference by manipulating inputs on which the model is making predictions [5]

3.2.8

adversary's knowledge

adversary's knowledge of the AI module, e.g. including training data, model architectures, hyperparameters, numbers of layers, activation functions, model weights [6]

3.2.9

attack detection

the action of differentiating between anomalous and normal behaviour, or between an adversarial example and a benign example [3]

3.2.10

availability violation

a compromise of the normal system functionalities available to legitimate users, such as accuracy, quality, or access, resulting in inaccessible or unusable model output [3]

3.2.11

blackbox model

the adversary has limited or no knowledge about the model under attack and might not be allowed to probe or query the model while constructing adversarial examples [3]

3.2.12

defence strategy

a set of mitigation techniques or countermeasures to secure an AI module against adversarial attacks

3.2.13

individual perturbation scope

property of an adversarial perturbation, namely that the perturbation is linked to one specific input point and tries to fool the AI module on this chosen input point

3.2.14

integrity violation

to induce a particular output or behaviour of the adversary's choosing [3]

3.2.15

non-targeted attack

an attack that causes any misclassification as opposed to causing classification into a specific (incorrect) class [3]

3.2.16

privacy violation

revealing personal information about an individual included in the training data [3]

3.2.17**targeted attack**

the adversary tries to produce inputs that force the output of the classification model to be a specific target class [7]

3.2.18**universal perturbation scope**

property of an adversarial perturbation, namely that the perturbation is image-agnostic and tries to fool the AI module on the majority of inputs [7]

3.2.19**whitebox model**

the adversary has full knowledge of the model including model type, model architecture and values of all parameters and trainable weights [8]

3.3 Terminology – Corruption Robustness**3.3.1 Corruption Robustness – General terminology****3.3.1.1****anomaly**

a pattern in the data that does not conform to the expected behaviour [9]

3.3.1.2**anomaly detection**

strategies and methods to discover anomalies and outliers [9]

3.3.1.3**change detection**

strategies and methods to discover changes in single inputs or the input distribution [9]

3.3.1.4**data corruption**

the process by which data is changed from its original format to one regarded as noisy, flawed, or defective [10]

3.3.1.5**error set**

the set of points in the input space on which the classifier makes an incorrect prediction [13]

3.3.1.6**noisy environment**

environment of an AI module where variables are observed with an error [14]

3.3.1.7**non-stationary environment**

environment of an AI module where the underlying data distribution changes over time [14]

3.3.1.8**outliers**

a datapoint that differs significantly from other observations [15]

3.3.1.9

prediction consistency

a property guaranteeing that, as the sample size increases, the sampling distribution of the prediction becomes increasingly concentrated at the true value. [10]

3.3.1.10

prediction reliability

an AI module is said to have a high prediction reliability if it produces similar results under consistent conditions [10]

3.3.1.11

spatial robustness

robustness of an AI module to geometric transformations, in particular translation and rotation [16]

3.3.1.12

unknown unknowns

data points which did not occur in the training or test data of the AI module (that the AI module cannot process correctly) [17]

3.3.2 Distributional Shift/ Dataset Shift

3.3.2.1

distributional shift / dataset shift

joint distribution of inputs and outputs differs between training or test and deployment stage [11]

3.3.2.2

concept shift

concept shift happens when the relationship between input and output changes [12]

3.3.2.3

covariate shift

distributional shift of the input variables between the training or test and deployment stage [12]

3.3.2.4

prior probability shift

change of distribution of the output variable between the training or test and deployment stage [12]

3.3.3 Sample Selection Bias

3.3.3.1

sample selection bias

difference between the distributions of the training or test data and the deployment distribution as a result of an unknown sample rejection process [11]

3.3.3.2

missing at random

missing at random occurs when the sampling method depends on the input features given the condition that the input features are independent of the output. This kind of bias can potentially produce covariate shift [12]

3.3.3.3

missing at random-class

missing at random-class occurs when the sampling method depends on the output given the condition that the output is independent of the output of the input features. This kind of bias can potentially produce prior probability shift [12]

3.3.3.4

missing completely at random

missing completely at random occurs when the sampling method is completely independent of the input features and the output [12]

3.3.3.5

missing not at random

missing not at random occurs when there is no independence assumption between the sampling method, the input features, and the output. This kind of bias can introduce one or more covariate shifts, prior probability shifts and concept shifts [12]

4 AI Quality Metamodel

The AI quality metamodel is introduced in DIN SPEC 92001-1, Figure 1. It includes the most important aspects that need to be taken into account to facilitate the design of high quality AI modules. The life cycle of an AI module, illustrated by the circle at the centre of Figure 1, consists of the life cycle stages "concept", "development", "deployment", "operation", and "retirement". This quality metamodel recognizes three major groups of life cycle processes that need to be considered during the life cycle stages. These are organizational project-enabling, technical management, and technical processes. The processes assigned to these three groups can be found in [14]. The quality requirements for AI modules, even more so than in classical software, need to be linked to the life cycle. This becomes particularly apparent when considering ML applications which evolve through acquiring information while being deployed in the real world, known as online learning. Thus, certain requirements concerning, for example, the prevention of harmful bias, are required for prototyping as well as during the deployment and operation of the AI module's life cycle.

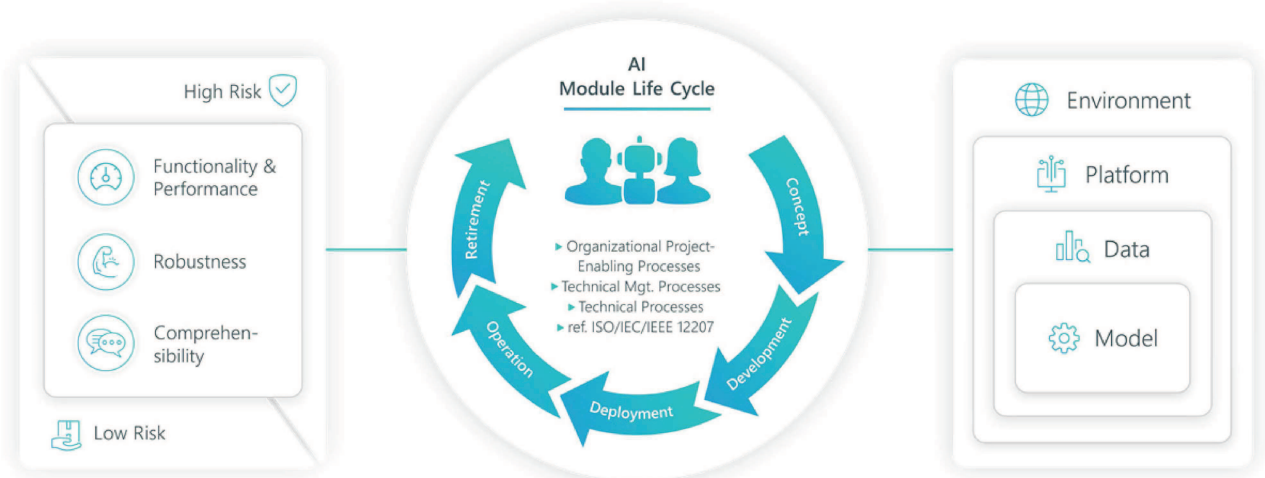


Figure 1 — AI quality metamodel

The key quality characteristics, the so-called quality pillars, that need to be taken into account throughout the whole life cycle of an AI module are functionality & performance, robustness, and comprehensibility. They are depicted in the left-hand part of Figure 1. These quality pillars entail the most pressing AI-specific quality challenges. They indicate central deployment issues of AI modules in comparison to classical software, such as training set construction and robustness with respect to adversarial examples.

In this DIN SPEC, each AI module is considered to be either of high or low risk. AI modules with safety, security, privacy, and ethical relevance are classified as modules with (potentially) high risk. AI modules without such relevance are considered to be of low risk. For high risk AI modules, this document requires the consideration of all listed quality requirements. Potential deviations of high-risk AI modules from the listed quality requirements need a sound justification. This requirement is relaxed for low risk AI modules. This risk classification is shown in Table 1.

Table 1 – Categorization of requirements concerning AI modules with high or low risk in mandatory, highly recommended, and recommended

AI module class Requirement class	High risk	Low risk
Mandatory	No deviation from requirements allowed	No deviation from requirements allowed
Highly recommended	Deviation from requirements with justification only	Deviation from requirements with justification only
Recommended	Deviation from requirements with justification only	Deviation from requirements without justification allowed

In Figure 1, this underlying risk evaluation is visualized as a rectangle placed behind the three quality pillars. The first step in the concept stage of an AI module is to assess whether the AI module in question has potentially high or low risk concerning safety, security, privacy, and ethical aspects. This is not dealt with in this DIN SPEC. A body of experts within an organization responsible for the AI module is to be convened for this task. The risk assessment has an influence on the rigour that needs to be applied when implementing and evaluating the properties and quality of the AI module.

5 Robustness

5.1 Introduction to AI Robustness

In the following, robustness shall denote the capability of an AI module to cope with erroneous, noisy, unknown, or adversarially constructed input data. Due to the fact that an AI module is situated in a potentially highly complex environment (the environment may be non-stationary in nature, but also high-dimensional in terms of the data-generating process), robustness is a key AI quality issue. Therefore, robustness is introduced and treated in this document as a fundamental pillar of AI quality.

In this document, a distinction is made between adversarial robustness (AR), i.e. robustness with respect to adversarial perturbations, and corruption robustness (CR), i.e. robustness with respect to noisy signals or changes in the underlying data distribution. These two specific robustness objectives give substance to the rather general definition of robustness by encompassing two essential topics and sets of criteria of AI quality.

AR requires measures that protect the AI module from so-called adversaries attempting to deceive the AI module by means of carefully chosen harmful input. The last years of AI research and development have shown that state-of-the-art AI modules are vulnerable to adversarial examples. This poses a grave risk when AI solutions are applied in safety- and security-critical situations.

On the other hand, CR requires measures that reduce or even prevent complications arising from differences between the data sets used in the development and deployment stage, respectively. There are various types of distributional or data set shifts that have to be addressed within the AI lifecycle, for instance covariate shifts and prior probability shifts. The most prominent causes of distributional shifts are selection bias and the non-stationarity of the environment or platform.

Note that there are some measures that improve both AR and CR. This is due to the fact that both AR and CR ensure performance of the AI module when it is exposed to data points that have been thought of as highly unlikely or even data points outside the distribution that was seen during training. An adversarial example may even be interpreted as a special kind of distributional shift that was specifically designed to cause the AI module to deteriorate in performance.

Nevertheless, a distinction is still made between AR and CR. There are various reasons for this. Most prominently, since an active adversary is assumed in the AR setting, a continuous "arms race" between attack and defence needs to be taken into account, whereas CR is usually addressed as an issue arising from natural causes such as hardware deterioration or otherwise compromised input. Therefore AR and CR are respectively cast as a security and a safety problem.

While CR corresponds to and is limited to situations where there is no intentional design behind critical environments that lead to performance deterioration of the AI module (hence CR is dubbed "optimization-free"), AR is an essentially optimization-driven field of research. However, both CR and AR may have to take into account, and be tested in, three different test environments, i.e. the digital, simulated, and physical domain.

The remainder of this document is organized as follows. At the highest level, clauses 5.3 and 5.4 contain the main contributions of this document, i.e. requirements concerning AR and requirements concerning CR. Each clause is split into four subclauses referred to as threat model analysis, likelihood and impact analysis, robustness evaluation, and mitigations. While the threat model analysis supports awareness or increases understanding of potential vulnerabilities of an AI module, the likelihood and impact analysis quantifies the likelihood of malperformance and the associated damage or cost. The subsequent robustness evaluation prescribes a test plan that may be taken into account as mitigations are finally formulated